UNIVERSIDAD DE SALAMANCA

DEPARTAMENTO DE ESTADÍSTICA

TESIS DOCTORAL

Doctorado en Estadística Multivariante Aplicada

ANÁLISIS SPARSE DE TENSORES MULTIDIMENSIONALES

NEREA GONZÁLEZ GARCÍA

DIRECTOR: María Purificación Galindo Villardón

Ana Belén Nieto Librero

2019

ANÁLISIS SPARSE DE TENSORES MULTIDIMENSIONALES

DEPARTAMENTO DE ESTADÍSTICA

Memoria para optar al Grado de Doctor
en Estadística Multivariante Aplicada
por el Departamento de Estadística de la
Universidad de Salamanca, presenta:

Nerea González García

Salamanca

2019

VNiVERSiDAD D SALAMANCA

CAMPUS DE EXCELENCIA INTERNACIONAL

DEPARTAMENTO DE ESTADÍSTICA

DRA. MARÍA PURIFICACIÓN GALINDO VILLARDÓN

Catedrática de la Universidad de Salamanca del Área de Estadística e Investigación Operativa

y

DRA. ANA BELÉN NIETO LIBRERO

Ayudante Doctor del Departamento de Estadística de la Universidad de Salamanca

CERTIFICAN:

Que **Dña. Nerea González García,** graduada en Matemáticas, ha realizado en el Departamento de Estadística de la Universidad de Salamanca, bajo su dirección, el trabajo para optar al Grado de Doctor en Estadística Multivariante Aplicada, que presenta con el título "***Análisis Sparse de Tensores Multidimensionales"***, autorizando expresamente su lectura y defensa. Y para que conste, firman el presente certificado en Salamanca a 25 de Noviembre de 2019.

M.ª Purificación Galindo Villardón

Ana B. Nieto Librero

“Lo que no se hace sentir
no se entiende,
lo que no se entiende
no interesa”

Simón Rodríguez

“Esfuerzo hoy,

éxito mañana”

AGRADECIMIENTOS

A mis directoras, Dra. Purificación Galindo y Dra. Ana Belén Nieto por todos sus conocimientos, sus consejos, sus enseñanzas y su confianza depositada en mí. Juntas me habéis guiado por este mar de incertidumbre. Puri, a nivel profesional te agradezco que me hayas abierto el mundo de la estadística, todo lo que me has enseñado y todos los buenos momentos vividos, desde mi primer congreso hasta las horas de despacho con desarrollos matemáticos. A nivel personal, agradezco que hayas confiado en mí desde el principio y que me hayas dado la oportunidad de formar parte de este equipo, con todas las herramientas que estaban a tu disposición, siempre que te fue posible. Ana, lo más bonito del trabajo doctoral es encontrar a una persona que te entienda, comprenda y comparta contigo tantas cosas. Siempre estaré eternamente agradecida por tus cuidados, consejos, por las innumerables veces que me has prestado tu ayuda, horas incontables, y por haberme acompañado a lo largo de todo el camino que estoy segura, solo acaba de empezar. No tengo palabras de agradecimiento.

A todos los profesores del departamento de estadística, por sus enseñanzas, su disposición, y por hacerme sentir como una más; particularmente a Carmen, por los momentos compartidos, su apoyo desinteresado y estar disponible siempre que lo he necesitado. A Lola y Ana Belén, por su disponibilidad y colaboración.

A mis amigos del doctorado, que pronto estoy segura de que viajaremos juntos por el mundo y brindaremos por todos los momentos vividos, recordando cada uno de ellos desde la sonrisa más grande. Vicky, Zaira, Carmen, Mitzi, María, Joel, Cinthia, Guille, John, Carlos.. y en especial, a mi buen amigo Greibin. Tú me has enseñado que la constancia siempre llega a buen puerto y que la amistad no habla de edad.

A mis compañeros de laboratorio, Álex, Marta, Violeta, Rodri. Con vosotros descubrí una Salamanca diferente y me habéis hecho formar parte de una familia desde el primer día, a pesar de que cada uno vaya comenzando su propio camino. Álex, te agradezco tu amistad, tus momentos de locura y también cordura, por ponerme los pies en la tierra y con tu forma de ver la vida hacer que yo sea hoy un poquito más yo, de lo que antes era. A mis amigas de Salamanca y de Vitoria, mis compañeras de equipos; con especial cariño, Nieves, Clara,

Víctor, gracias por todo lo que habéis hecho por mí, mucho más de lo que os podáis imaginar. Que dura es Salamanca sin vosotros. Sara, mi otra gemela, sin darte cuenta hiciste de nuestra vida en Salamanca un hogar.

Para ti, mi coach, no tengo palabras. Ni este trabajo, ni mi vida, sería así sin ti. Cada vez tengo más claro que mi decisión fue totalmente acertada. La palabra más bonita, del baloncesto y de la vida, la he comprendido a tu lado, equipo. Y como bien dijo una sabia hace relativamente poco, esta es la forma más bonita de recorrer la vida.

A mis padres, Belén, Dani y toda mi familia, que aun sin comprender este mundo han hecho todo lo posible para que este trabajo, y yo misma, llegase a su destino. Emma, sé que dentro de unos años miraremos atrás y juntas comprenderemos el camino recorrido. Sé que contigo los momentos no serán momentos perdidos.

A todas las personas que, sin saberlo, han contribuido a que este proyecto viera su fin.

Gracias.

A mis padres

A Emma

RESUMEN

Una de las áreas más importantes de la investigación actual en el análisis de datos multivariantes se centra en el desarrollo de técnicas eficientes para el estudio matrices de datos de altas dimensiones. En disciplinas como la genética o el procesamiento de imágenes, las bases de datos están formadas por miles de variables. Para su análisis, se requieren técnicas que las simplifiquen y que no pierdan la información clave de la muestra. El análisis de componentes principales, mediante la descomposición en valores singulares, es la técnica más implementada para la reducción de la dimensión de matrices de datos y extracción de variables características, lo que se logra a través de la extracción de nuevas variables latentes conocidas como componentes principales. Sin embargo, presenta el inconveniente de que cada componente principal es una combinación lineal de todas las variables originales y esto dificulta su interpretación. A lo largo de los años se han desarrollado distintos enfoques para paliar su principal desventaja, pero es en la última década en la que este método se ha modificado para producir componentes principales sparse; es decir, componentes que envuelvan únicamente un pequeño subconjunto de las variables originales más importantes. Todo ello ha dado lugar a la aparición del análisis de componentes principales sparse, un método de selección automática de variables características extremadamente útil en las aplicaciones modernas donde el número de variables originales es enorme.

El proyecto aquí propuesto investiga y propone una nueva herramienta de análisis aplicable a un tipo especial de datos, conocidos en estadística como datos multivía o, más recientemente en minería de datos, como tensores. Hasta ahora, los estudios recogían la información en matrices bidimensionales, pero en la actualidad existen ocasiones en las que es interesante y necesario englobar dicha información en bloques de más vías, incluyéndose más cantidad de información en el estudio. Analizándose estos tensores mediante los métodos de descomposición pertinentes se obtendrían respuestas de manera más eficaz (en términos de solución, tiempo y área de aplicación) que si dichas matrices se analizasen por separado. Estos métodos, como los modelos Tucker o el método STATIS, aplicados en diversas áreas, basan su fundamento teórico en el análisis

de componentes principales clásico, cuya deficiencia es bien conocida, y en la descomposición en valores singulares, que no tiene una definición única en el caso de tensores.

Nuestro trabajo se enfoca en una línea de investigación que acaba de comenzar: el desarrollo de un método sparse generalizado, adaptado al análisis de datos multidimensionales. Para ello, se desarrolla una nueva formulación matemática de la descomposición en valores singulares, C_{enet}SVD, restringida para la obtención de vectores singulares ortogonales y sparse al mismo tiempo mediante la penalización Elastic net. Dicha implementación es extendida al análisis de dos vías, proponiendo C_{enet}PCA, análisis de componentes principales restringido, sparse y ortogonal, y C_{enet}Biplot, métodos Biplot con componentes sparse y ortogonales. Finalmente, la formulación se generalizará a los modelos Tucker de tres vías, para producir matrices de componentes sparse y ortogonales en el conocido como C_{enet}Tucker. Las metodologías propuestas serán aplicadas en distintos campos de conocimiento, examinando así su utilidad en disciplinas tan diversas como psicología y genética. Gracias a las herramientas matemáticas, se abre así un nuevo camino en la estadística multivariante, con importantes aplicaciones prácticas en cualquier área de la actividad humana.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

NOTACIÓN

$x \in \mathbb{R}$	Número real
$x \in \mathbb{R}^+$	Número real no negativo
$\text{sign}(x)$	Función signo (= 1 si $x > 0$ o $= -1$ si $x < 0$)
$\lvert x \rvert$	Función valor absoluto
$\boldsymbol{x} \in \mathbb{R}^I$	Vector de números reales y longitud I
x_i	i-ésimo elemento del vector $\boldsymbol{x}$
$\lVert \boldsymbol{x} \rVert_p = \left(\sum_j \lvert x_j^p \rvert \right)^{\frac{1}{p}}$	Norma $L_p = \ell_p$ de un vector $\boldsymbol{x}$
$\lVert \boldsymbol{x} \rVert_2 = \sqrt{\boldsymbol{x}^T \boldsymbol{x}}$	Norma $L_2 = \ell_2$ de un vector $\boldsymbol{x}$
$\lVert \boldsymbol{x} \rVert_1 = \sum_{j=1}^{J} \lvert x_j \rvert$	Norma $L_1 = \ell_1$ de un vector $\boldsymbol{x}$
$\boldsymbol{x} / \lVert \boldsymbol{x} \rVert_2$	Vector $\boldsymbol{x}$ normalizado
$\boldsymbol{X} = (x_{ij}) \in \mathbb{R}^{IxJ}$	Matriz de números reales de dimensión IxJ
I	Número de observaciones
J	Número de variables
$\boldsymbol{X}^T$	Matriz traspuesta
$\boldsymbol{X}^{-1}$	Matriz inversa de una matriz cuadrada y no singular $\boldsymbol{X}$
$\boldsymbol{X}^\dagger$	Matriz pseudo-inversa de Moore-Penrose de $\boldsymbol{X}$
$det(\boldsymbol{X})$	Determinante de una matriz X
$\text{tr}(\boldsymbol{X})$	Traza de una matriz $\boldsymbol{X}$
$\lVert \boldsymbol{X} \rVert_F^2 = tr(\boldsymbol{X}^T \boldsymbol{X})$	Norma de Frobenius de una matriz $\boldsymbol{X}$ al cuadrado
$\mathfrak{B}_\tau^{\ell_2}(\boldsymbol{x}) = \{\boldsymbol{x} \,/\, \lVert \boldsymbol{x} \rVert_2^2 \leq \tau\}$	Región restringida al balón Ridge (norma L2)
$\mathfrak{B}_\tau^{\ell_1}(\boldsymbol{x}) = \{\boldsymbol{x} \,/\, \lVert \boldsymbol{x} \rVert_1 \leq \tau\}$	Región restringida al balón Lasso (norma L1)
$\mathfrak{B}_\tau^{\ell_1+\ell_2}(\boldsymbol{x}) = \{\boldsymbol{x} \,/\, (1-\alpha)\lVert \boldsymbol{x} \rVert_1 + \alpha \lVert \boldsymbol{x} \rVert_2^2 \leq \tau\}$	Región restringida al balón Elastic net para algún $\alpha \in [0,1]$
$\text{diag}()$	Matriz diagonal
$\underline{\mathbf{X}} = (x_{ijk}) \in \mathbb{R}^{IxJxK}$	Tensor de orden 3 (matriz de tres vías) de dimensión $IxJxK$

K	Número de condiciones (modo 3)
$\underline{\mathbf{G}} = (x_{ijk}) \in \mathbb{R}^{PxQxR}$	Matriz Core del modelo Tucker de orden $PxQxR$
P, Q, R	Dimensiones de los modos 1 (individuos), 2 (variables) y 3 (condiciones)
$\boldsymbol{A} \in \mathbb{R}^{IxP}$	Matriz del primer modo (individuos x componentes)
$\boldsymbol{B} \in \mathbb{R}^{IxP}$	Matriz del segundo modo (variables x componentes)
$\boldsymbol{C} \in \mathbb{R}^{IxP}$	Matriz del tercer modo (condiciones x componentes)
$*$	Producto de Hadamard
$\otimes$	Producto de Kronecker
$\odot$	Producto Khatri-Rao
X_n	Producto n-modo de un tensor matriz

ABREVIATURAS

SVD	*Singular Value Decomposition*
CSVD	*Constrained Singular Value Decomposition*
C_{enet}SVD	*Constrained to Elastic net Singular Value Decomposition*
C_{enet}PCA	*Constrained to Elastic net Principal Component Analysis*
C_{enet}Biplot	*Constrained to Elastic net Biplot*
C_{enet}Tucker	*Constrained to Elastic net Tucker models*
FA	*Factor Analysis*
PCA	*Principal Component Analysis*
PC	*Principal Component*
NMF	*Nonnegative Matrix Factorization*
MFA	*Multiple Factor Analysis*
JIVE	*Joint and Individual Variation Explained*
MDS	*Multidimensional Scaling*
CCA	*Canonical Correspondence Analysis*
SPCA	*Sparse Principal Component Analysis*
PARAFAC	*Parallel Factor Analysis*
CP	*CANDECOMP/PARAFAC*
OLS	*Ordinary Least Squares*
CV	*Cross-Validation*
MSE	*Mean Square Error*
BIC	*Bayesian Information Criterion*

INTRODUCCIÓN

La obtención de conocimiento a través de la extracción y comprensión de conjuntos de datos experimentales, observaciones y medidas es el día a día de múltiples disciplinas. La recolección de bases de datos analíticas da lugar a matrices de información. Un análisis apropiado de ellas permitirá caracterizar los objetos en estudio, traduciendo esta información en conocimiento. Ahora bien, 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, bien por desconocimiento del comportamiento de la población o simplemente para un uso exploratorio.

La estadística descriptiva univariante queda muy lejos de proporcionar óptimos resultados, pues su enfoque tradicional partía de pequeñas muestras de una población. Es un requisito imprescindible el uso de técnicas estadísticas multivariantes que, acompañadas de la informática, asuman un rol principal dentro del 'corpus científico'. Deben aprovechar toda la información de los datos para generar una buena reproducción de la realidad, pues el poder de los estudios individuales y la comprensión de todos los fenómenos radica en la visión multivariante del mundo (Smilde, Bro, & Geladi, 2004).

En la actualidad, los recientes avances en tecnología, la generalización de Internet y de las TIC, la disminución de los costes de procesamiento y almacenamiento de los grandes repositorios de datos, están provocando un cambio paradigmático en la sociedad que podríamos definir como la 'Revolución de los Datos', que dan lugar a bases de datos formadas por miles de variables, conocidas como matrices de datos de altas dimensiones o, por su término en inglés, *high-dimensional data*. Sin embargo, al considerar un número elevado de variables, su análisis no es simple y directo. Las interacciones entre ellas hacen aún más evidente la dificultad de visualizar el comportamiento de la muestra y por ello son necesarias técnicas estadísticas multivariantes. Esto ha

provocado en los últimos años la necesidad de un cambio teórico a gran escala pues, para el estudio de esta información, son necesarias técnicas capaces de manejar las relaciones entre todas las variables almacenadas. La definición de modelos matemáticos basados en datos de alta dimensionalidad, con el objetivo global de analizarlos de forma correcta, nunca ha sido más importante: se requiere una metodología lógica de análisis con técnicas capaces de manejar grandes cantidades de información, que sean a su vez capaces de resumir el conjunto de variables observadas en unas pocas nuevas variables hipotéticas, construidas como transformaciones de las originales, con la mínima pérdida de información. Dado el carácter multidimensional de los datos, la geometría, el álgebra matricial, el análisis numérico y la teoría de la probabilidad juegan un papel fundamental.

I1. Evolución de los métodos clásicos factoriales

Una de las ramas centrales de la metodología estadística multivariante se centra en la extracción de información relevante de entre una gran cantidad de datos, con el fin de detectar patrones, relaciones y diferencias entre los objetos y/o variables estudiadas. Se trata de las técnicas de reducción de la dimensión, entre las que destacan el Análisis Factorial (FA, por sus siglas en inglés) (Hotelling, 1933) y el Análisis de Componentes Principales (PCA, por sus siglas en inglés) (Jolliffe, 2002). Estos métodos tienen su origen en la búsqueda de patrones que permitan la proyección de modelos definidos en un hiperespacio a uno de menor dimensión, donde la interpretación sea factible. Basándose en las relaciones de las variables medidas y en la búsqueda de estructuras latentes subyacentes en la matriz de datos, estos mecanismos de descomposición (conocidos también como métodos de proyección) tratan de obtener un subespacio de menor dimensión donde representar la información inicial de manera más clara y práctica para su estudio con una pérdida mínima de información. Esta idea se basa en la existencia de un subespacio de menor dimensión subyacente a los datos debido al hecho de que la matriz no sea de rango completo (en otras palabras, las características medidas presentan

relaciones entre ellas). Mediante la proyección de los datos de entrada en dichas nuevas direcciones, se logra representar la mayor cantidad de información posible eliminando el ruido existente en los datos.

El descuido completo de las covarianzas existentes que realiza la estadística univariante y la dificultad de ajustar modelos en su presencia para algunas técnicas estadísticas, como la regresión, podría dar lugar a ignorar características importantes de las unidades muestrales (Bro y Smilde, 2014). Así, estas relaciones son aprovechadas por las técnicas factoriales para generar una visión global de la realidad, sin tener en cuenta el ruido que la defina.

A pesar de que el FA es una técnica factorial muy implementada en todas las ciencias sociales, como psicología y educación (Bandalos & Finney, 2018; Briz-Ponce & García-Peñalvo, 2015; Mellers et al., 2015), de entre las técnicas de reducción, la más empleada es el PCA por su diversidad y gran utilidad en disciplinas de muy diverso calibre (Alonso-Gutierrez et al., 2015; Andrés, Asongu, & Amavilah, 2015; Săndica, Dudian, & Ştefănescu, 2018; S. H. Wang et al., 2016; Zahedi & Rounaghi, 2015). El PCA trata de extraer la información más relevante de una matriz de datos y proyectarla en un espacio de menor dimensión al original, absorbiendo la mayor variabilidad posible de la información inicial (Jolliffe & Cadima, 2016). Se basa en el cálculo de un conjunto de nuevas variables (componentes principales, PCs) como combinación lineal de todas las variables originales, conocidas como componentes principales. La interpretación de las PCs radica justamente en los coeficientes de estas combinaciones lineales, que denotan la contribución de cada variable original a la formulación de la nueva variable latente. Habitualmente, la implementación del PCA se realiza a partir de un proceso de descomposición de la matriz original, como la descomposición en valores y vectores propios. Sin embargo, la formulación en el año 1936 de la Descomposición en Valores Singulares (SVD, por sus siglas en inglés) (Eckart & Young, 1936) supuso un gran avance en la estadística multivariante. A partir de entonces, la forma más habitual de implementar el PCA es derivarlo directamente de los resultados de la SVD, obteniendo cada una de las PCs de manera directa a partir de los vectores singulares de la SVD.

La SVD es la herramienta por excelencia del análisis estadístico multivariante (Abdi, 2007; Puntanen, 2011). Diversos campos de la ciencia hacen

uso de ella: clasificación de textos (Thara & Sidharth, 2017), genética (Franceschini, Lin, von Mering, & Jensen, 2016), neuroimagen (Juneja, Rana, & Agrawal, 2016) y política (Skillicorn & Leuprecht, 2015),… Se trata de una herramienta algebraica de descomposición matricial, que aproxima una matriz de datos X_{nxp} mediante el producto de tres matrices que contienen los vectores singulares y valores singulares de la matriz de datos. La SVD debe su gran potencial a que proporciona la mejor aproximación de bajo rango de una matriz de datos, en el sentido de los mínimos cuadrados (Björck, 2015).

I1.1.- Problemática de la metodología clásica

Como se decía anteriormente, el PCA es la técnica más utilizada de la estadística multivariante. Sin embargo, presenta una serie de inconvenientes que deben ser considerados. Como muchos otros métodos clásicos, el PCA puede derivarse de la SVD y, por ello, algunos de sus inconvenientes se deben al uso de esta descomposición. Cada PC se calcula a partir de los vectores singulares obtenidos en la SVD de la matriz de datos, y dado que estos se expresan en base a todas las variables originales, cada PC se calcula como una combinación de todas también. Esto es independiente de que alguna de las variables contribuya menos a su formación. En la práctica, la situación ideal sería aquella que llevase a la obtención de coeficientes exactamente nulos (coeficientes *sparse*, de ahora en adelante), de manera que la interpretación de las PCs solo dependiera de un subconjunto de las variables originales. Desafortunadamente, entendiendo las combinaciones lineales como las abstracciones matemáticas que son y debido a que las cargas en la práctica real suelen ser no nulas, no existe garantía de proporcionar un significado a estos conceptos matemáticos y, con ello, significado a las PCs. Esto dificulta la capacidad informativa de los datos y se genera el principal inconveniente del PCA: su interpretación. Más aún, en el caso de matrices de altas dimensiones donde el número de variables excede ampliamente el número de observaciones de la muestra. En los últimos años, la alta dimensionalidad de los datos se ha convertido en una característica común de múltiples disciplinas (Gligorijević, Malod-Dognin, & Pržulj, 2016; Sch, 2005). Por alta dimensionalidad se hace referencia a bases de datos formadas por un número de observaciones (muestras) mucho menor al número de variables consideradas ($I \ll J$), al

contrario de la ideología de los métodos clásicos, diseñados para matrices en que $I \gg J$.

En segundo lugar, las operaciones matemáticas realizadas en la SVD no respetan la estructura inicial de los datos. Ocasiones en las que las matrices de partida poseen una gran cantidad de coeficientes nulos[1] o incluso casos en los que es necesario mantener la estructura "positiva" de los datos, se ven afectadas por el descuido de la SVD de dichas propiedades.

En tercer lugar, como se mencionaba hace unas líneas, el PCA es inconsistente para datos de altas dimensiones en los que $J \gg I$, puesto que se agrava la dificultad de encontrar las nuevas direcciones con los métodos tradicionales cuando se trabaja con una gran cantidad de datos. El cálculo de la SVD, y de la matriz de covarianzas, supone un alto coste computacional y no es óptima en todas las ocasiones. En el caso de matrices de altas dimensiones, además, se subraya el problema de interpretación de las PCs, puesto que es de gran dificultad dar significado a variables formadas por una combinación de miles de características. Este es el caso, por ejemplo, de estudios genómicos, donde se recoge información de unos 10.000-20.000 genes simultáneamente.

En cuarto lugar, hay que tener en cuenta que el PCA se basa en relaciones lineales de los datos. Un coeficiente de correlación bajo o casi nulo entre las variables de estudio, no garantiza que sean independientes y no tengan otro tipo de asociación, pues el coeficiente de correlación sólo mide el grado de asociación lineal.

Esta problemática es arraigada por todas las metodologías implementadas mediante la SVD, algunas de ellas mencionadas en la sección anterior, como los métodos Biplot o el Análisis de Correspondencias. Las componentes obtenidas para la reducción de la dimensión en el caso de los métodos Biplots se obtienen como en el PCA, combinación de todas las variables de partida. Este hecho dificulta la interpretación de los resultados obtenidos, pues a pesar de poder interpretar individuos y variables, y las relaciones entre

[1] Las matrices con una gran cantidad de coeficientes nulos son conocidas como matrices *sparse*. Debe tenerse en cuenta que esta definición no es la que se dará aquí al hablar de métodos *sparse*.

ellos, los ejes sobre los que se realiza la proyección del subespacio de dimensión reducida son combinación lineal de todas las variables originales.

Problemática de las bases de datos de altas dimensiones.

A *nivel estadístico* las bases de datos de altas dimensiones suponen un reto en lo referente a las estimaciones de los parámetros de un modelo, por los denominados grados de libertad. Al pensar en un caso trivial, como por ejemplo un modelo simple en el que se quiera explicar la relación entre dos variables relacionadas de manera lineal, la recta de regresión empleada para explicar su relación viene dada por la estimación previa de dos parámetros (coeficiente de regresión y ordenada en el origen) calculados mediante el método de los mínimos cuadrados ordinario. Ahora bien, dicho método necesita al menos i puntos para la estimación de i parámetros. En este caso, para calcular los parámetros del modelo bidimensional bastaría con conocer dos puntos u observaciones. Sin embargo, si solo se dispusiera de datos de una observación, no sería posible resolver y estimar ambos parámetros puesto que no sería posible conocer cuál de todas las factibles es la recta de mejor ajuste, pues todas minimizarían la suma de errores al cuadrado. Algo similar ocurre con los datos de altas dimensiones y las técnicas clásicas de la estadística; para estimar un modelo de 1.000 parámetros, se necesitan datos de al menos 1.000 observaciones. Desafortunadamente, no en todas las disciplinas es posible obtener tantos datos, porque es excesivamente caro y costoso en tiempo. ¿Qué hacemos si necesitamos estimar un modelo de 1.000 parámetros, pero sólo se dispone de 200 observaciones? La respuesta está en los métodos de selección de variables, como paso previo a la posterior implementación de una técnica clásica de análisis.

I1.2.- Propuestas de mejora: selección de variables

A lo largo de los años, han surgido distintos enfoques para mejorar la interpretación de las variables latentes, mediante la generación de coeficentes sparse (nulas). En 1995, Jolliffe propuso el que quizá sea el enfoque más antiguo para solventar este problema: los métodos de rotación (Jolliffe, 1995). Esto ya había sido propuesto anteriormente por Thurstone (1935) para el Análisis Factorial, con la finalidad de que los nuevos ejes siguiesen el conocido como

principio de estructura simple. Los métodos de rotación, como la rotación Varimax (Kaiser, 1958) o la rotación oblicua, tratan de girar los nuevos ejes, obtenidos tras la reducción de dimensionalidad, hasta conseguir aproximarlos lo máximo posible a las variables que cargan en ellos. Sin embargo, estos métodos simplifican la matriz de cargas, pero no producen coeficientes exactamente nulos. Es por este último motivo que, también en 1995, Cadima y Jolliffe (1995) consideraron la idea propuesta por Jeffers (1967) acerca del uso de la umbralización. Esta consistía en ignorar aquellos coeficientes cuya magnitud estuviera por debajo de un umbral establecido y suponerlos como nulos de manera artificial. Actualmente, la umbralización es una de las técnicas más utilizadas en la práctica; sin embargo, puede dar lugar a resultados engañosos (Trendafilov, 2014). Vines (2000) y más adelante Anaya-Izquierdo, Critchley y Vines (2011), siguiendo las ideas de Hausmann (1982), plantean la opción de restringir el valor de las cargas, de manera que adquieran un valor de entre un cierto conjunto de enteros, como por ejemplo $\{-1,0,1\}$. Sin embargo, esta metodología proporciona la misma importancia a las variables que contribuyen a la formación de los ejes independientemente de que unas aporten en su formación más que otras (Anaya-Izquierdo, Critchley, & Vines, 2011; Vines, 2000).

Con los avances en tecnología y la aparición de bases de datos de altas dimensiones, los enfoques más modernos se centran en los métodos de regularización, que incluyen penalizaciones que promueven la anulación de cargas (*sparsing-promoting penalties*) a la formulación del problema de optimización (Hastie, Tibshirani & Wainwright, 2015). Los métodos de regularización han ganado popularidad debido a su control de sobreajuste en la estimación de parámetros del modelo, así como a la selección de variables (Huang, Liu y Liang, 2016; Liu et al., 2019). Su motivación principal es lograr modelos mejor adaptados e interpretables y que no sean inconsistentes para matrices de grandes dimensiones (problemática de la metodología clásica).

Para ello, los métodos de regularización o penalización introducen una restricción sobre la norma de un vector en el problema de optimización de la función de pérdida para lograr que algunos de sus coeficientes sean nulos.

El método de penalización más utilizado es el operador Lasso (*Least Absolute Shrinkage and Selection Operator*) (Tibshirani, 1996). Este penaliza la norma L1 de un vector, restringiendo la suma de los valores absolutos del vector considerado. Lasso realiza una selección de variables automática, al producir coeficientes exactamente nulos. Debido a esto, se ha convertido en una herramienta muy útil en el análisis de datos de altas dimensiones, donde la identificación **automática** de variables importantes en un modelo es uno de los principales propósitos (Wong, Rostomily, & Wong, 2019). A pesar de sus muchas ventajas, Lasso presenta algunos inconvenientes que han sido considerados, dando lugar al desarrollo matemático de otros tipos de regularizaciones. Primero, Lasso no cumple ser un procedimiento Oracle (Fan & Li, 2001); en otras palabras, Lasso realiza una selección de variables inconsistente porque no selecciona el conjunto correcto de variables con una probabilidad que converja a 1, permitiendo que aparezcan características redundantes en el modelo estimado (el lector puede encontrar más información en capítulos posteriores). Por eso, en el año 2006 Zou presenta *adaptive* Lasso (Zou, 2006), una versión ponderada de Lasso que sí verifica ser un procedimiento *oracle*. Por otro lado, la esencia de los métodos multivariantes radica en aprovechar la relación entre las variables para explicar los patrones en los datos. En este sentido, si hay un grupo de variables correlacionadas, Lasso tiende a seleccionar una variable del grupo. Esto supone una inconsistencia práctica en diversas disciplinas, como en el análisis de expresión génica de microarrays, donde es importante tener en cuenta la actividad conjunta de los genes en múltiples mecanismos biológicos (Hore et al., 2016; Wang, Yuan, & Montana, 2015) o en el análisis de las propiedades psicométricas de cuestionarios, donde cada constructo latente está conformado por la relación de un conjunto de ítems (Barahona, García, Sánchez-García, Barba, & Galindo-Villardón, 2018; Vega-Hernández, Patino-Alonso, & Galindo-Villardón, 2018). Por ello, para solventar este defecto Zou y Hastie (2005) proponen Elastic net (enet, L1+L2), una combinación de la penalización L1 (Lasso) y L2 (Ridge) que se centra en el efecto de agrupación, permitiendo que las variables relacionadas aparezcan juntas en el modelo sparse. En el año 2009 aparece *adaptive* Elastic net (Zou & Zhang, 2009), una combinación de las penalizaciones *adaptive* Lasso y Ridge, que da lugar a un procedimiento *oracle*

que permite la selección de varias variables de entre un conjunto de relacionadas.

La inclusión de este tipo de penalizaciones en el PCA ha dado lugar a una diversa variedad de técnicas de Análisis de Componentes Principales Sparse (Sparse Principal Component Analysis, Sparse PCA), convirtiéndose en la primera técnica de reducción de la dimensión y selección de variables automática, simultáneamente (Li, Tian, & Liu, 2016; Trendafilov, 2014). La revisión bibliográfica genera múltiples trabajos con un objetivo común: generar vectores de proyección en un subespacio de dimensión menor con parte de sus cargas nulas. Al igual que su antecesor PCA, el Sparse PCA cuenta con distintas formulaciones del problema: maximización de la varianza, minimización del error y enfoque probabilístico. Dentro de estos enfoques, principalmente, existen dos formas de lograr la nulidad en las cargas: mediante la restricción de las cargas o a través de su contracción. Desde el punto de vista de maximización del error, SCoTLASS es el primer método de Sparse PCA conocido (Jolliffe, Trendafilov, & Uddin, 2003). Posteriormente, d'Aspremont, Ghaoui, Jordan y Lanckriet (2007) proponen una relajación convexa del problema inicial, haciendo que el problema sea computacionalmente factible. Dentro de los métodos de minimización del error que tratan de contraer los vectores de cargas, el principal es el SPCA (Zou, Hastie, & Tibshirani, 2006) y el *sparse PCA via regularized SVD* (sPCA-rSVD) (Shen & Huang, 2008), que utiliza los conceptos de la selección de variables en regresión. Formulan el problema del SPCA como un problema de optimización del tipo regresión penalizada con Elastic net (Zou & Hastie, 2005).

Estrechamente relacionado con esto, otra de las ramas de investigación actual centra su atención en el desarrollo de técnicas de selección automática de características relevantes desde el punto de vista de los métodos de factorización matricial. La literatura recoge diferentes alternativas, como la descomposición matricial penalizada (PMD) (Witten, Tibshirani, & Hastie, 2009), la descomposición CUR (Mahoney & Drineas, 2009) o la factorización matricial no negativa (Lee & Seung, 1999) o incluso la mezcla metodológica de varias de estas técnicas, como la descomposición en valores singulares sparse (Lee, Shen, Huang, & Marron, 2010) o la factorización matricial no negativa sparse (Kim & Park, 2008; Ye & Jin, 2013). Centrándose en las propiedades óptimas de

ortonormalidad de las matrices de la SVD, las técnicas mencionadas anteriormente presentan propiedades óptimas en cuanto a no-negatividad y coeficientes sparse, en detrimento de la ortogonalidad de las variables generadas. Esto provoca que su extensión a otras técnicas de análisis de datos de dos vías se dificulte. Por este motivo, recientemente a comienzos del año 2019, Guillemot et al. (2019) proponen la SVD restringida (*constrained* SVD, CSVD): un método que integra simultáneamente resultados sparse haciendo uso de la penalización Lasso y ortogonalidad. Este técnica será de gran interés de ahora en adelante en lo que concierne a la contribución teórica de este trabajo.

Al igual que el PCA ha sido modificado para lograr PCs sparse, otros tipos de penalizaciones han sido consideradas recientemente en la literatura. Aunque no se entrará en detalle aquí, ya han sido puestas en práctica penalizaciones de no negatividad, cuyo objetivo común es mantener la estructura inicial positiva de la matriz de datos. Una revisión más exhaustiva de estos métodos puede verse en (Cichocki, Zdunek, Phan, & Amari, 2009).

Además de las penalizaciones de dispersión (*sparsity*) y no negatividad, han sido puestas en práctica las penalizaciones de suavizado de los datos. Su finalidad es suavizar y mantener la estructura no lineal de los datos medidos. Estas penalizaciones aparecen en el ámbito del Análisis de Componentes Principales Funcional (*Functional Principal Component Analysis, FPCA)* (Ramsay & Silverman, 2005). El FPCA es una técnica basada en la idea de que la correlación entre las variables de partida no tiene por qué ser lineal, uno de los problemas del PCA clásico que considera relaciones lineales entre ellas. Plantea el cálculo de las PCs funcionales a partir de una base de funciones cuadráticas o *splines*, de forma que recojan correctamente el tipo de relación de las variables originales y reproduzcan adecuadamente los datos.

La reducción de la dimensionalidad de un problema es uno de los dos objetivos primordiales al tratar con estructuras de datos multivariantes. El segundo de ellos se centra en el reconocimiento de patrones y la clasificación de observaciones con comportamientos similares. En este sentido, la literatura recoge técnicas de biclustering, cuyo fin es clasificar objetos a la vez que mejorar la interpretación de los ejes factoriales atendiendo a distintas definiciones (Hartigan, 1972; 1975; Li, 2005). Algunas de las técnicas más actuales de

biclustering, que se acercan a la teoría de los métodos Sparse, son el *Sparse K-Means Clustering* propuesto por Witten y Tibshirani en el año 2010 (Witten & Tibshirani, 2010) y el *Clustering and Disjoint Principal Component Analysis* (CDPCA) de Vichi y Saporta (2009), que tratan de detectar las variables más influyentes en el modelo paralelamente a la búsqueda de la clasificación de los objetos. Cuando el objetivo es representar observaciones y variables en un mismo sistema de referencia donde las relaciones entre ellos sean visualmente interpretables de manera que puedan determinarse subgrupos de observaciones con patrones similares, lo adecuado es implementar un método de representación Biplot (Gabriel, 1971; Galindo, 1986). En el caso de métodos Biplots sparse, la literatura tan solo recopila dos técnicas cuyo objetivo sea la producción de ejes factoriales sparse: para producir cargas nulas: i) *Clustering Disjoint Biplot* (CDBiplot), que combina los métodos factoriales de reducción de la dimensión y los métodos de clasificación, produciendo ejes factoriales disjuntos computados bajo la restricción de que cada variable original tan solo puede contribuir a la formación de un eje factorial y ii) Elastic net HJ-Biplot (Cubilla-Montilla, Galindo-Villardón, Nieto-Librero, Vicente, & García-Sánchez, 2019), que produce componentes sparse incorporando las penalizaciones Ridge, Lasso y Elastic net a las componentes principales del HJ-Biplot. Ambas metodologías presentan la desventaja de que las componentes principales generadas comparten información entre ellas, al no ser ortogonales.

I2. Extensión al análisis multivía

La forma en la que se representen los datos es fundamental para extraer la información latente subyacente. Habitualmente, la información recogida de muestras de datos queda almacenada en matrices de dos vías, descritas por modelos lineales. El resultado recogido en dichas matrices de datos representa la medición de variables relacionadas entre sí, o de factores latentes subyacentes que facilitan el análisis del comportamiento de una muestra. Se ha recalcado hasta ahora la importancia de la reducción de la dimensión de los datos y selección de variables características.

Aunque los métodos clásicos del análisis multivariante trabajan con matrices de dos vías, en las últimas décadas ha surgido la necesidad de manipular datos

descritos en múltiples dimensiones, lo que se conoce como matrices multidimensionales/multivía o tensores. La integración de múltiples matrices de datos de diferentes fuentes se ha vuelto cada vez más popular (Kolda & Bader, 2009). En el campo de la investigación sobre cáncer, por ejemplo, ha implicado una mejora de las condiciones de estudio que permiten explicar diferentes asociaciones biológicas de tipos de tumores (Hore et al., 2016). Con la ayuda de dicho análisis, los investigadores clínicos pueden optimizar sus recursos para generar nuevos conocimientos que mejoren la prevención, el diagnóstico y el tratamiento de este tipo de enfermedades y de otras. Desde el punto de vista analítico, la necesidad de manipular los datos descritos en múltiples dimensiones ha dado como resultado un gran marco de técnicas factorización y descomposición de datos de múltiples vías (Kroonenberg, 2008). Estas deben poder integrar información de diferentes plataformas o disciplinas (proteómica, metabolómica, bioquímica ...) como en el trabajo de Hijazi y Chan (2013) o incluso desde una única plataforma en diferentes condiciones de tiempo. Como se decía anteriormente, en lo referente a su análisis estadístico, algunas técnicas tradicionales de dos vías podrían trabajar con cada una de las matrices por separado, para la unificación y comparación de sus resultados posteriormente. Sin embargo, en un contexto de bases de datos múltiples esto sería ineficiente al no tener en cuenta las posibles asociaciones entre las diferentes matrices de datos. Es por este motivo por el que surgen técnicas de análisis de datos dispuestos en múltiples vías con el fin de identificar las relaciones latentes que explican las diferencias y/o similitudes entre los diferentes escenarios planteados (Smilde et al., 2017).

En cuanto a su definición general, las matrices multivía o los tensores no son más que la generalización multilineal de matrices y vectores a espacios de mayor dimensión, donde los datos se organizan en tres o más direcciones. El reemplazo de estos tensores por una aproximación de menor dimensión permite observar estructuras que a priori no podrían ser observadas, al igual que ocurría en el caso bidimensional. La factorización de matrices y tensores, la Descomposición en Valores Singulares, el Análisis de Componentes Principales y los métodos de análisis de matrices de tres vías, como los modelos Tucker o STATIS, son en la actualidad los mecanismos de descomposición lineal y multilineal de bases de

datos que permiten representar la información de una manera más práctica para su estudio. Se tomará como referencia en este proyecto el análisis de datos de tres vías: conjunto de datos que se representa en un bloque tridimensional (individuos, variables y condiciones).

Las investigaciones sobre este tipo de datos muestran diferentes métodos de análisis. Los trabajos de integración de matrices, cuyo cimiento teórico tiene origen en la SVD o PCA clásicos de dos vías, quedan recogidos en dos vertientes fundamentales: los métodos franceses y los anglosajones. Estos métodos también son conocidos como métodos asimétricos (franceses) y simétricos (anglosajones). Los métodos franceses se centran en el análisis de la matriz *unfolding;* es decir, de la matriz desconcatenada. De esta forma, este tipo de métodos suponen una extensión inmediata de las técnicas de dos vías. En este sentido, la literatura incluye métodos como el análisis factorial múltiple (MFA, por sus siglas en inglés) (Escofier & Pagès, 1983) y los métodos de la familia STATIS de L'Hermier des Plantes (1976). El lector puede encontrar una revisión fundamental y excelente de los métodos de la familia STATIS en (Abdi, Williams, Valentin, & Bennani-Dosse, 2012). Sin embargo, la dificultad de algunos de estos métodos para distinguir la variación sistemática del ruido e identificar la variación compartida versus específica entre matrices, ha dado lugar a nuevos avances en la investigación. En este sentido, atendiendo a otro tipo de factorizaciones diferentes a la SVD, puede mencionarse la factorización matricial no negativa (NMF, por sus siglas en inglés) (Lee & Seung, 1999) y sus extensiones multivía como la NMF conjunta (*Joint Non-negative Matrix Factorization* (jNMF)) (Zhang et al., 2012) o la factorización matricial no negativa que integra Ortogonalidad y regularización *Integrative Orthogonality-regularized Nonnegative Matrix Factorization* (ioNMF), propuesta por Strazar, Zitnik, Zupan, Ule, & Curk (2016). Cuando el interés radica en detectar la variabilidad compartida entre matrices de diferentes fuentes, así como la variación específica de cada uno de ellos y la variabilidad residual no identificada, cabe destacar el análisis de Componentes Simultáneo con rotación para la búsqueda de componentes comunes y distintas (*Simultaneous Component Analysis with rotation to Common and Distinctive Components*, DISCO-SCA) y la descomposición en valores singulares generalizada (GSVD) (Katrijn van Deun et al., 2012) o el método de variabilidad

explicada individual y conjunta JIVE (*Joint and Invidivual Variation Explained*) (Lock, Hoadley, Marron, & Nobel, 2013). JIVE se define como la extensión PCA para múltiples matrices de datos como una descomposición de datos de múltiples vías, teniendo en cuenta estos desde el punto de vista de la familia francesa; esto es, desde el desdoblamiento de matrices. Esta técnica reciente ya ha sido aplicada eficazmente en diferentes contextos, como las redes ferroviarias en las ciudades modernas (Jere et al., 2014), en el campo ómico para caracterizar el cáncer de pulmón (Sandri et al., 2018), en análisis de supervivencia de glioblastomas multiformes, uno de los tumores cerebrales más agresivos por sus altas tasas de mortalidad (Kaplan & Lock, 2017) o incluso en casos de autismo infantil (Chawarska, Ye, Shic & Chen, 2016).

De entre los métodos anglosajones, caracterizados por ajustar modelos que reproduzcan lo mejor posible los datos originales, destacan los modelos de Tucker (Tucker, 1966; 1972), el método PARAFAC-CANDECOMP (Harshman, 1970 ; Carrol & Chang, 1970) y los métodos Tuckals (Kroonenberg & Leeuw, 1980). Estos métodos han sido exitosamente aplicados en el análisis de datos multidimensionales (Cichocki et al., 2009). Su principal diferencia con los métodos franceses es que son técnicas de descomposición/factorización de tensores, mientras que los métodos anglosajones trabajan con técnicas de descomposición de matrices de dos vías sobre la matriz desdoblada. Esto último supone una desventaja pues se pierde uno de los modos de información de la matriz tridimensional. Por otro lado, los métodos de descomposición de tensores tienen su cimiento teórico en el PCA y en la SVD; tanto es así que los modelos Tucker son conocidos también como el Análisis de Componentes Principales Generalizado y el modelo PARAFAC como la Descomposición en Valores Singulares Generalizada.

Al igual que ocurría en el Análisis de Matrices, la problemática de interpretación y no selección de variables relevantes de la SVD y, en consecuencia, del PCA, es arraigado al caso multidimensional. Los métodos de descomposición multivía no proporcionan una estructura de soluciones sparse, ni en el caso de los métodos anglosajones ni franceses. En la práctica, estas son deseables y necesarias cuando existe una gran cantidad de variables en estudio y las componentes subyacentes tienen interpretación, pero también en pequeñas

bases de datos para facilitar el análisis de los resultados. El desarrollo de la teoría y las aplicaciones de las descomposiciones multilineales sparse en este ámbito apenas acaba de comenzar y nuestro objetivo es producir un método multivía que unifique los métodos sparse de dos dimensiones y los principales métodos de descomposición/factorización del análisis multivía. En todo el proceso, habrá que tener en cuenta que la generalización a una dimensión mayor no es trivial, al igual que ocurrió con la generalización del PCA clásico a múltiples dimensiones, y la no unicidad en la solución de la SVD Generalizada.

I3. Justificación de la investigación

Una de las áreas más importantes de la investigación actual en el análisis de datos multivariantes se centra en el desarrollo de técnicas eficientes para el estudio de grandes matrices de datos (*high dimensional data*). En disciplinas como la genética o el procesamiento de imágenes, las bases de datos están formadas por miles de variables. Para su análisis, se requieren técnicas que las simplifiquen y que no pierdan la información clave de la muestra. El PCA, formulado a partir de la SVD, es la técnica más implementada para la reducción de la dimensión de matrices de datos y extracción de variables características, lo que se logra a través de la extracción de nuevas variables latentes conocidas. Sin embargo, el PCA presenta el inconveniente de que cada componente principal es una combinación lineal de todas las variables originales y esto dificulta su interpretación. A lo largo de los años se han desarrollado distintas metodologías para paliar su principal desventaja, pero es en la última década en la que este método se ha modificado para producir componentes principales sparse; es decir, componentes que envuelvan únicamente un pequeño subconjunto de las variables originales más importantes. Todo ello ha dado lugar a la aparición de los métodos de penalización o regularización, técnicas de selección automática de variables características extremadamente útiles en las aplicaciones modernas donde el número de variables originales es muy elevado y, por ello, la necesidad de simplificar la interpretación de los resultados es imprescindible.

Este proyecto investiga y propone una nueva herramienta de análisis aplicable a los datos multivía o denominados en minería de datos, tensores.

Hasta ahora, los estudios recogían la información en matrices bidimensionales, pero en la actualidad existen ocasiones en las que es interesante y necesario englobar dicha información en bloques de más vías, incluyéndose más cantidad de información en el estudio. Estos métodos, como los modelos Tucker o el método STATIS, basan su fundamento teórico en el PCA clásico, cuya deficiencia es bien conocida, y en la Descomposición en Valores Singulares (SVD), que no tiene una definición única en el caso de tensores. Nuestro trabajo se enfoca en una línea de investigación reciente en la literatura: el desarrollo de un método sparse generalizado adaptado al análisis de datos multidimensionales. Gracias a las herramientas matemáticas, se abre así un nuevo camino en la estadística multivariante, con importantes aplicaciones prácticas en cualquier área de la actividad humana.

OBJETIVOS Y METODOLOGÍA

Objetivo principal

El propósito general de este trabajo consiste en la propuesta de una metodología sparse para el análisis de datos multivía, a partir de la generalización de las técnicas penalizadas.

Objetivos secundarios

Los objetivos directrices de esta investigación son:

(1) Realizar una revisión bibliográfica crítica y evaluar los principales métodos de penalización para datos bidimensionales, para su posible generalización a tensores o datos multidimensionales.
(2) Evidenciar las dificultades en el análisis de matrices con las técnicas clásicas, tanto bidimensionales como multidimensionales y mostrar la utilidad de las técnicas de selección de variables, como Sparse PCA y descomposición CUR; así como la extensión de su uso a nuevos campos, como áreas del ámbito social.
(3) Desarrollar un modelo sparse de dos vías generalizable y proponer una metodología para el cálculo de los parámetros de regularización del modelo.
(4) Hacer una revisión de los métodos para el análisis de tablas de tres vías penalizados, con énfasis principalmente sobre los modelos Tucker.
(5) Proponer un nuevo modelo para el análisis sparse de datos de tres vías.
(6) Implementar el software necesario en R para facilitar el uso de las metodologías propuestas.

Material y Métodos

1.-Bases de datos

Las técnicas que se describen a lo largo de este trabajo son compatibles con bases de datos de disciplinas muy diversas, como pueden ser educación, psicología o economía. Ahora bien, el uso de técnicas automáticas de selección de variables tiene especial interés en el ámbito del análisis de conjuntos de datos de altas dimensiones. Entre otros campos, la genómica es una disciplina que envuelve la información de un conjunto de genes trabajando de manera conjunta como ejes principales de todo el mecanismo regulador del cuerpo humano. Por todo ello, las diversas metodologías que se presentan a lo largo de este trabajo tendrán un sentido doble:

- **Análisis exploratorio psicométrico de la validez factorial de cuestionarios**. Mostrar su utilidad en disciplinas como psicología y educación donde se convierten en herramientas útiles para simplificar la interpretación de las técnicas clásicas o incluso pueden superponerse a situaciones poco habituales donde las técnicas clásicas fracasan estrepitosamente. Estas contribuciones prácticas se podrán encontrar en los capítulos 1 y 2.
- **Clasificación de tumores por medio del análisis de expresión génica**. Recientemente, el análisis de datos biológicos ha sufrido un cambio de paradigma por la posibilidad de analizar, recoger y almacenar grandes cantidades de datos cruciales para entender los mecanismos de los códigos genéticos. El cambio de paradigma que ha dado lugar a las bases de datos de altas dimensiones en esta disciplina ha marcado un punto clave a nivel biológico. En el proceso de clasificación de tumores, el estudio de microarrays génicos que recogen la actividad de miles de genes permite investigar los vínculos existentes entre el desarrollo de la enfermedad, malignidad, supervivencia, factores pronósticos, … con la expresión de los genes. Esto es útil en el diagnóstico de casos, pues una gran cantidad de casos no presentan un diagnóstico ni pronóstico claro.

A continuación se hace una breve reseña sobre ambas aplicaciones.

Ámbito social: psicología y educación.

En disciplinas como la psicología o educación las variables de interés son constructos latentes no tangibles, medidos a través de cuestionarios. La capacidad de los instrumentos de medida para cuantificar los rasgos latentes para los cuales han sido diseñados debe ser evaluada a través de sus propiedades psicométricas: fiabilidad y validez, con el fin de que las puntuaciones factoriales obtenidas tengan sentido en sí mismas. La forma clásica de estudiar la validez factorial de un constructo latente es utilizar un Análisis Factorial Exploratorio/Confirmatorio, en la mayoría de los casos, en su versión Análisis de Componentes Principales. Ahora bien, dichas técnicas surgen con el propósito de captar la variabilidad latente en los datos y explicar así la existencia de los rasgos evaluados. En otras palabras, este tipo de procedimientos es muy útil cuando las dimensiones latentes presentan alta variabilidad o cuando el número de variables observables no es muy alto. Sin embargo, fracasan estrepitosamente cuando los datos presentan efecto techo/suelo, o cuando el número de variables observables es muy alto, poniendo en entredicho la validez factorial de la escala e impidiendo la inferencia sobre las puntuaciones de los sujetos.

En este sentido se contribuye al nivel práctico de análisis factorial de cuestionarios en situaciones favorables y desfavorables para las técnicas clásicas.

1- **Actitud hacia la didáctica y enfoques de aprendizaje.** Presentamos este estudio, con el principal objetivo de encontrar el punto de enlace entre los distintos estilos de aprendizaje del alumnado universitario y su actitud a la hora de abordar el estudio de la materia, puesto que cuando su predisposición es favorable, estos estarán motivados a realizar esfuerzos fuertes para aprender y afianzar sus conocimientos. La muestra analizada está formada por 146 estudiantes, mayoritariamente mujeres, del grado de Educación Social de la Universidad de Salamanca, entre los años 2011 y 2013, que cursaron la asignatura de Didáctica. Se recogió la respuesta de la muestra a dos cuestionarios diferentes. Por un lado, el cuestionario

de actitud hacia la didáctica, que sirve para valorar las componentes afectivas y valorativas de los estudiantes, con una estructura factorial de cuatro dimensiones: interés, ansiedad, utilidad profesional y utilidad presente. Por otro lado, se utilizó el cuestionario R-SPQ-2F para examinar los enfoques de aprendizaje del alumnado universitario, estructurado con respecto a dos factores latentes: aprendizaje profundo y superficial.

2- **Burnout.** El burnout se define como un síndrome psicológico asociado al acelerado nivel de vida de las personas, que está relacionado con el agotamiento personal, la realización personal y la despersonalización que sienten las personas en sí mismas. En esta investigación se plantea el análisis de la estructura factorial del cuestionario MBI (*Maslach Burnout Inventory*) en un colectivo de farmacéuticos. Para ello se recogió la respuesta al cuestionario por un grupo de 51 farmacéuticos de Castilla y León que desarrollaban su empleo en una farmacia.

Ámbito genómico. Bases de datos de alta dimensionalidad.

Conceptos preliminares. Todas las funciones biológicas llevadas a cabo en una célula están gobernadas por las proteínas. La célula es capaz de producir todas las proteínas que necesita a partir del ADN. El ADN es una secuencia ordenada y altamente regulada que almacena toda la información genética. Esta cadena de gran tamaño se compacta formando unas estructuras denominadas cromosomas. Aproximadamente un 2% de toda esta secuencia se compone de regiones limitadas denominadas genes. Un gen es la unidad básica funcional del ADN, es decir, cada gen tiene información para codificar una proteína. El proceso de producción de proteínas implementado por la célula se compone de dos pasos. En primer lugar, debe producirse la transcripción que se refiere a la conversión de ADN en ARN mensajero (ARNm). Este proceso es fundamental ya que la maquinaria efectora de proteínas sólo es capaz de leer secuencias de ARNm. El segundo proceso, o traducción, está llevado a cabo por los ribosomas y culmina en la síntesis de proteínas a partir del ARNm.

Importancia de la regulación de la expresión génica. La expresión génica es una medida de la cantidad de ARN transcrito a partir de un gen. Por ello, se estudia el comportamiento de esta medida como referencia de los niveles de

variación experimentados en el proceso de transcripción de un gen. Cada vez se detectan más genes implicados en distintas patologías, tanto en su desarrollo como en su evolución. Concretamente, las alteraciones en la expresión génica pueden ser diferenciales para algunas de dichas enfermedades; por ejemplo, en los últimos años han ganado una importancia crucial en la clasificación tumoral (Algamal & Lee, 2015).

Estudio de la expresión de los genes: *microarrays*. La biología molecular dispone de diversas técnicas para medir los niveles de ARN o ADN. Algunas de ellas caracterizadas por la cantidad de mediciones simultáneas que pueden realizar, como es el caso de los *microarrays* (Wit & McClure, 2004). Un *microarray* es una placa de un número variable de pocillos, donde en cada uno de ellos se adhieren pequeñas cantidades de ADN, denominadas sondas, que pueden unirse de forma complementaria al ARNm procedente de diversos genes. En cada pocillo hay secuencias de ADN diseñadas para unirse al ARNm originado a partir de un único gen de modo que al incorporarse el ARNm total de la célula (de todos los genes) sólo quedarán unidos aquellos fragmentos propios del gen en cuestión, produciendo una señal luminosa. Estas señales son almacenadas en imágenes digitales que son resumidas en ficheros de tipo CEL por el software de los fabricantes. Posteriormente, los archivos CDF son los encargados de traducir los pixeles de la imagen obtenida en archivos de expresión que nos reportarán una medida de expresión para cada gen.

Estructura de los datos ómicos. Aunque existen distintos tipos de datos ómicos, en este caso se trabajará sobre datos genómicos que recogen la información genética de los humanos. La estructura de este tipo de datos sigue el siguiente patrón: un gran número de características colocadas en filas y un número pequeño de muestras colocadas en columnas. Las matrices de expresión con las que se trabajará en las distintas aplicaciones contendrán en su interior información de una medida numérica de la expresión génica de cada gen en cada muestra recogida. Una vez obtenidos los datos de la matriz de expresión, previo a su análisis, debe realizarse un proceso de normalización de estos, pues los datos brutos no son inicialmente comparables. Para que los valores de expresión génica de los *microarrays* sean comparables entre sí es necesario eliminar el efecto de los errores sistemáticos acumulados debidos a la

fluorescencia, impresión y experimento biológico. Es decir, es necesario eliminar diferencias sistemáticas observadas en los niveles de expresión, existentes por el procesamiento experimental de los *microarrays*, a la vez que preservar la variabilidad biológica que posteriormente será analizada mediante técnicas multivariantes. Este proceso de normalización consta de tres pasos: i) corrección de fondo (estimar y eliminar la intensidad del ruido de fondo); ii) normalización (conseguir una misma variación para todas las sondas) y iii) sumarización (conversión de las sondas a información de expresión de genes). En todos los casos, los datos fueron normalizados de acuerdo al proceso RMA (*Robust Multi-array Average*). Mediante este método se realiza una corrección de fondo, normalización y sumarización de los datos para obtener una matriz numérica de intensidades de expresión génica logarítmicas. El objetivo de este trabajo no es entrar a valorar estos métodos. Por ello, se dirige al lector para obtener más información a (Bolstad, Irizarry, Åstrand, & Speed, 2003; Irizarry et al., 2003). Una vez corregida la variabilidad experimental mediante este tipo de procesos, se obtendrá la matriz de expresión: matriz numérica con señales de genes/sondas (en filas) en cada una de las muestras (colocadas en columnas). A partir de ahora, el objetivo es descubrir niveles de expresión diferenciales de genes en distintas muestras.

Software. En este estudio se hace uso del software libre R para el análisis de los datos y del repositorio Bioconductor, proyecto de desarrollo de programación libre destinado al análisis de datos genéticos. Una guía referente para el análisis estadístico de datos ómicos mediante R y Bioconductor puede encontrarse en (Ayala, 2018).

Aplicaciones. Una de las potencialidades de los datos genómicos es que deben ser públicos, con el fin de que toda la comunidad científica pueda realizar aportes significativos. Todos los datos de expresión de RNAm de las muestras de tumores en las aplicaciones que se mostrarán a continuación fueron obtenidas del repositorio funcional genómico público *Gene Expression Omnibus* (GEO, https://www.ncbi.nlm.nih.gov/geo/), del centro nacional de información biotecnológica (NCBI). A continuación se muestran las tres bases de datos genómicas analizadas en este estudio, empleadas para mostrar la utilidad de las técnicas de selección de variables multivariantes en la detección de patrones en

datos de altas dimensiones. Las dos primeras aplicaciones serán utilizadas para observar la utilidad de las técnicas de selección ya existentes en la literatura. La tercera será utilizada para la puesta en práctica de las contribuciones metodológicas realizadas en este documento.

1- Tumores cerebrales benignos: meningiomas.

Los meningiomas son tumores desarrollados en las meninges (Figura 1). En este caso, se analiza la información genética de 38 muestras de pacientes con meningiomas, con el objetivo de establecer perfiles de expresión genética asociados a la aparición de recidivas. La información de las muestras es la referente a la serie GSE43290 extraída de la plataforma GPL96 (HG-U133A Affymetrix Human Genome U133A Array) de la base de datos GEO. La matriz de expresión genética ha sido obtenida mediante la plataforma Affymetrix, almacenando información de expresión genética, en un total de 22283 sondas genéticas en las 38 muestras analizadas, 11 de las cuales han sufrido aparición de recidivas de grado I, II y III (Figura 2).

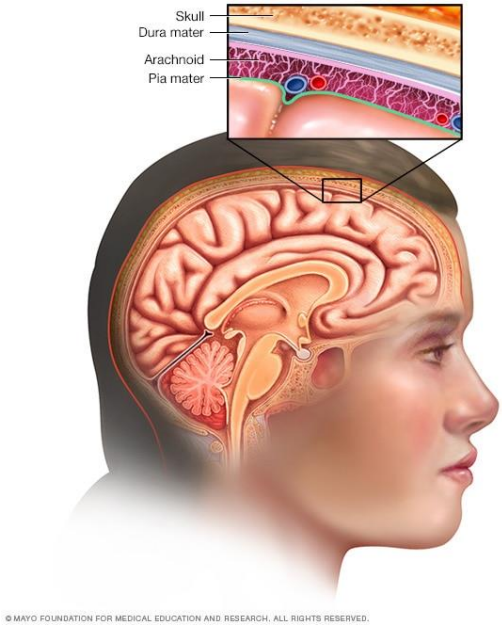


Figura 1. *Meninges. Fuente: Mayo Foundation for Medical Education and Research*

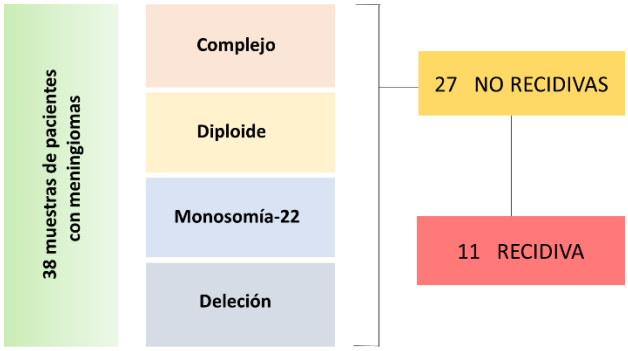


Figura 2. *Muestra de pacientes con meningiomas*

2- **Tumores cerebrales malignos: gliomas astrocíticos difusos.**

Los gliomas son los tumores cerebrales más comunes. En concreto, la línea astrocítica, desarrollada de células gliales, representa el 80% de las neoplasias cerebrales diagnosticadas. La relevancia de la patología es tal que se han generado volúmenes de información inmensos y que ha desencadenado su clasificación en múltiples sentidos: a nivel histológico, por su localización, supervivencia, … dando lugar a una idea de su gran heterogeneidad. La clasificación de los tumores astrocíticos más utilizada es la proporcionada, a nivel histológico, por la Organización Mundial de la Salud (WHO, por sus siglas en inglés) (Louis et al., 2016; Wesseling & Capper, 2018). De acuerdo con la WHO, los gliomas astrocíticos se pueden clasificar en dos subtipos: tumores difusos y tumores no difusos. Además, atendiendo a su grado de malignidad pueden clasificarse como tumores de grado I, II, III y IV, siendo los tumores de grado IV los más agresivos. En nuestro estudio nos centraremos en los tumores astrocíticos de tipo difuso (WHO grado II (denominados astrocitomas difusos, DA), WHO grado III (conocidos como astrocitomas anaplásicos, AA) y WHO grado IV (glioblastoma, GBM)), excluyendo los astrocitomas pilocíticos (PA; WHO grado I). Los subtipos de tumores astrocíticos difusos son muy heterogéneos y su pronóstico y tratamiento muy diferentes. El diagnóstico patológico es la prueba de oro estándar del examen clínico en la práctica, pero la variabilidad intratumoral e interobservador complican la clasificación histológica de los tumores. Por ello, el reconocimiento de las alteraciones genéticas en los subtipos de gliomas astrocíticos podría contribuir a refinar su diagnóstico histológico actual. Para ello, se analizará la información de expresión génica de un total de 176 gliomas astrocíticos difusos y se validarán posteriormente los resultados obtenidos en una muestra de 113 tumores adicional. Se obtuvieron datos de un total de 12 series públicas del repositorio GEO, respectivas a datos de la plataforma HG-U133Plus2.

3- **Leucemias.** La leucemia es un conjunto de enfermedades que se caracteriza por la acumulación de células de la sangre anormales en la médula ósea debido a un desorden en su proceso de maduración,

denominada hematopoyesis. De este modo, la producción normal de dichas células queda interferida y las células malignas pueden llegar a diseminarse a sangre o a otros órganos. Mayoritariamente se distinguen dos grupos generales: leucemia aguda, que evoluciona rápidamente, y leucemia crónica, que se desarrolla en un proceso más pausado. En este estudio se considerarán datos respectivos a la primera etapa del proyecto MILE (*Microarray Innovations in Leukemia*) (Haferlach et al., 2010), destinado al progreso en la investigación del diagnóstico y subclasificación de las leucemias. Se recogieron un total de 2.096 muestras, de 11 laboratorios de diferentes estados, de microarrays de expresión de la plataforma HG-U133Plus2. Se almacenaron muestras de pacientes con diferentes tipos de leucemias y pacientes control. Del total de las muestras proporcionadas, en nuestro estudio se seleccionaron aleatoriamente 216 muestras: 32,9% (n=71) muestras de leucemia linfoblástica aguda (ALL), 34,3% de muestras de leucemia linfoblástica crónica (n=74) y 32,9% de muestras sin leucemia (n=71). Se recogió información de 44.692 sondas de genes referentes a 21.336 genes.

2.-Metodología

Con el fin de cumplir los objetivos planteados en el marco de este trabajo, la metodología estadística de la que se hará uso a lo largo de todo el documento estará enfocada a métodos de selección de variables desde distintas perspectivas: métodos de factorización matricial y métodos de reducción de la dimensionalidad tanto desde la perspectiva del análisis de datos de dos vías, como de su extensión multivía, haciendo especial hincapié en el análisis de datos de tres vías. Se realizará una revisión de las principales técnicas diseñadas para dicho fin en la literatura y se propondrá una nueva metodología de selección de variables, cuyo eje central será la descomposición en valores singulares.

Estructura del trabajo

El Capítulo 1 comienza con una descripción de los métodos clásicos de dos vías, desde las técnicas de descomposición matricial, como la descomposición en valores singulares y la descomposición CUR, hasta las técnicas de reducción de la dimensión clásicas, como el Análisis de Componentes Principales o los métodos Biplot. La segunda parte del capítulo se centra en la definición formal del problema de mínimos cuadrados ordinario y penalizado, y se presentan las principales técnicas de penalización utilizadas en la literatura, como Lasso.

El capítulo 2 recoge las principales contribuciones prácticas de la tesis al análisis de datos de dos vías mediante técnicas clásicas y técnicas sparse.

En el Capítulo 3 se presentan las nociones básicas en cuanto a la convexidad de conjuntos y se realiza una exhaustiva revisión de los métodos de proyección de un vector sobre un conjunto de restricciones convexas y de sus correspondientes algoritmos.

En el capítulo 4 se presenta la que es la principal contribución teórica de esta tesis: la propuesta de C$_{enet}$SVD. Para ello, previamente se presentará el método de proyección de un vector sobre la región $\mathfrak{B}_{\ell_1+\ell_2}(\tau) \cap \mathfrak{B}_{\ell_2}(1)$ y a continuación se formulará el problema de optimización que plantea C$_{enet}$SVD. Así mismo, se presenta en este capítulo la extensión de la SVD penalizada al ámbito de las técnicas de reducción de la dimensión, presentando C$_{enet}$PCA y C$_{enet}$Biplot. Finalmente, las metodologías propuestas se aplican al análisis de datos reales de alta dimensión.

En el capítulo 5 se realiza una revisión de los métodos sparse implementados para el análisis de datos multivía, desde el punto de vista de la concatenación de matrices y desde el punto de vista de los tensores multidimensionales, revisando principalmente los modelos Tucker. Posteriormente, se plantean los modelos C$_{enet}$JIVE y C$_{enet}$Tucker, una contribución enfocada a la extensión de la C$_{enet}$SVD al análisis de matrices multivía.

Finalmente se recogen posibles líneas futuras de trabajo y se presentan las principales conclusiones de este trabajo de tesis doctoral.

CAPÍTULO 1

ANÁLISIS DE DATOS DE DOS VÍAS: MÉTODOS CLÁSICOS

1.1 Problema de mínimos cuadrados ordinario

La optimización por mínimos cuadrados ordinarios (OLS por sus siglas en inglés *Ordinal Least Squares*) es una herramienta necesaria para la modelización lineal generalizada que se utiliza para modelizar una respuesta $Y = (y_1, \dots, y_I)^T$ a partir de un conjunto de variables explicativas $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_J$ almacenadas en $\boldsymbol{X}_{IxJ}$. Para entender la motivación, considérese el modelo de regresión para una variable respuesta $\boldsymbol{y}_i$:

$$\boldsymbol{y}_i = \beta_1 \boldsymbol{x}_{i1} + \beta_2 \boldsymbol{x}_{i2} + \cdots + \beta_J \boldsymbol{x}_{iJ} + e_i$$

Supóngase sin pérdida de generalidad que las variables predictoras están centradas y estandarizadas y que la variable respuesta está centrada. Matricialmente, $\boldsymbol{X}_{IxJ}$ es la matriz de variables explicativas**,** $\hat{Y}_{Ix1}$ la matriz respuesta y $\widehat{\boldsymbol{\beta}}_{Jx1}$ la matriz de coeficientes de regresión a estimar.

$$\widehat{\boldsymbol{Y}}_{Ix1} = \boldsymbol{X}_{IxJ} \widehat{\boldsymbol{\beta}}_{Jx1}$$

Matemáticamente, la solución $\widehat{\boldsymbol{\beta}}_{Jx1}$ a este problema viene dada por la estimación mínimo-cuadrática, que trata de minimizar la función objetivo que se plantea en el problema de mínimos cuadrados ordinarios. Este define la estimación $\widehat{\boldsymbol{\beta}}_{Jx1}$ como la solución a un problema de optimización sin restricciones en términos de la suma de cuadrados haciendo mínimo el error residual:

$$\widehat{\boldsymbol{\beta}}_{Jx1} = \min_{\boldsymbol{\beta}} \left\| \boldsymbol{Y} - \widehat{\boldsymbol{Y}} \right\|_F^2 = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_F^2 = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$$

La solución que minimice la expresión anterior (denotada de ahora en adelante como SSE (*squared sum of errors*)) de mínimos cuadrados $\widehat{\boldsymbol{\beta}}$ se conoce como estimación ordinaria de mínimos cuadrados.

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*

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

La varianza de los coeficientes estimados mediante mínimos cuadrado es

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

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

Como se sigue de (Hastie et al. 2009), derivando con respecto a $\boldsymbol{\beta}$ se tiene:

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

y además,

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

Si $\boldsymbol{X}$ es de rango completo, la solución viene dada por:

$$0 = -2\boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{X\beta})$$

$$\boldsymbol{X^T X\beta} = \boldsymbol{X^T Y}$$

$$\widehat{\boldsymbol{\beta}}^{OLS} = (\boldsymbol{X^T X})^{-1} \boldsymbol{X^T Y}$$

Es trivial que para que la estimación OLS exista, $\boldsymbol{X^T X}$ tiene que tener inversa, lo que ocurrirá cuando $det(\boldsymbol{X^T X})$ no sea 0; es decir, cuando $\boldsymbol{X}$ sea de rango completo. Además, si $J > I$ la solución no será única. Tal y como expone Hastie et al. (2009) en su trabajo, hay dos puntos esenciales que no deben perderse de vista: i) las estimaciones por mínimos cuadrados tienden a presentar un sesgo bajo, pero una gran varianza, de manera que si se contrajesen algunos coeficientes a 0 la exactitud de la predicción podría verse mejorada haciendo un sacrificio de sesgo para reducir la varianza de los valores predichos; y ii) con un gran conjunto de variables explicativas, sería óptimo escoger un conjunto menor de variables que recojan los mayores efectos y con ellos obtener a una interpretación factible.

En este sentido los métodos de selección de variables tenderán una mano en aquellos casos donde se den algunos de los inconvenientes anteriores.

Algunas de estas técnicas engloban la selección de variables hacia adelante y hacia atrás, pero en este trabajo se situará el foco de atención en los métodos de optimización penalizados, cuyo objetivo será incluir algún tipo de restricción sobre los parámetros del modelo a definir. De esto se hablará más adelante, en el capítulo 2.

1.2 Métodos de factorización matricial

En el ámbito del álgebra lineal se entiende por factorización de una matriz a su descomposición como producto de dos o más matrices que aproximen esta de manera adecuada. De entre los métodos de descomposición matricial destacan la factorización LU, la factorización QR, la factorización de Schur… Sin embargo, en el ámbito de la estadística multivariante el método de descomposición matricial por excelencia es la Descomposición en Valores Singulares.

Los métodos de factorización matricial constituyen la base de las principales técnicas aplicadas de la estadística multivariante. Por todo ello, en este capítulo se presentan las principales técnicas de descomposición matricial tradicionales, así como algunos de los principales desarrollos encontrados en la literatura en los últimos años en torno al ámbito de las técnicas de regularización, convirtiéndose en técnicas adaptadas a las necesidades del análisis de datos de altas dimensiones.

1.2.1 Descomposición en valores singulares

La Descomposición en Valores Singulares (*Singular Value Decomposition*, SVD) (Eckart & Young, 1936) es una de las herramientas más potentes aportadas por el álgebra lineal y uno de los métodos de descomposición matricial más utilizado debido a sus importantes propiedades. Como se mencionará más adelante, constituye el grosso de muchas de las técnicas más relevantes de la estadística multivariante tanto en el análisis de datos de dos vías como de tres o más: análisis de componentes principales, análisis de correlación canónica, análisis de correspondencias múltiple, HOSVD, Tucker, … Se trata de la generalización de la *eigen-decomposition* para matrices no cuadradas.

Dada una matriz cualquiera $\boldsymbol{X}_{IxJ}$ de rango $r \leq \min(I, J)$ e $I \neq J$, la SVD de $\boldsymbol{X}$ es única y trata de aproximar ésta mediante el producto de tres matrices:

$$\boldsymbol{X}_{IxJ} = \boldsymbol{U}_{Ixr}\boldsymbol{D}_{rxr}\boldsymbol{V}^{T}{}_{rxJ} \tag{1.1}$$

siendo $\boldsymbol{U} = [\boldsymbol{u}_1, \dots, \boldsymbol{u}_I]$ y $\boldsymbol{V} = [\boldsymbol{v}_1, \dots, \boldsymbol{v}_J]$ las matrices ortonormales cuyos vectores columna son los vectores singulares a izquierda y derecha respectivamente, $U^T U = I \; y \; V^T V = I$, y D la matriz diagonal que almacena los valores singulares de X, expresados convenientemente de forma que: $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r \geq 0$.

En el desarrollo de la SVD, las matrices de correlación y covarianzas de X, $X^T X$ y XX^T, juegan un papel muy importante (Figura 3). Mientras que es posible el cálculo de la SVD de cualquier matriz, sólo es posible la diagonalización ortogonal de matrices simétricas y es por eso que $X^T X$ y XX^T son necesarias en el cálculo de la descomposición (Eckart & Young, 1936). Por otro lado, la matriz D es una matriz diagonal que contiene los valores singulares de la descomposición; esto es, las raíces cuadradas no negativas de los valores propios asociados a los vectores propios de las matrices $X^T X$ y XX^T. Esto se debe a que la SVD es consecuencia de la eigen-decomposicion de una matriz semidefinida positiva. A partir de la SVD de $X^T X$ y XX^T se obtienen las siguientes ecuaciones:

$$\boldsymbol{X}\boldsymbol{X}^T = (\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T)(\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T)^T = (\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T)(\boldsymbol{V}\boldsymbol{D}^T\boldsymbol{U}^T) = \boldsymbol{U}\boldsymbol{D}^2\boldsymbol{U}^T$$

$$\boldsymbol{X}^T\boldsymbol{X} = (\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T)^T(\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T) = (\boldsymbol{V}\boldsymbol{D}^T\boldsymbol{U}^T)(\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T) = \boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^T$$

Así, $\boldsymbol{D}$ es la raíz cuadrada (elemento a elemento) de $\boldsymbol{D}^2$, que contiene los valores propios, $\boldsymbol{V}$ contiene los vectores propios de $\boldsymbol{X}^T\boldsymbol{X}$ y $\boldsymbol{U}$ los vectores propios de $\boldsymbol{X}\boldsymbol{X}^T$.

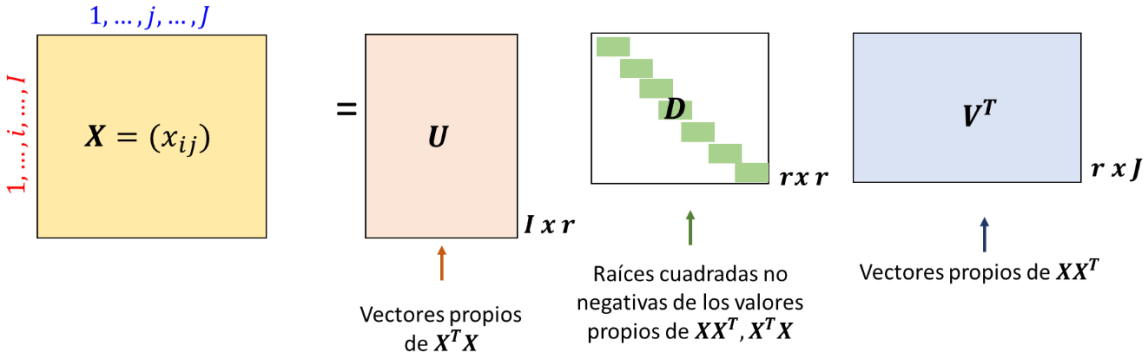


***Figura 3**. Descomposición en valores singulares de una matriz $\boldsymbol{X}$*

La importancia de la SVD se debe a una importante propiedad matemática y es que genera la mejor aproximación de bajo rango de una matriz rectangular cualquiera (Abdi & Valentin, 2006). En concreto, proporciona la mejor aproximación de rango Q de la matriz original en el sentido de los mínimos cuadrados, minimizando la norma de Frobenius al cuadrado entre la matriz original y la matriz de rango Q reconstruida $\widehat{\boldsymbol{X}}_Q$ (Eckart & Young, 1936; Shen & Huang, 2008).

$$\left\|\boldsymbol{X}-\widehat{\boldsymbol{X}}_Q\right\|_{\mathrm{F}}^2=\|\boldsymbol{X}-\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\mathrm{T}}\|_{\mathrm{F}}^2=traza((\boldsymbol{X}-\widehat{\boldsymbol{X}}_Q)(\boldsymbol{X}-\widehat{\boldsymbol{X}}_Q)^{\boldsymbol{T}}) \tag{1.2}$$

La matriz $\widehat{\boldsymbol{X}}_Q$ óptima se obtiene como:

$$\widehat{\boldsymbol{X}}_Q=\sum_{q=1}^{Q} d_{\mathrm{q}}\boldsymbol{u}_q\boldsymbol{v}_q{}^T$$

con $\boldsymbol{u}_q{}^T\boldsymbol{u}_q=\boldsymbol{v}_q{}^T\boldsymbol{v}_q=1$ y $\boldsymbol{u}_q{}^T\boldsymbol{u}_{q'}=\boldsymbol{v}_q{}^T\boldsymbol{v}_{q'}=0\ \forall q\neq q'$, resolviendo el problema de optimización de minimización del error:

$$\underset{d,\boldsymbol{u},\boldsymbol{v}}{\operatorname{argmin}}\ \frac{1}{2}\left\|\boldsymbol{X}-d_q\boldsymbol{u}_q\boldsymbol{v}_q{}^T\right\|_F^2$$

$$\text{s.a. }\{\boldsymbol{u}_q{}^T\boldsymbol{u}_q=\boldsymbol{v}_q{}^T\boldsymbol{v}_q=1,\boldsymbol{u}_q{}^T\boldsymbol{u}_{q'}=\boldsymbol{v}_q{}^T\boldsymbol{v}_{q'}=0\quad\forall q\neq q'$$

Nótese que, como los valores singulares se presentan en orden decreciente y dado que la varianza absorbida por cada componente se calcula a partir de ellos, se mantiene la propiedad de que las PCs capturen secuencialmente la máxima variabilidad entre las columnas de X.

Frecuentemente, la SVD es implementada a nivel computacional mediante el algoritmo Power Iteration (método de potencia) (Tabla 1), desarrollado inicialmente para calcular la *eigen-decomposition* de una matriz cuadrada. A

pesar de ser un algoritmo inestable y cuya convergencia es muy lenta es sencillo de implementar y efectivo en casos de altas dimensiones.

Tabla 1. *Método Power Iteration para el cálculo de la SVD*

Algoritmo: SVD	
Entrada:	$\boldsymbol{X} \in \mathbb{R}^{IxJ}$, rango Q, $\varepsilon \approx 0$
Salida:	$\boldsymbol{U} \in \mathbb{R}^{IxQ}, \boldsymbol{D} \in \mathbb{R}^{QxQ}, \boldsymbol{V} \in \mathbb{R}^{JxQ}$
1:	**Para** q en 1: Q hacer:
2:	Inicializar $\boldsymbol{u}_0, \boldsymbol{v}_0$ aleatoriamente
	$\boldsymbol{X}_1 = \boldsymbol{X}$
	$\lambda_0 = 0$
3:	$\lambda_1 = \boldsymbol{u}_0^T \boldsymbol{X} \boldsymbol{v}_0$
	$t = 0$
	Mientras $\lvert\lambda_{t+1} - \lambda_t\rvert \geq \varepsilon$ hacer:
	$\boldsymbol{u}_{t+1} = \text{normalizar}(\boldsymbol{X}\boldsymbol{v}_t)$
	$\boldsymbol{v}_{t+1} = \text{normalizar}(\boldsymbol{X}^T \boldsymbol{u}_{t+1})$
	$\lambda_{t+1} = \boldsymbol{u}_{t+1}^T \boldsymbol{X} \boldsymbol{v}_{t+1}$
	$t = t + 1$
6:	$\boldsymbol{X}_{t+1} = \boldsymbol{X}_t - \lambda_t \boldsymbol{u}_t \boldsymbol{v}_t^T$
9:	**Fin**

Los vectores singulares obtenidos en la SVD se corresponden con una combinación entre todas las variables de partida de la matriz rectangular. Las combinaciones lineales presentan propiedades óptimas de síntesis de un conjunto de variables en un subespacio de menor dimensión. Ahora bien, en la práctica, las operaciones matemáticas realizadas en la SVD no respetan la estructura inicial de los datos. Esto provoca que en aquellas ocasiones en las que las matrices de partida posean una gran cantidad de coeficientes nulos o incluso casos en los que es necesario mantener la estructura “positiva” de los datos, la SVD hace desaparecer las propiedades iniciales de las matrices, llegando incluso a generar resultados que no son interpretables. Por estos motivos (y otros) la literatura engloba diversas técnicas de factorización matricial que tratan de paliar este tipo de inconvenientes. Entre otras existentes, en este trabajo se hará referencia a la factorización matricial no negativa y la descomposición CUR.

1.2.2 Factorización matricial no negativa

La factorización matricial no negativa (NMF, por sus siglas en inglés) (Lee & Seung, 1999) es un método no supervisado de aproximación de una matriz en bajo rango que se posiciona como un método de biclustering.

Se define como la descomposición matricial que aproxima una matriz de valores no negativos como el producto de dos matrices de bajo rango, también no negativas. En el caso de la SVD o PCA, los datos son aproximados de manera que pueden involucrar "cancelaciones" entre valores positivos y negativos. Esta situación, en algunas aplicaciones, contradice el sentido físico de los datos, como puede ser el caso de datos de expresión génica, donde un valor no negativo no es interpretable. Este tipo de situación hace que tenga sentido la propuesta de técnicas en las que tanto los factores como sus coeficientes sean restringidos en un sentido no negativo.

Como se decía, la NMF es una técnica de biclustering y reducción de la dimensión que trata de encontrar Q combinaciones lineales de las variables originales (conocidos como módulos) que expliquen los datos de partida. Sin embargo, a diferencia de lo que ocurre en el PCA, estas tienen que ser positivas. Las nuevas dimensiones latentes generadas son conocidas como módulos (por ser una técnica de agrupación) o componentes base. Matemáticamente, la NMF factoriza una matriz $\boldsymbol{X} \in \mathbb{R}_{+}^{I\,x\,J}$ positiva, con I observaciones y J variables medidas, en el producto de dos matrices de coeficientes no negativos:

$$\boldsymbol{X} \approx \boldsymbol{H}\boldsymbol{W}^{\boldsymbol{T}}$$

con $\boldsymbol{H} \in \mathbb{R}_{+}^{I\,x\,Q}$ y $\boldsymbol{W} \in \mathbb{R}_{+}^{JxQ}$. Lo más frecuente es estimar las matrices $\boldsymbol{W}$ y $\boldsymbol{H}$ como las soluciones que minimizan el problema de optimización:

$$\min_{\boldsymbol{H},\boldsymbol{W}>0} \|\boldsymbol{X} - \boldsymbol{H}\boldsymbol{W}^{\boldsymbol{T}}\|_F^2$$

En la NMF, las variables de partida de $\boldsymbol{X}$ son sustituidas por combinaciones lineales no negativas de las mismas, cuyos coeficientes se almacenan en las columnas de $\boldsymbol{W}$. Esta factorización proporciona una interpretación directa debida a las combinaciones positivas de los vectores no negativos. Se trata de una técnica de selección de variables natural puesto que debido a sus restricciones

no negativas habitualmente genera vectores sparse (con parte de sus coeficientes nulos). En este sentido, es considerada una técnica de biclustering. La matriz $\boldsymbol{H}$ es utilizada para agrupar las I observaciones en Q módulos (o clusters). El hecho de que la matriz $\boldsymbol{W}$ tenga coeficientes exactamente nulos permite descubrir patrones en las variables asociados a subconjuntos de observaciones.

El problema anterior es un problema convexo únicamente para $\boldsymbol{H}$ o $\boldsymbol{W}$, pero no para ambas simultáneamente. Esto quiere decir que no existe un algoritmo que encuentre un mínimo global de (1) (Cai, He, Wu, & Han, 2008). Lee y Seung (1999) proponen para su resolución un algoritmo iterativo basado en las siguientes reglas de actualización multiplicativas de las matrices $\boldsymbol{H}$ y $\boldsymbol{W}$.

$$h_{ij} \leftarrow h_{ij}^{T} \frac{(\boldsymbol{XW^T})_{ij}}{(\boldsymbol{HW^TW})_{ij}}$$

$$w_{ij} \leftarrow w_{ij}^{T} \frac{(\boldsymbol{X^TH})_{iq}}{(\boldsymbol{WH^TH})_{iq}}$$

A pesar de que frecuentemente la función de pérdida asociada al problema de optimización está basada en la norma de Frobenius, otros autores proponen el uso de otras divergencias, como la divergencia de Kullback-Leibler (D D Lee & Seung, 2001). Otra de las cuestiones a destacar es que la convergencia de los algoritmos existentes para las reglas multiplicativas, iterativos, es lenta (Salakhutdinov, Roweis & Ghahramani, 2002). En parte esto puede ser debido a la naturaleza iterativa de dichos algoritmos, muy dependiente de la inicialización de las matrices $\boldsymbol{H}$ y $\boldsymbol{W}$. Por eso, la búsqueda de buenas inicializaciones es otro punto abierto en la investigación. En este sentido, Boutsidis y Gallopoulos, (2008) proponen NNDSVD (*nonnegative double singular value decomposition*), un método para la inicialización de las matrices $\boldsymbol{H}$ y $\boldsymbol{W}$ a partir de los resultados de la SVD clásica.

La popularidad de esta técnica de descomposición factorial se debe a que presenta una serie de propiedades óptimas en el análisis de datos (Gaujoux & Seoighe, 2010). Entre otras razones, los factores latentes son generados por una combinación de coeficientes positivos, lo que da lugar a dimensiones latentes cuya interpretación puede ser más intuitiva. Esta primera propiedad da lugar a la

segunda que le caracteriza: genera resultados sparse. Dado que los datos de partida se reconstruyen a partir de dimensiones puramente positivas, dicha aproximación provoca que algunos de los coeficientes de las combinaciones lineales deban ser nulos para que la aproximación tenga sentido. No se convierte en una técnica penalizada para producir cargas nulas, pero sí es considerada una técnica sparse por generar valores exactamente cero. Esta propiedad define la NMF como una técnica de biclustering, pues produce resultados sparse automáticamente tanto en la matriz $\boldsymbol{H}$, de puntuaciones de las observaciones en las dimensiones latentes, como en la matriz $\boldsymbol{W}$, matriz de configuración de los ejes. Esto hace que sus resultados sean fácilmente interpretables (Gillis, 2014). Ahora bien, a diferencia de las técnicas puramente sparse la cantidad de coeficientes exactamente cero que desean generarse es completamente arbitraria y no puede controlarse en este contexto.

Equivalentemente a lo que ocurrirá con los métodos que promueven la penalización sparse, y a diferencia de técnicas como el PCA, los factores (o módulos) generados en la NMF no serán ortogonales. Esta propiedad se pierde al imponer la restricción de no negatividad sobre las matrices $\boldsymbol{H}$ y $\boldsymbol{W}$. En campos como la genética esto no supone una desventaja puesto que pueden así encontrarse genes correlacionados con más de un módulo a la vez (Brunet, Tamayo, Golub, & Mesirov, 2004). Sin embargo, en otras ocasiones, resulta natural imponer la ortogonalidad en las matrices $\boldsymbol{H}$ y $\boldsymbol{W}$, dando lugar a la denominada ONMF (*orthogonal nonnegative matrix factorization*) (Choi, 2008; Yoo & Choi, 2010):

$$\min_{\boldsymbol{H},\boldsymbol{W}>0} \|\boldsymbol{X} - \boldsymbol{H}\boldsymbol{W}^{\boldsymbol{T}}\|_F^2$$
$$s.a. \boldsymbol{H}^{\boldsymbol{T}}\boldsymbol{H} = \boldsymbol{I}, \boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{W} = \boldsymbol{I}$$

Esta formulación se corresponde con el algoritmo K-Means (Ding, He, & Simon, 2005; Yoo & Choi, 2010). Otras extensiones de la NMF incluyen la non-smooth NMF (NS-NMF) (Carmona-Saez, Pascual-Marqui, Tirado, Carazo, & Pascual-Montano, 2006), LS-NMF (Wang, Kossenkov, & Ochs, 2006), KernelNMF (Li & Ngom, 2012), Semi-NMF, Convex-NMF,… Además, esta factorización ha generado resultados fructíferos en muchas aplicaciones: en el análisis de datos de microarrays y expresión génica en cáncer (Brunet et al.,

2004), en arte con el estudio de cuadros (Alfeld et al., 2014), análisis de música (Févotte, Bertin, & Durrieu, 2009), … De hecho ha sido considerada más útil que la SVD en algunas disciplinas, como reconocimiento facial o análisis de texto (Li, Hou, Zhang, & Cheng, 2001; Xu, Liu, & Gong, 2003).

Una revisión más exhaustiva de estos métodos puede verse en (Cichocki et al., 2009). Los distintos algoritmos para ejecutar la NMF de una matriz están disponibles en la librería *NMF* de R (Gaujoux & Seoighe, 2010). También existen implementaciones en MATLAB.

1.2.3 Descomposición CUR

En los últimos años, y siguiendo la idea de las variables principales de McCabe (1984) aparece la descomposición CUR (Mahoney & Drineas, 2009). Estos autores identifican los vectores singulares de la SVD como la causa de la dificultad en la interpretación de las PCs. Por eso, proponen otro tipo de factorización conocida como descomposición CUR. El objetivo principal de CUR es aproximar la matriz de datos original mediante un menor número de observaciones y/o variables, convirtiéndose en una técnica de selección de variables. Esta no es más que una aproximación de bajo rango de la matriz original expresada en un pequeño número de filas y/o columnas de esta, más interpretables que los vectores singulares de la SVD. Dada $\boldsymbol{X} \in \mathbb{R}^{IxJ}$, se define la descomposición CUR de $\boldsymbol{X}$ como $\boldsymbol{X} \approx \boldsymbol{CUR}$, donde la matriz $\boldsymbol{C} \in \mathbb{R}^{Ixc}$ contiene un subconjunto de las columnas de $\boldsymbol{X}$, la matriz $\boldsymbol{R} \in \mathbb{R}^{rxJ}$ contiene un número reducido de filas de $\boldsymbol{X}$ y la matriz $\boldsymbol{U} \in \mathbb{R}^{cxr}$, calculada a partir de $\boldsymbol{C}$ y $\boldsymbol{R}$, garantiza que el producto matricial CUR aproxime $\boldsymbol{X}$ satisfactoriamente (Figura 4).

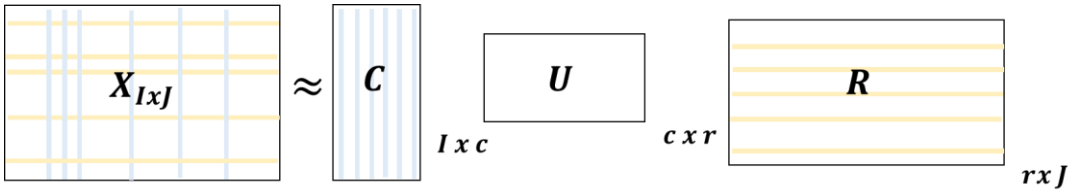


***Figura 4**. Descomposición matricial CUR de una matriz X*

Las filas y/o columnas de la matriz original se seleccionan en base a su nivel de influencia o *leverage*. Para cada fila y/o columna de la matriz de datos, estos factores de importancia se definen a partir de la SVD de la matriz original. Como método de selección de variables, CUR busca proporcionar a cada una de las variables originales un factor de importancia que clasifique las variables en torno a la información que aporten en el modelo final. Sean $\boldsymbol{v}_j$ $(j = 1, \dots, J)$ los vectores singulares a derecha obtenidos en la SVD de la matriz original, se definen para cada variable el leverage correspondiente de la siguiente forma:

$$l_j = \frac{1}{Q}\sum_{j=1}^{Q}(v_{ij})^2$$

donde Q es el número de nuevos ejes en la reducción de la dimensión. El cálculo de los leverage para cada uno de los individuos de la matriz, para la selección de los más influyentes, se realiza de la misma manera que en el caso de las variables, pero con la diferencia de que la SVD se calcula para la matriz original traspuesta y no para la matriz original (como era el caso de las variables) (Figura 5).

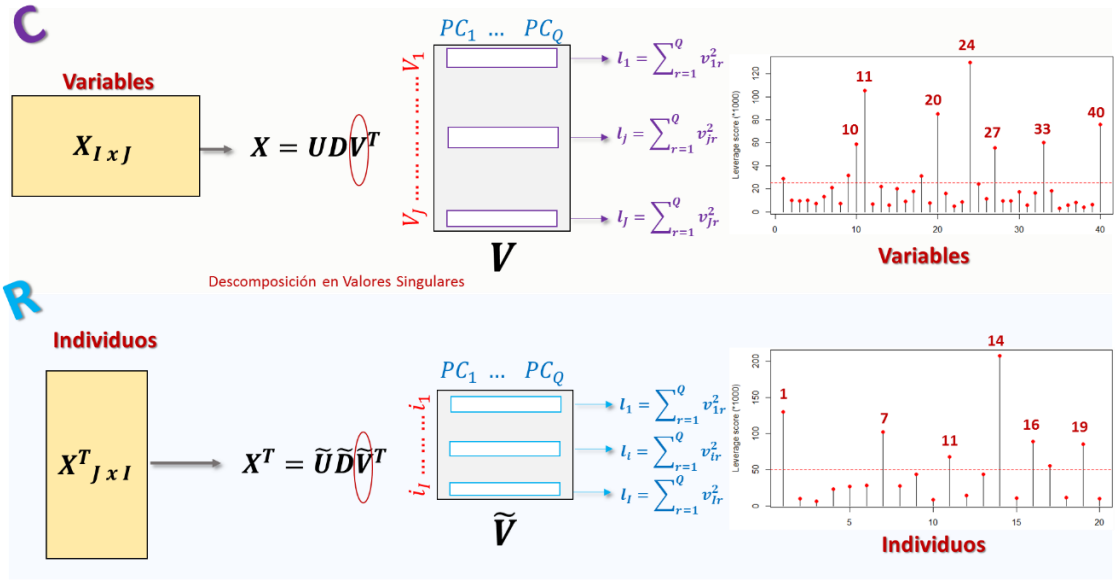


Figura 5. *Cálculo de los leverage para cada variable y/o individuo para la creación de las matrices C y R respectivamente*

De manera general, los leverage se interpretan naturalmente como sensores de la influencia de cada columna en la mejor aproximación de menor rango de la matriz de datos. Esto es, representan la cantidad de varianza aportada por cada variable en el modelo factorial definido. Debe considerarse que CUR muestra, a través de estos valores, las más importantes, a nivel general, en el modelo factorial supuesto y no las más importantes en cada componente como el PCA o *Sparse* PCA.

Una vez definidos los factores de influencia para cada variable y/o individuo de la matriz original, las columnas y/o filas de la matriz son seleccionadas según distintos criterios que tienen en cuenta la distribución establecida por los leverage.

En primer lugar, se definen las matrices $\boldsymbol{C}$ y $\boldsymbol{R}$ de la descomposición a partir de estos valores calculados. Las variables seleccionadas como más influyentes en el modelo a partir de los leverage son las que conforman la matriz $\boldsymbol{C}$. Los individuos seleccionados como más influyentes en el modelo, a través de distintos criterios, son los que conforman la matriz $\boldsymbol{R}$. Las matrices $\boldsymbol{C}$ y $\boldsymbol{R}$ no son más que una submatriz de la matriz original formada por las variables y/o individuos seleccionados, respectivamente. La aproximación CUR no es única y existen múltiples algoritmos para calcular dicha aproximación que se diferencian en las cotas de error obtenidas y en los criterios para seleccionar las filas y las columnas, una vez definidos los leverage. Aunque el más intuitivo es el método *top.scores* (Bodor, Csabai, Mahoney, & Solymosi, 2012) basado en, una vez definidos los leverage para cada columna y/o fila de la matriz inicial, seleccionar aquellas variables con leverage más altos. Ahora bien, existen en la literatura otras alternativas para la selección de filas y/o columnas como el método *random*, método original definido en (Mahoney & Drineas, 2009), en el que las columnas y/o filas son seleccionadas teniendo en cuenta la distribución de probabilidad generada por los leverage. Existen otras muchas alternativas, como el método *ortho.top.scores*, que selecciona filas y/o columnas teniendo en cuenta la combinación de los leverage con la ortogonalidad del subespacio formado por las variables y/o individuos seleccionados anteriormente, de manera que sea máxima (Bodor et al., 2012; Drineas, Kannan, & Mahoney, 2006; Drineas, Mahoney, & Muthukrishnan, 2007; Frieze, Kannan, & Vempala, 2013). Puede

entenderse así que, aunque el cálculo de los leverage es común para todos los métodos, la selección de las filas originales que conformarán la matriz $\boldsymbol{R}$, o de las columnas que conformarán la matriz $\boldsymbol{C}$, se realiza por métodos diferentes que tienen en cuenta estos factores de influencia.

Una vez identificadas las variables y/o individuos más importantes por CUR, la matriz $\boldsymbol{C}$ se define como la submatriz de la matriz original $\boldsymbol{X}$ formada por las columnas seleccionadas por CUR. De igual manera, una vez seleccionadas las filas más relevantes (como se ha visto, en función de distintos criterios teóricos), la matriz $\boldsymbol{R}$ se define como la submatriz de la matriz original $\boldsymbol{X}$ formada por las filas seleccionadas. Con las matrices $\boldsymbol{C}$ y $\boldsymbol{R}$ determinadas, se define la matriz $\boldsymbol{U}$ de manera que CUR aproxime de la mejor forma posible $\boldsymbol{X}$. La matriz $\boldsymbol{U}$ se define $\boldsymbol{U} = \boldsymbol{C}^{\dagger}\boldsymbol{X}\boldsymbol{R}^{\dagger}$, con $\boldsymbol{C}^{\dagger}$ y $\boldsymbol{R}^{\dagger}$ las matrices pseudoinversas de $\boldsymbol{C}$ y R respectivamente (Figura 6). Así, demuestran que el error cometido en la aproximación es mínimo y verifica $\|\boldsymbol{X} - CUR\| \leq (2+\epsilon)\left\|\boldsymbol{X} - \widehat{\boldsymbol{X}}_Q\right\|$, con $\widehat{\boldsymbol{X}}_Q$ la mejor aproximación de rango Q de $\boldsymbol{X}$ y ϵ un parámetro de precisión establecido (Mahoney & Drineas, 2009).

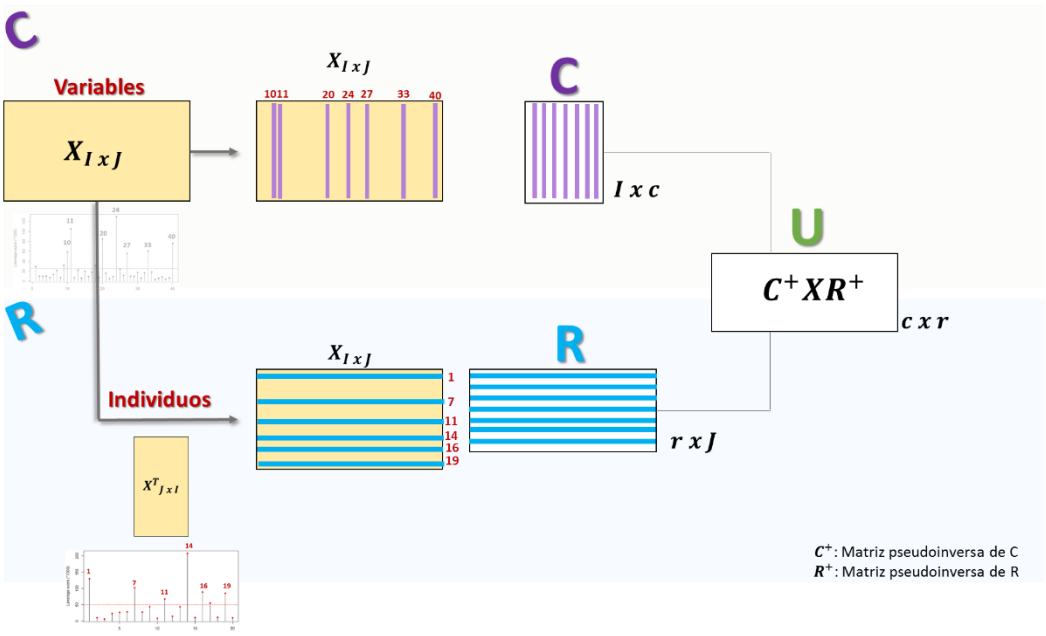


***Figura 6**. Algoritmo CUR mediante el método top.scores*

Analizando las matrices $\boldsymbol{C}$ y $\boldsymbol{R}$ obtenidas en la descomposición CUR, el investigador identifica las variables e individuos característicos de la muestra en

estudio (los que más información aportan). Así, al igual que el PCA y el *Sparse* PCA ayudan a establecer las variables con mayor importancia dentro de cada latente, a través del valor de las cargas, CUR permite al analista determinar las variables importantes dentro de la muestra completa.

CUR aproxima la matriz de partida mediante un número menor de sus columnas y/o filas, de manera que construye una representación informativa que facilita la interpretación. Al no tratarse de una técnica de reducción de la dimensión como tal, no puede verificarse la estructura teórica subyacente, pero permite identificar las variables importantes de la matriz inicial y manejar grandes volúmenes de datos. La ventaja que tiene sobre la SVD es que permiten expresar la matriz de datos respecto de un número reducido de columnas y/o filas en lugar de extraer ejes factoriales que son combinaciones lineales de todas las variables de partida.

En el capítulo 2 el lector puede encontrar una aplicación de la NMF y descomposición CUR aplicada al análisis de datos genómicos junto a otras técnicas de selección de variables.

1.3 Técnicas multivariantes clásicas de reducción de la dimensión

Habitualmente el uso en la estadística multivariante de técnicas de descomposición matricial está íntimamente ligado al objetivo de reducir la dimensión del espacio vectorial donde residen los datos de partida. Se presentan a continuación algunas de las principales técnicas de reducción de la dimensionalidad y que posteriormente tendrán un peso importante en la contribución de la tesis doctoral.

1.3.1 Análisis Factorial

El Análisis Factorial (*Factorial Analysis*, FA) (Spearman, 1904) es una técnica de reducción de la dimensión que surgió para responder a necesidades prácticas en ámbito social. Sea $\boldsymbol{X}_{IxJ}$ una matriz que recoge las mediciones de J variables sobre una muestra de I individuos u objetos, que se supondrá centrada por columnas sin pérdida de generalidad de aquí en adelante. El análisis factorial

de $\boldsymbol{X}_{IxJ}$ se define como una técnica de análisis estadístico no supervisada diseñada para analizar y explicar las relaciones entre las J variables originales en términos de un nuevo conjunto de Q variables latentes comunes, conocidas como factores. Esto es posible debido a que muchas de las cuestiones analizadas son en realidad aspectos de una misma característica latente, que resulta de la relación entre las variables originales. Considérense $\boldsymbol{x}_1, \dots, \boldsymbol{x}_J$ las variables de partida. El FA trata de estimar un conjunto de Q factores latentes comunes $\boldsymbol{F}_1, \dots, \boldsymbol{F}_Q$ y J factores únicos $\boldsymbol{U}_1, \dots, \boldsymbol{U}_J$, cuya combinación lineal defina cada una de las variables originales:

$$\begin{gathered}\boldsymbol{x}_1 = a_{11}\boldsymbol{F}_1 + \cdots + a_{JQ}\boldsymbol{F}_Q + d_1\boldsymbol{U}_1 \\ \dots \\ \boldsymbol{x}_J = a_{J1}\boldsymbol{F}_1 + \cdots + a_{JQ}\boldsymbol{F}_Q + d_J\boldsymbol{U}_J\end{gathered}$$

donde $Q < J$, para explicar las variables observadas con un número reducido de factores subyacentes, y las $Q + J$ dimensiones definidas son no correlacionadas. Los coeficientes a_{ij} son conocidos como saturaciones y su cuadrado denota la contribución del factor $\boldsymbol{F}_j$ a la variabilidad total de $\boldsymbol{x}_j$ y los coeficientes d_j, conocidos como unicidades, denotan la contribución del factor único $\boldsymbol{U}_j$ a la variable respectiva. Además, las puntuaciones f_{iq} son las respectivas del individuo i en la dimensión q. Matricialmente:

$$\boldsymbol{X} = \boldsymbol{AF} + \boldsymbol{DU}$$

con $\boldsymbol{X} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_J), \boldsymbol{F} = (\boldsymbol{F}_1, \dots, \boldsymbol{F}_Q)$, $\boldsymbol{U} = (\boldsymbol{U}_1, \dots, \boldsymbol{U}_J), \boldsymbol{A} = (a_{ij})$ y $\boldsymbol{D} = \mathrm{diag}(d_i)$.

1.3.2 Análisis de Componentes Principales

El Análisis de Componentes Principales (*Principal Component Analysis*, PCA) (Jolliffe & Cadima, 2016) es un método clásico de reducción de la dimensión, pre-procesamiento de datos y extracción de características relevantes, ampliamente utilizado en ingeniería, ciencias sociales o biología (Benigni & Giuliani, 1994; Deth, Montero, & Westholm, 2007; Paul & Al Sumam, 2012; Sanguansat, 2012). Su versatilidad se debe a que no tiene en cuenta el objetivo de los datos, lo que facilita su uso en áreas del conocimiento muy dispares. El objetivo principal del PCA consiste en identificar y extraer Q variables latentes (conocidas como componentes principales, PCs), y transformar las J

variables originales correlacionadas en nuevas variables no correlacionadas, con $Q \ll J$, conociendo con ellas el comportamiento de la muestra en un espacio de baja dimensión.

Según la literatura, puede decirse que la primera publicación científica relacionada con el PCA fue realizada por Pearson (1901), que propuso la búsqueda PCs mediante la búsqueda del subespacio de mejor ajuste en el sentido de los mínimos cuadrados. Más tarde, en los años treinta, Hotelling (1933) presentó la idea de definir las componentes principales buscando aquella combinación lineal que absorba la máxima variabilidad posible. Este evidenció que las cargas asociadas a las componentes principales eran justamente los vectores propios de la matriz de covarianzas de la matriz original.

Dada la matriz $\boldsymbol{X}_{IxJ}$, se define la q −ésima componente principal como una combinación lineal de las J variables observadas $\boldsymbol{x}_q$, con $j = 1, \dots, J$, así:

$$\boldsymbol{y}_q = v_{1q}\boldsymbol{x}_1 + \cdots + v_{Jq}\boldsymbol{x}_J$$

En forma matricial:

$$\boldsymbol{Y}_{IxJ} = \boldsymbol{X}_{IxJ}\,\boldsymbol{V}_{JxJ}$$

donde $\boldsymbol{Y}_{IxJ}$ es la matriz que contiene las puntuaciones de cada uno de los individuos sobre el espacio de las nuevas variables y $\boldsymbol{V}_{JxJ}$, conocida como matriz de cargas, es la matriz de proyección en el nuevo espacio que contiene en columnas los coeficientes de cada combinación lineal. Las cargas denotan la contribución de las variables iniciales a cada una de las PCs y son las encargadas, por tanto, de darles un significado físico a los nuevos ejes.

El éxito del PCA se debe a que las PCs son no correlacionadas y capturan secuencialmente la máxima variabilidad entre las columnas de $\boldsymbol{X}$, lo que garantiza que haya una mínima pérdida de información al desechar la información de las ultimas PCs que son las que menos variabilidad aportan al modelo global. Por esto, la reducción de la dimensionalidad se logra al descartar aquellas componentes que absorben menor información y reteniendo el subespacio generado por las primeras $Q < J$ PCs.

$$\boldsymbol{Y}_{IxQ} = \boldsymbol{X}_{IxJ}\,\boldsymbol{V}_{JxQ}$$

Este procedimiento se define, en el fondo, como la rotación del espacio J-dimensional original, seguido de la proyección en el subespacio Q-dimensional que menor distorsión provoca sobre los datos originales (Trendafilov, Unkel, & Krzanowski, 2013).

El enfoque de cálculo de las PCs proporcionado por Hotelling (1933) trata de buscar los nuevos ejes en base a que sean no correlacionados y de manera que la varianza explicada por ellos sea máxima. Este enfoque plantea el problema de optimización siguiente para la búsqueda de las cargas óptimas:

$$\begin{matrix}\max \operatorname{tr}(\operatorname{Cov}_{\boldsymbol{Y}}) \\ \text{s. a.}\, \boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}\end{matrix} \leftrightarrow \begin{matrix}\max \operatorname{tr}(\boldsymbol{Y}\boldsymbol{Y}^T) \\ \text{s. a.}\, \boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}\end{matrix} \leftrightarrow \begin{matrix}\max ||\boldsymbol{Y}||^2 \\ \text{s. a.}\, \boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}\end{matrix}$$

El planteamiento del problema seguido por Pearson tenía como fin estimar la matriz de cargas que define las PCs a partir de un problema de minimización del error de reconstrucción de $\boldsymbol{X}$,

$$\begin{matrix}\min \left\| \boldsymbol{X} - \widehat{\boldsymbol{X}}_Q \right\|^2 \\ \text{s. a.}\, \boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}\end{matrix}$$

siendo $\widehat{\boldsymbol{X}}_Q$ la matriz de coordenadas de las proyecciones sobre el subespacio de las componentes, en el sistema de referencia original.

En 1936, el desarrollo de la SVD 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. Las puntuaciones de los individuos en el nuevo subespacio generado en el PCA vienen dadas por la expresión matricial:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{V} \tag{1.2}$$

Utilizando esta factorización puede verse, dado que la matriz V verifica ser ortonormal:

$$\boldsymbol{Y} = (\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\boldsymbol{T}})\boldsymbol{V} = \boldsymbol{U}\boldsymbol{D}\,(\boldsymbol{V}^{\boldsymbol{T}}\boldsymbol{V}) = \boldsymbol{U}\boldsymbol{D}\boldsymbol{I} = \boldsymbol{U}\boldsymbol{D} \tag{1.3}$$

Así, en la ecuación (1.2) las puntuaciones de los individuos en los nuevos ejes son generadas por la proyección de los datos de entrada en la dirección de los vectores columna $\boldsymbol{v_j}$, mientras que en la ecuación (1.3) se obtienen a partir

de la proyección de los vectores propios a izquierda $\boldsymbol{u_i}$, escalados por los valores singulares λ_i.

1.3.2 Métodos Biplot

Uno de los principales propósitos en el análisis de datos es obtener representaciones gráficas que permitan establecer patrones de comportamiento de las observaciones. El PCA es uno de los métodos utilizados para esto. Sin embargo, permite representar sólo las observaciones o las variables del estudio, pero no ambas a la vez. En otras palabras, representa grupos de observaciones similares, pero no la causa de dicha agrupación. Para evitar estas desventajas, aparecen los métodos de representación Biplot.

Los métodos Biplot son técnicas de reducción de la dimensión y representación simultánea de observaciones y variables en un espacio de baja dimensión donde las interrelaciones entre ellos son capturadas. Permiten representar observaciones y variables conjuntamente en un mismo gráfico interpretable, como si de un gráfico de dispersión multivariante se tratase, con buena calidad de representación para columnas o filas, respectivamente. Gabriel (1971) propone los conocidos como métodos Biplot clásicos: JK-Biplot y GH-Biplot. Se trata de técnicas de factorización matricial que permiten representar observaciones y variables conjuntamente a partir de matrices de marcadores fila y columna. Más tarde, Galindo-Villardón (1986) desarrolla el HJ-Biplot, técnica que recoge las ideas de los Biplot de Gabriel (1971), con el objetivo principal de representar en el mismo sistema de referencia gráfico tanto las unidades como las variables del conjunto de datos, ambas interpretables con buena calidad de representación (Figura 7). Para ello, modifica la definición de las matrices de marcadores para filas y columnas consideradas anteriormente por Gabriel (1971).

El HJ-Biplot ha sido, y sigue siendo, aplicado con éxito en múltiples disciplinas como economía (Amor-Esteban, García-Sánchez, & Galindo-Villardón, 2017; Ortas, Álvarez, Jaussaud, & Garayar, 2015), genética (Frutos, Galindo, & Leiva, 2014), ciencias marinas (Mendes et al., 2009) y calidad del agua (Carrasco et al., 2019), bibliometría (Díaz-Faes, González-Albo, Galindo, & Bordons, 2013; Torres-Salinas, Robinson-García, Jiménez-Contreras, Herrera,

& López-Cózar, 2013), estudios de seguridad (Vázquez, Vicente-Galindo, & Galindo, 2011), y muchos más.

Los métodos Biplot reducen la dimensionalidad de las matrices de datos haciendo uso de las componentes del PCA y, como ocurre en él, las dimensiones latentes se obtienen de tal forma que son combinaciones lineales de todas las variables de partida y además no estén correlacionadas; esto es, que sean ortogonales. Esto facilita la interpretación de las componentes, así como el cálculo de la variabilidad explicada por cada uno de los ejes retenidos.

Formulación teórica: JK, GH y HJ-Biplot.

Un Biplot es una representación gráfica de una matriz mediante dos matrices de marcadores fila y columnas elegidas de manera que su producto aproxime el elemento original de la matriz de datos de la mejor forma posible. En otras palabras, la base de datos original $\boldsymbol{X} \in \mathbb{R}^{IxJ}$ es descompuesta como el producto de dos matrices $\boldsymbol{A} \in \mathbb{R}^{IxK}$ y $\boldsymbol{B} \in \mathbb{R}^{KxJ}$ de marcadores fila y columna respectivamente:

$$\boldsymbol{X} \approx \boldsymbol{A}\boldsymbol{B}^{T}$$

donde $\boldsymbol{A} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_I)$ y $\boldsymbol{B} = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_J)$ dos matrices de marcadores tales que el producto $\boldsymbol{x}_{ij} \approx \boldsymbol{a}_i^T \boldsymbol{b}_j$. Las coordenadas fila de la matriz $\boldsymbol{A}$ se corresponden con los puntos que representan a las observaciones de la matriz $\boldsymbol{X}$**,** y las coordenadas de los vectores columna de $\boldsymbol{B}$ representan las variables originales en un mismo espacio de menor dimensión formado por componentes latentes (PCs) ortogonales. La factorización Biplot de la matriz $\boldsymbol{X}$ en el producto de dos matrices $\boldsymbol{A}$ y $\boldsymbol{B}$ puede llevarse a cabo mediante métodos de descomposición matricial. Habitualmente, la factorización escogida es la SVD pues proporciona una solución única. Sea la SVD de la base de datos original:

$$\boldsymbol{X} \approx \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\boldsymbol{T}}$$

donde $\boldsymbol{U}$ y $\boldsymbol{V}$ son matrices ortonormales y $\boldsymbol{D}$ la matriz diagonal de valores singulares $\lambda_1, \ldots, \lambda_Q$ de $\boldsymbol{X}$.

La formulación teórica de las matrices de marcadores fila y columna para cada uno de los métodos Biplot clásicos (GH, JK, HJ) a partir de la SVD queda

resumida en la Figura 7. Las distintas combinaciones de las matrices $\boldsymbol{U}, \boldsymbol{D}$ y $\boldsymbol{V}$ dan lugar a los tipos de Biplot más comunes. El JK-Biplot establece los marcadores fila y columna como $\boldsymbol{A} = \boldsymbol{UD}$ y $\boldsymbol{B} = \boldsymbol{V}$, el JK-Biplot como $\boldsymbol{A} = \boldsymbol{U}$ y $\boldsymbol{B} = \boldsymbol{VD}$ y por último el HJ-Biplot los establece como $\boldsymbol{A} = \boldsymbol{UD}$ y $\boldsymbol{B} = \boldsymbol{DV}$**.** Es inmediato comprobar que tanto en el JK-Biplot como en el GH-Biplot el producto $\boldsymbol{AB}^T$ reproduce la matriz original $\boldsymbol{X}$**.** Esto no es así en el caso del HJ-Biplot pues a cambio de lograr máxima calidad de representación para filas y columnas de una matriz, el producto $\boldsymbol{AB}^T$ no reproduce la matriz de partida sino que $\boldsymbol{X} \approx \boldsymbol{AD}^{-1}\boldsymbol{B}^T$.

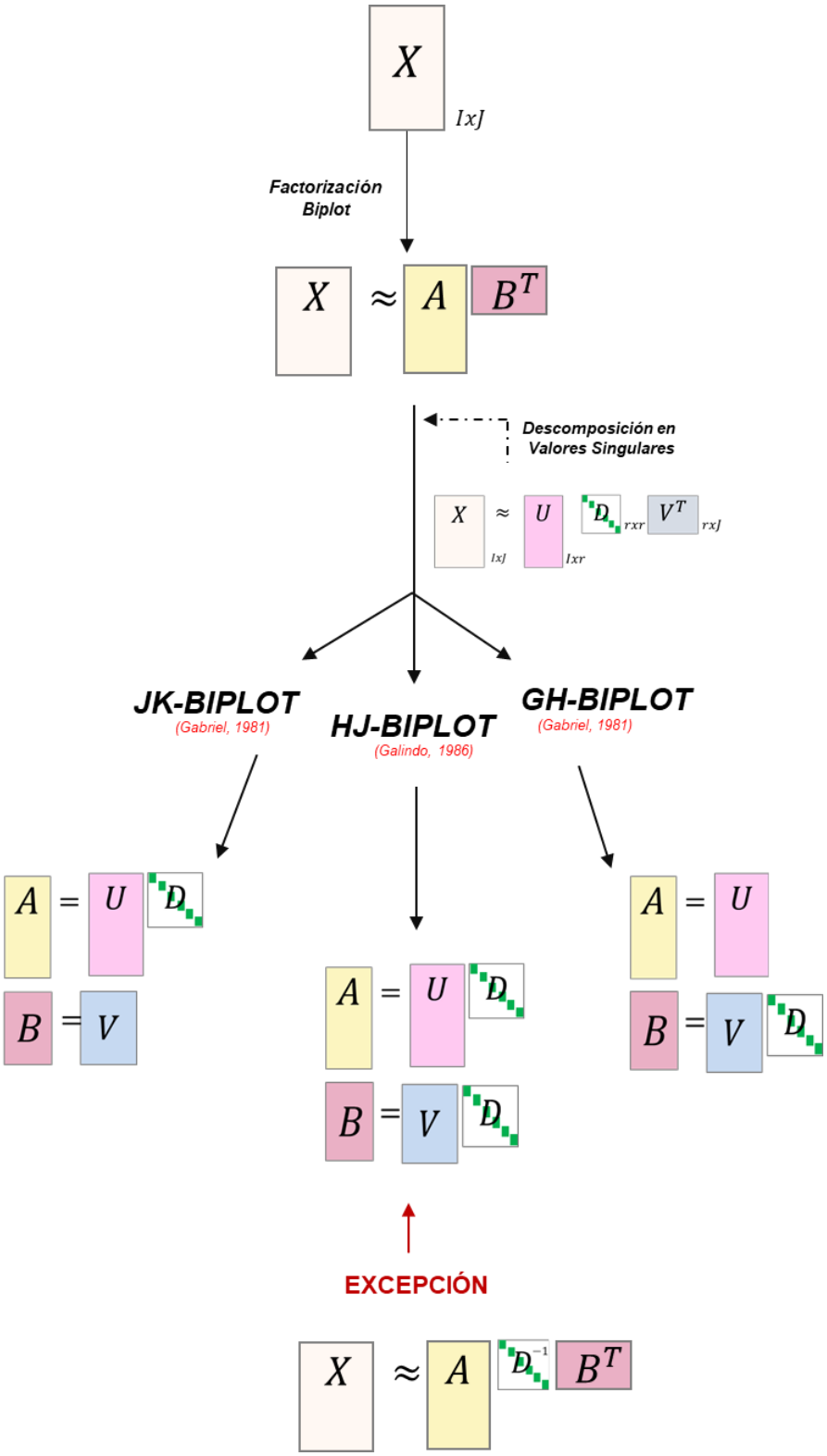


***Figura 7**. Factorización Biplot y definición de las metodologías JK-Biplot, GH-Biplot y HJ-Biplot*

El lector puede acudir a (Nieto-Librero, 2015) para leer una revisión exhaustiva de los métodos Biplot y de las propiedades de sus marcadores.

Existen softwares específicos para la ejecución de los métodos Biplot. Si se desea implementar los métodos Biplot clásicos o métodos Biplot con remuestreo Bootstrap para analizar la estabilidad de los resultados puede utilizarse el paquete BiplotbootGUI (*Bootstrap on Classical Biplots and Clustering Disjoint Biplot*), desarrollado por (Nieto, Galindo-Villardón, Leiva, & Vicente-Galindo, 2014) y disponible en el software R. Otra librería en R para implementar los Biplot clásicos, canónico o logístico externo es MultBiplotR. Existen en R otras librerías que incorporan funciones para ejecutar distintos tipos de Biplot como el paquete GGEBiplotGUI para Biplots de interacción genotipo/ambiente (Frutos & Galindo, 2014) o dynBiplotGUI para Biplots dinámicos (Egido, 2017).

Interpretación geométrica y bondad de ajuste.

Una apropiada interpretación de los resultados de los métodos Biplot clásicos lleva implícita el seguimiento de una serie de pautas (Figura 8). En primer lugar, los marcadores fila $\boldsymbol{a}_i$ se representan como puntos y los marcadores columna $\boldsymbol{b}_j$ como vectores. Esta representación permitirá estudiar las relaciones entre observaciones, variables, y entre ambos como se describe a continuación:

- La distancia entre puntos se interpreta en términos de similaridad entre observaciones, de manera que marcadores fila cercanos en el gráfico harán referencia a observaciones con patrones de comportamiento comunes. Las observaciones más alejadas (con mayor distancia euclídea entre ellas) apuntan a unidades con comportamientos diferentes con respecto a las variables consideradas.
- La longitud del vector aproxima la variabilidad de la variable correspondiente. A mayor longitud, mayor variabilidad. Cuanto más cercano al origen se encuentre un marcador columna, menor variabilidad presenta en dicho plano factorial. En consecuencia, pueden no estar bien representadas en las dimensiones latentes consideradas.
- La covariación entre variables es medida a partir de los ángulos entre los vectores de marcadores columna correspondientes. En este sentido, si los ángulos son agudos las variables correspondientes se

encontrarán relacionadas de manera directa, con una mayor relación cuanto menor sea el ángulo entre ambas. Si el ángulo es obtuso, las variables estarán correlacionadas de manera inversa y en el caso de ángulos rectos, las variables correspondientes serán independientes linealmente.

- La dirección del vector o marcador columna apunta el gradiente en el que el nivel de la variable aumenta.
- La covariación entre observaciones y variables es medida a partir del producto escalar de sus marcadores. Por la definición del producto escalar, la relación entre observaciones/variables se cuantifica como la proyección ortogonal de los marcadores fila sobre los marcadores columna correspondientes. Así es posible establecer grupos de observaciones con comportamientos similares. En el caso de los métodos GH y JK, la proyección aproxima el valor x_{ij} original. Esto no es así en el caso del HJ-Biplot.
- La proporción de variabilidad de cada una de las variables originales explicada por el factor latente se conoce como contribución relativa del factor al elemento. Por otro lado, las contribuciones del elemento al factor permiten conocer las variables responsables de la formación de los ejes. Gráficamente, esto puede traducirse a partir de los ángulos generados entre los vectores de los marcadores columna y los ejes factoriales.

La cantidad de variabilidad total explicada por el modelo de componentes latentes retenidas viene dada por la división entre la suma de los Q primeros valores propios entre la suma de todos los valores propios. Cuanto mayor sea la variabilidad explicada por los primeros ejes, mejor representada estará la estructura de la matriz de datos en las primeras dimensiones.

Anteriormente se ha mencionado como la calidad de representación de filas y columnas difiere en un tipo de Biplot u otro. Así, la calidad de representación (que es una medida del ajuste global de la aproximación) en filas para el JK-Biplot es máxima, mientras que en el caso del GH-Biplot lo es la calidad de representación para columnas, pero no para filas. En el caso del HJ-Biplot ambas

calidades son buenas. En la Tabla 2 se resumen la calidad de representación en cada tipo de Biplot.

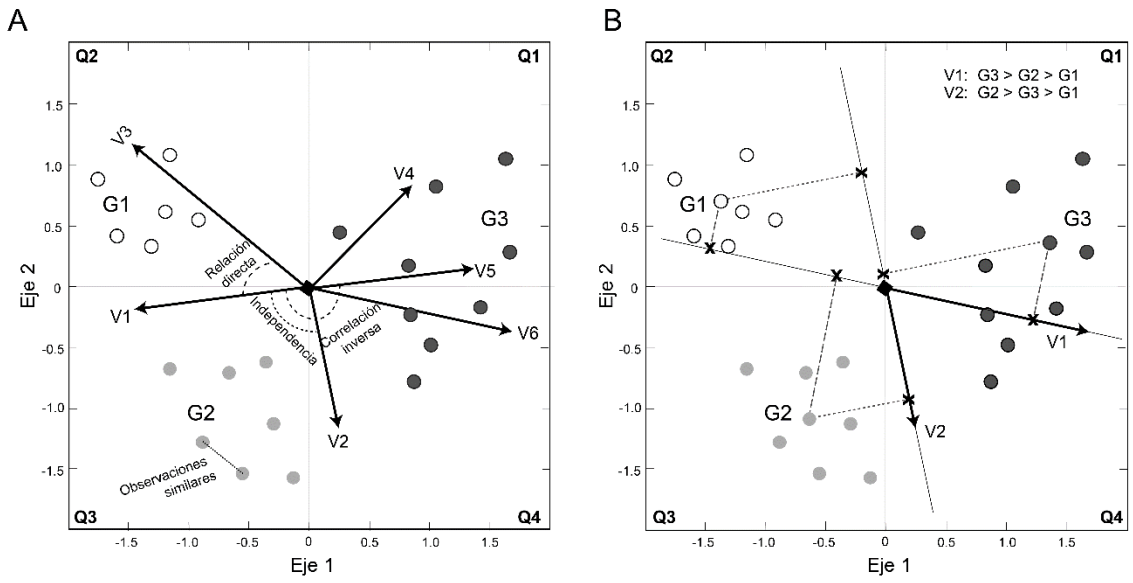


***Figura 8**. Interpretación Biplot para visualizar las diferencias entre distintas observaciones de una matriz multivariante. En el panel A las distancias entre puntos reflejan la similitud entre observaciones y los vectores representan las variables. En el panel B se muestran las diferencias para diferentes observaciones a través de la proyección perpendicular del punto sobre el vector*

***Tabla 2**. Calidades de representación de los marcadores fila y columna en el GH, JK, y HJ-Biplot*

	Calidad filas	**Calidad columnas**
GH-Biplot	$\frac{Q}{J}$	$\frac{\sum_{q=1}^{Q}\lambda_q^4}{\sum_{q=1}^{J}\lambda_q^4}$
JK-Biplot	$\frac{\sum_{q=1}^{Q}\lambda_q^4}{\sum_{q=1}^{J}\lambda_q^4}$	$\frac{Q}{J}$
HJ-Biplot	$\frac{\sum_{q=1}^{Q}\lambda_q^4}{\sum_{q=1}^{J}\lambda_q^4}$	$\frac{\sum_{q=1}^{Q}\lambda_q^4}{\sum_{q=1}^{J}\lambda_q^4}$

Por último, si lo que se desea evaluar es la calidad de ajuste de individuos y variables de manera independiente, se debe examinar las contribuciones relativas y absolutas de los elementos (Tabla 3). Las contribuciones relativas del elemento al factor sirven para examinar en qué medida el factor puede ser explicado por el elemento (observación/variable), mientras que las contribuciones relativas del factor al elemento evalúan la posible relación del factor con el elemento.

***Tabla 3**. Contribuciones absolutas y relativas de los individuos y variables*

Contribución	
Contribución absoluta de los individuos a la varianza del factor q	$CAE_i F_q = a_{iq}^2$
Contribución absoluta de las variables a la varianza del factor q	$CAE_j F_q = b_{jq}^2$
Contribución relativa del elemento i al factor q	$CRE_i F_q = CAE_i F_q / \sum_{i=1}^{I} a_{iq}^2$
Contribución relativa del elemento j al factor q	$CRE_j F_q = CAE_j F_q / \sum_{j=1}^{J} b_{iq}^2$
Contribución relativa del factor q al elemento i	$CRF_q E_i = a_{iq}^2 / d^2(a_i, 0)$
Contribución relativa del factor q al elemento j	$CRF_q E_j = b_{jq}^2 / d^2(b_j, 0)$

1.4 Contribuciones al análisis de escalas psicométricas: una aplicación en educación

1.4.1 Análisis de la actitud y enfoques de aprendizaje en estudiantes universitarios

En los últimos años se ha producido un cambio en el paradigma de la educación universitaria según el Espacio Europeo de Educación Superior en el sentido de la enseñanza-aprendizaje, con el alumno posicionado en el eje central de todo el proceso y asumiendo que estos no son meros receptores del conocimiento. A raíz de ello, el concepto teórico de la actividad docente de los profesores universitarios se ha visto modificada, ganando importancia las habilidades para diseñar e impartir docencia adaptada a los estudiantes, en lugar de valorar únicamente los conocimientos del usuario sobre la asignatura. Además, en todo proceso educativo no sólo prima la repercusión docente en el desarrollo de las habilidades de los estudiantes, sino también la actitud adoptada por ellos al enfrentarse al estudio de una nueva materia. Asimismo, la predisposición del alumno constituye un elemento primordial en este proceso, puesto que cuando esta es favorable están motivados a realizar esfuerzos para aprender y afianzar sus conocimientos.

Actitud. La actitud es la predisposición del individuo para responder de manera favorable o desfavorable a un determinado objeto o situación (Schwarz &

Bohner, 2001). Una actitud negativa bloqueará su uso en el futuro profesional de los estudiantes, mientras que una respuesta positiva ayudará al alumno a sentirse seguro en su trabajo, a creer y confiar en sus habilidades para poder enfrentarse a diversas situaciones y a estar motivados para conseguir los objetivos que se planteen. Las definiciones clásicas coinciden en identificar tres factores relevantes que componen la actitud de una persona: lo que se siente (el factor afectivo), lo que se piensa (el factor cognitivo) y lo que se dice o hace (que constituiría el factor conductual) (Muñoz & Mato, 2008; Triandis, 1971)(Figura 9).

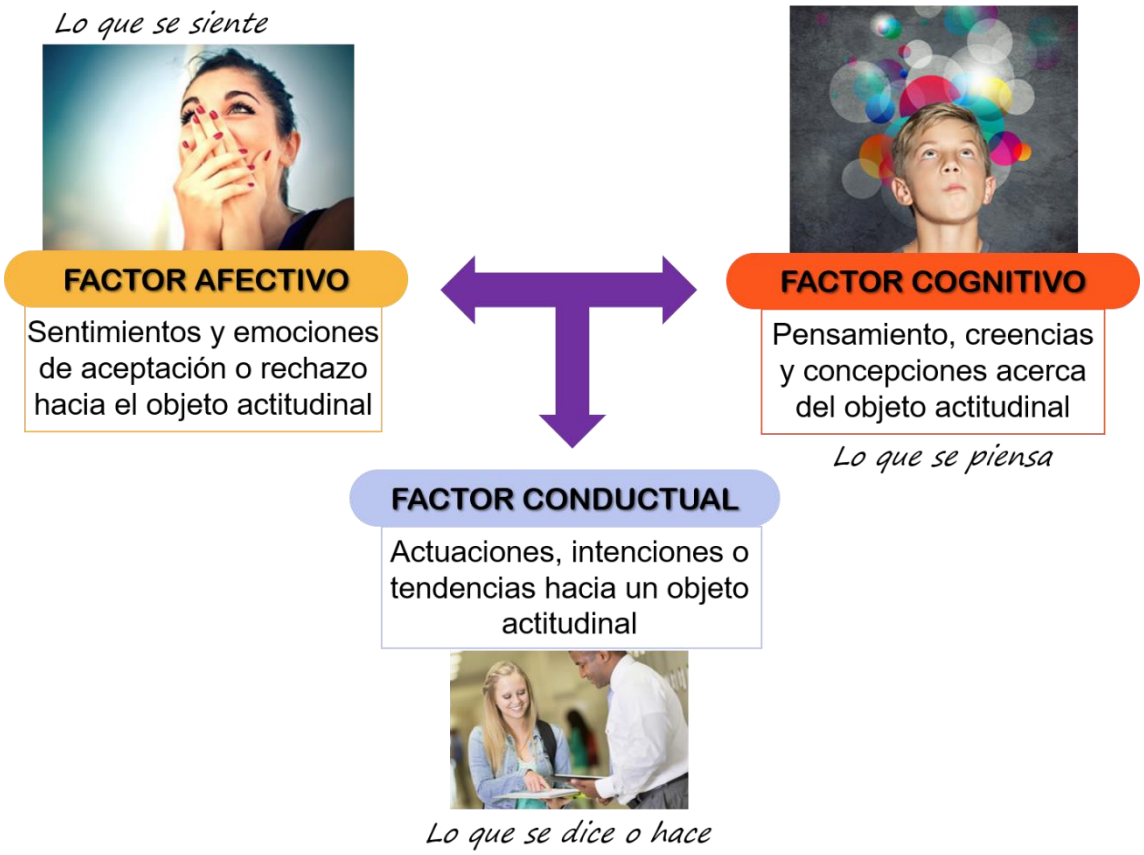


***Figura 9**. Modelo multidimensional de la actitud*

Enfoques de aprendizaje. Un enfoque de aprendizaje puede entenderse como la forma en que el estudiante se enfrenta a la tarea de aprendizaje. Apoyan la creencia que defiende que en la educación de un estudiante no solo es importante qué se aprende, sino también cómo se aprende. En este sentido uno de los modelos que ha suscitado mayor interés es el *Students Approaches Learning* (SAL) de Marton & Säljö (1976), en el cual se introducen por primera vez los conceptos de enfoques de aprendizaje superficial y profundo (Figura 10).

El enfoque profundo corresponde a estudiantes que manifiestan una elevada motivación y un alto nivel de implicación en relación con lo que están aprendiendo. Por otro lado, el enfoque superficial describe a los estudiantes que tienden a cumplir los requisitos mínimos para aprobar las pruebas de evaluación con mínimo esfuerzo (Witriw, Molina, & Ferrari, 2014). Emplean estrategias dirigidas a aprender de forma mecánica y repetitiva. Es evidente que la adquisición de un mayor conocimiento está íntimamente ligada con el enfoque que los alumnos utilizan a la hora de aprender; además, no todos los estudiantes captan la información de la misma manera y, debido a ello, no todos aprenden igual.

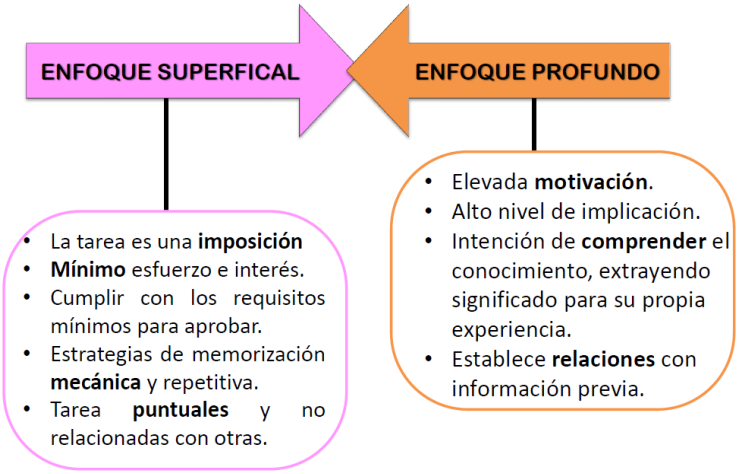


***Figura 10**. Características principales de los enfoques de estudio superficial y profundo, siguiendo los modelos planteados por Marton y Säljö (1976) y Biggs (1987)*

Asignatura. El análisis de la relación entre actitud y modo de aprendizaje del alumnado universitario se llevó a cabo en el ámbito de la materia “Didáctica General”, asignatura problemática por su contenido en la carrera universitaria. En el año 1969, Nérici definió la didáctica como “el estudio del conjunto de recursos técnicos que tienen por finalidad dirigir el aprendizaje del alumno, con el objeto de llevarlo a alcanzar un estado de madurez que le permita encarar la realidad de manera consciente, eficiente y responsable, para actuar en ella como ciudadano participante y responsable” (Nerici, 1969).

Diversos estudios han analizado la influencia de factores externos como el género, el rendimiento académico o los métodos docentes en la actitud. Pero hasta el momento no existía ninguna investigación que relacionara actitud y

enfoques de aprendizaje en dicho ámbito. La existencia de poca literatura al respecto puede deberse a la relativa reciente incorporación del grado en Educación Social al sistema universitario español.

Instrumentos utilizados. Los instrumentos utilizados para evaluar la actitud y los enfoques de aprendizaje fueron:

1. *Cuestionario de Medición de la Actitud hacia la Estadística*. Este cuestionario surge con el objetivo de captar la estructura latente de los cuestionarios ATS (*Attitudes Toward STATIStics*) de 29 ítems (Wise, 1985) y EAE (Escala de Actitud hacia la Estadística) 25 ítems de Auzmendi (1992), dos de las escalas más utilizadas y que muestran la relación de la actitud hacia esta disciplina y los resultados académicos o rendimiento. El cuestionario de medición de la actitud hacia la Estadística es propuesto por Mondéjar Jimenez, Vargas, & Bayot Mestre (2008) y está compuesto por 27 ítems, medidos en escala Likert de 5 categorías, que plantean situaciones en torno al dominio afectivo y valorativo de los estudiantes (véase el Anexo 1). Los 27 ítems se distribuyen en torno a una estructura factorial de cuatro dimensiones latentes: interés y ansiedad, que conforman la escala afectiva del cuestionario, utilidad presente y utilidad profesional, que conforman la escala cognitiva o valorativa. Los enunciados de los ítems se adaptaron a la asignatura considerada.

2. *Cuestionario de Procesos de Estudio Revisado de Dos Factores (Revised two-factor Study Process Questionnaire, R-SPQ-2F)*. Los enfoques de aprendizaje de los estudiantes son evaluados mediante la versión validada al castellano y adaptada al ámbito universitario por (Hernández-Pina, García-Sanz, & Maquilón-Sánchez, 2004) del cuestionario R-SPQ-2F (Biggs, Kember & Leung, 2001), que surge como revisión del *Cuestionario de Procesos de Estudio* (*Study Process Questionnaire, SPQ*) (Biggs, 1987a; Biggs, 1987b). Evalúa las actuaciones de los estudiantes a diversas situaciones en relación con su proceso de aprendizaje, analizando su grado de conformidad con 20 ítems, medidos en escala Likert de 5 categorías de respuesta (Anexo 1). El instrumento está compuesto por 2 subescalas latentes: aprendizaje profundo (con ítems como “Me doy cuenta de que

estudiar me proporciona un sentimiento de profunda satisfacción personal") y superficial (por ejemplo, "Creo que los profesores no deberían esperar que los alumnos dedicaran mucho tiempo a estudiar cosas que no van a caer en el examen"), cada una de ellas definida por un total de 10 ítems (Tabla 4).

***Tabla 4**. Estructura factorial del cuestionario de Medición de la Actitud (Mondéjar et al. 2008) y R-SPQ-2F (Biggs et al. 2001)*

Instrumento	Constructo	Definición	Ítems
Medición de la actitud hacia la materia	Interés	Agrado y satisfacción de los alumnos con la materia	13,14,15,17, 18,24
	Ansiedad	Nerviosismo o temor ante la utilización de esta materia	1,7,9,12,21, 22,23
	Utilidad presente	Valor que los alumnos atribuyen a la materia dentro de sus estudios	3,10,16,25
	Utilidad profesional	Percepción de utilidad de la materia para su futuro profesional	2,4,5,6,11,19 ,20,26,27
R-SPQ-2F	Aprendizaje profundo	Enfoque profundo de estudio, con mayor interés y esfuerzo por parte del alumno	1,2,5,6,9,10, 13,14,17,18
	Aprendizaje superficial	Planteamiento más superficial del estudio, enfocado a superar la asignatura realizando el mínimo esfuerzo	3,4,7,8,11,12 ,15,16,19,20

Análisis de datos. La consistencia interna de ambos cuestionarios se midió mediante el coeficiente α de Cronbach para cada una de las dimensiones del instrumento utilizado. En la aplicación 1, estos resultados se compararán con otras medidas, como el índice ω (McDonald, 1999). En cuanto a la validez factorial de los cuestionarios, en primer lugar se realizó un estudio de las correlaciones entre ítems de cada uno de los cuestionarios por separado y un FA exploratorio de los datos a partir del FA con rotación Varimax. En segundo lugar, se confirmó la estructura de ambas escalas mediante el análisis factorial confirmatorio (*Confirmatory Factor Analysis*, CFA por sus siglas en inglés) mediante el método de máxima verosimilitud, valorando el ajuste de los modelos haciendo uso de las medidas RMSEA (*Root Mean Square Error of Approximation*), CFI (*Comparative Fit Index*), SRMR (*Standardized Root Mean*

Square Residual), CF (C*omposite Fiability Index*) y AVE (A*verage Variance Extracted*).

Los estudiantes fueron clasificados de acuerdo a su enfoque de aprendizaje predominante, siguiendo la metodología propuesta en (Hernández-Pina, Rodríguez, Ruiz, & Esquivel, 2010). Se obtuvo una puntuación de enfoque de aprendizaje profundo y superficial a partir de la combinación lineal de los ítems que conforman cada uno de los factores; esto es, la suma de sus puntuaciones. De esta manera, el rango para cualquiera de las subescalas oscilaba entre 10 y 50 puntos, puesto que todos los ítems son puntuados en la misma dirección. Esta fue posteriormente categorizada según tres niveles: aprendizaje profundo, superficial y mixto. Se consideró que el estudiante adoptaba aquel enfoque de aprendizaje para el cual hubiera obtenido una mayor puntuación, definiéndose el enfoque mixto para aquellos individuos que obtenían la misma puntuación en ambos factores (Hernández-Pina et al., 2010).

Se realizó un análisis descriptivo de cada uno de los ítems para conocer la actitud y el modo de aprendizaje de los alumnos universitarios. Con el fin de profundizar en la comprensión de la relación entre las distintas actitudes de los universitarios con base en el enfoque de aprendizaje, se llevó a cabo un análisis multivariante mediante HJ-Biplot. Posteriormente, estos resultados se combinaron con el análisis de clúster jerárquico por el método de Ward sobre las coordenadas Biplot obtenidas para describir patrones de estudiantes con un mismo perfil multivariante, en sentido actitudinal y de aprendizaje.

Los datos se han analizado utilizando el programa IBM SPSS STATIStical Package, versión 23,0. El CFA se lleva a cabo mediante el módulo AMOS, versión 23,0, de IBM SPSS (Arbuckle, 2014) y el análisis HJ-Biplot y el análisis de clúster se realizan a través de la librería BiplotbootGUI (Nieto et al., 2014) del software libre R (Team, 2019).

Participantes. La muestra de participantes estaba formada por 146 estudiantes, con una edad media de 21±2,65 años y mayoritariamente por mujeres (n=129, 88%), estudiantes del Grado de Educación Social de la Universidad de Salamanca matriculados en la asignatura entre los años 2011 y 2014. Se trata de una materia obligatoria, semestral, cursada en el segundo año del Grado en

Educación Social. Se recogió su respuesta a un total de 74 ítems, 20 respectivos a la escala R-SPQ-2F y 27 de la escala de medición de la actitud, estos últimos en dos ocasiones diferentes. Previamente a la recogida de datos del cuestionario de medición de la actitud es importante mencionar que se realizó una adaptación del cuestionario original de Mondéjar Jimenez, Vargas y Bayot Mestre (2008) al ámbito de la Didáctica General. Se mantuvieron los 27 ítems propuestos originalmente por los autores, pero adaptando los epígrafes a esta asignatura y se realizaron dos mediciones del mismo al inicio de la asignatura y tras la docencia recibida para poder examinar cambios en las puntuaciones en la actitud hacia la materia.

Resultados. Los resultados del análisis de consistencia interna (Tabla 5) reflejan índices de fiabilidad adecuados de las dimensiones del R-SPQ-2F e índices más fuertes en el caso del cuestionario de actitud tanto al inicio de la asignatura como tras la docencia recibida, a excepción de la dimensión de la utilidad presente.

Tabla 5*. Consistencia interna de los cuestionarios de Medición de la Actitud hacia la Didáctica General y R-SPQ-2F*

Medición	Autor	Número de ítems	α de Cronbach	Índice ω
Pre	Interés	6	0,81	0,82
Pre	Ansiedad	7	0,84	0,84
Pre	Utilidad presente	4	0,57	0,63
Pre	Utilidad profesional	9	0,78	0,79
Post	Interés	6	0,85	0,85
Post	Ansiedad	7	0,84	0,84
Post	Utilidad presente	4	0,66	0,66
Post	Utilidad profesional	9	0,83	0,84
	Aprendizaje profundo	10	0,72	0,74
	Aprendizaje superficial	10	0,71	0,71

La medida de adecuación muestral KMO (Kaiser-Meyer-Olkin), (0,80 para Pre y 0,86 para Post del cuestionario de Medición de la Actitud hacia la Didáctica; 0,71 del cuestionario R-SPQ-2F) y el test de esfericidad de Barlett estadísticamente significativo (p = 0,000 en todos los casos) reflejan la existencia de relaciones entre variables y factores, siendo las técnicas factoriales aptas

para el análisis de datos. La Figura 11 muestra las correlaciones entre los ítems del cuestionario de medición de la actitud (Verde: utilidad profesional; rojo: ansiedad; azul: interés; morado: utilidad presente). Se observan correlaciones directas entre ítems de una misma dimensión, tanto antes de la docencia (Figura 11, izquierda) como tras ella (Figura 11, derecha), relaciones directas entre ítems de interés y utilidad presente, así como una relación inversa tanto de los ítems de la utilidad presente como del interés con la utilidad profesional. Los ítems de la dimensión ansiedad se muestran relacionados de manera directa, aunque con leves correlaciones, con la utilidad profesional; es decir, aquellos estudiantes con mayores niveles de ansiedad, creen en la utilidad profesional de la asignatura a largo plazo. En el caso del R-SPQ-2F (Figura 12) la correlación entre los ítems de ambas dimensiones es aún más evidente. Los ítems que valoran tanto el aprendizaje profundo como superficial muestran una relación directa y alta entre sí de manera respectiva, relacionándose de manera inversa entre ellos.

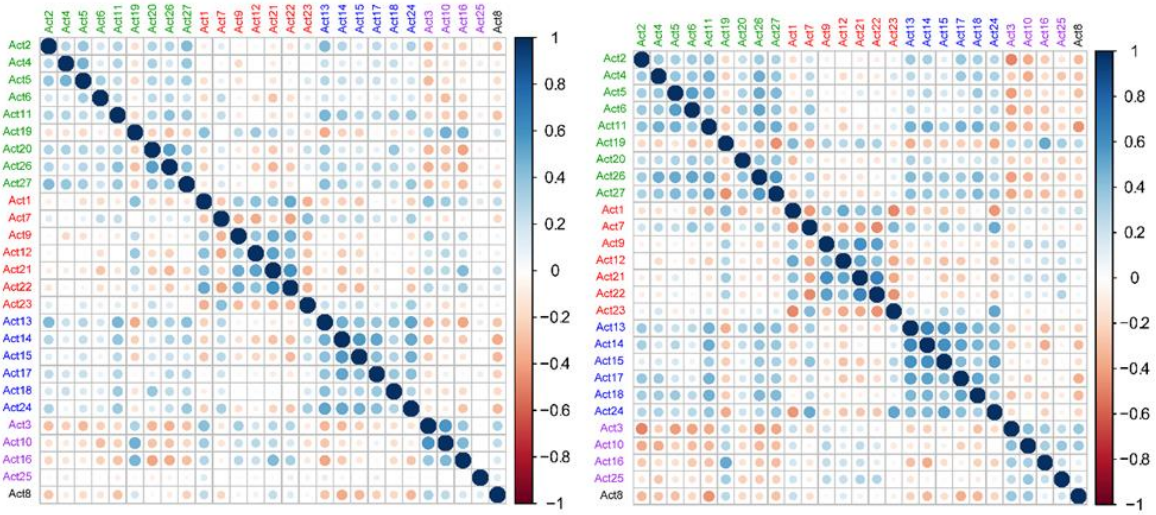


Figura 11. *Análisis de las correlaciones entre ítems del cuestionario de medición de la actitud hacia la didáctica antes de la docencia (izquierda) y después de la docencia (derecha) (Verde: utilidad profesional; rojo: ansiedad; azul: interés; morado: utilidad presente).*

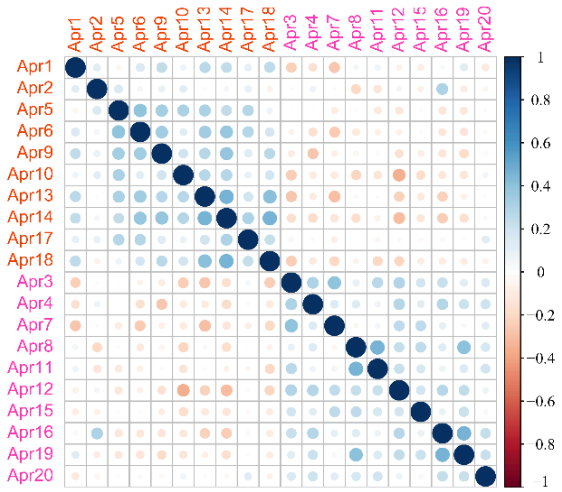


***Figura 12**. Análisis de las correlaciones entre ítems del cuestionario R-SPQ-2F (Naranja: aprendizaje profundo; rosa: aprendizaje superficial).*

El cuestionario de Medición de Actitud hacia la Didáctica revela la estructura latente teórica de 4 factores en el pre y post-test explicando el 40,1% y 46,8% de la varianza total, respectivamente (Tabla 6). Tanto en el momento previo como en el posterior a la docencia, se confirma la existencia de las cuatro dimensiones teóricas: interés (11,5%; 12,3%), ansiedad (10,7%; 12,9%), utilidad presente (7%; 6,9%) y utilidad profesional (10,9%; 14,7%). En el momento pre-test es la dimensión interés la que más variabilidad recoge en la muestra y en el post-test es la utilidad profesional. Por otro lado, el EFA con rotación Varimax sobre el cuestionario R-SPQ-2F permite vislumbrar la estructura bidimensional de estudio profundo (13,1%) y estudio superficial (10,5%), absorbiendo un 23,6% de la variabilidad global (Tabla 7). En ambos casos se observan un porcentaje de varianza bajo absorbido por cada uno de los ejes retenidos.

***Tabla 6**. Matriz de cargas factoriales del cuestionario Medición de la Actitud hacia la Didáctica por dimensiones: interés, ansiedad, utilidad presente, utilidad profesional obtenida mediante el FA con rotación Varimax y umbralización de 0,3*

Ítems	**Pre**				**Post**			
	F1	**F2**	**F3**	**F4**	**F1**	**F2**	**F3**	**F4**
Interés								
Act13	0,62	0,38					0,69	
Act14	0,69						0,83	
Act15	0,64					-0,31	0,70	
Act17	0,60						0,61	
Act18	0,41	0,32			0,38		0,50	
Act24	0,67				0,30	-0,49	0,46	
Ansiedad								
Act1			0,50	0,42		0,61		
Act7			-0,47		0,35	-0,62		
Act9			0,60			0,53		0,43
Act12			0,60			0,57		
Act21			0,69			0,70		0,43
Act22			0,74			0,72		
Act23			-0,47			-0,67		
Utilidad presente								
Act3		-0,36		0,48	-0,53			0,36
Act10				0,70	-0,39			0,57
Act16		-0,34		0,49				0,45
Act25								0,48
Utilidad profesional								
Act2		0,49			0,48			
Act4		0,57			0,54			
Act5		0,54			0,71			
Act6		0,36			0,74			
Act11	0,41	0,43			0,55		0,45	
Act19				0,59		0,33		0,39
Act20		0,62			0,42			
Act26		0,52			0,63		0,31	
Act27		0,59			0,53		0,34	
Act8	-0,43				-0,36			
Varianza explicada (%)	11,5	10,9	10,7	7,0	14,7	12,9	12,3	6,9

Tabla 7. *Matriz de cargas factoriales del cuestionario R-SPQ-2F por dimensiones: estudio profundo y estudio superficial obtenida mediante el FA con rotación Varimax y umbralización de 0,3*

Ítems	**F1**	**F2**
Aprendizaje profundo		
Apr1	0,38	
Apr2		
Apr5	0,43	
Apr6	0,53	
Apr9	0,50	
Apr10	0,33	
Apr13	0,68	
Apr14	0,66	
Apr17	0,40	
Apr18	0,50	
Aprendizaje superficial		
Apr3		0,40
Apr4		0,38
Apr7	-0,34	
Apr8		0,59
Apr11		0,46
Apr12		0,48
Apr15		0,34
Apr16		0,46
Apr19		0,58
Apr20		0,35
Varianza Explicada	16,7	7

La Tabla 8 resume los principales resultados del CFA. El ajuste de cada uno de los modelos fue aceptable. En el caso del constructo actitudinal, la AVE por dimensiones presenta valores medios más altos tras la docencia y los índices FC indican alto nivel de fiabilidad en el momento previo y tras cursar la materia. En el caso del R-PSQ-2F se reportaron altos índices de FC; sin embargo es importante tener en cuenta la baja AVE de ambas dimensiones. Esto debería tenerse en cuenta en investigaciones futuras, llevando a consideración el estudio específico de cada una de las cuestiones de este instrumento.

***Tabla 8**. Consistencia interna de los cuestionarios de Medición de la Actitud hacia la Didáctica General y R-SPQ-2F*

	Autor	χ^2/gl	RMSEA	CFI	SRMR	FC	AVE
Pre	Interés	1,52	0,06 (0,05, 0,07)	0,88	0,07	0,96	0,40
Pre	Ansiedad					0,90	0,41
Pre	Utilidad presente					0,93	0,40
Pre	Utilidad profesional					0,90	0,26
Post	Interés	1,71	0,07 (0,06, 0,08)	0,88	0,08	0,97	0,54
Post	Ansiedad					0,94	0,45
Post	Utilidad presente					0,92	0,32
Post	Utilidad profesional					0,97	0,39
	Aprendizaje profundo	1,31	0,05 (0,03, 0,06)	0,91	0,07	0,88	0,27
	Aprendizaje superficial					0,86	0,19

Se realizó un análisis descriptivo de las respuestas obtenidas para caracterizar la respuesta media de la muestra universitaria los ítems de ambos cuestionarios. Los resultados destacaron un interés medio por la didáctica al comienzo de la asignatura, que se ve disminuido en parte tras cursarla. Algo similar ocurre con los niveles de ansiedad de los alumnos, aumentados al finalizar el semestre. Los alumnos consideran útil esta asignatura tanto al comienzo como tras haber recibido la docencia, hecho acorde con la alta utilidad profesional que siente el alumnado al inicio y al final, donde incluso se ve ampliado.

En cuanto a su enfoque de aprendizaje, los estudiantes presentaron mayor tendencia al estudio de manera profunda. En concordancia con esto, el enfoque predominante resultó ser el enfoque profundo (87%), seguido de aquellos con uno superficial (9,6%) o mixto (3,4%). Este hecho concuerda con lo expuesto por Hernández-Pina et al. (2010), que informan sobre un predominio del enfoque profundo en el contexto universitario y que es percibido como síntoma de buena enseñanza (Prosser & Trigwell, 2014). La nota de acceso a la universidad fue significativamente diferente (p=0,003) entre los alumnos que aprenden de manera profunda y los que aprenden de manera superficial (6,97 y 6,18 puntos, respectivamente). A su vez, los estudiantes que adoptan un enfoque profundo presentaron mayores niveles de interés (p=0,000), mientras que los que estudian de manera superficial sentían mayores niveles de ansiedad (p=0,016). Las mujeres reportaron niveles de ansiedad significativamente mayores antes de recibir la docencia (p=0,014), en consonancia con los

resultados de otras investigaciones (Mondéjar-Jiménez & Vargas-Vargas, 2010). No se encontraron diferencias significativas en ninguna de las dimensiones de ambos cuestionarios en función de la edad de los estudiantes.

El HJ-Biplot parte de la SVD de la matriz de datos estandarizados. La representación de los ejes factoriales 1-2 caracteriza la relación entre los ítems de las distintas dimensiones (Figura 13), que retienen el 26% y el 26,1% de la variabilidad antes de la docencia (Figura 13, izquierda) y tras ella (Figura 13, derecha). El primer hecho a destacar es la alta relación entre ítems que conforman de manera teórica una misma componente latente, formando grupos de ítems correspondientes a cada una de las dimensiones de ambos cuestionarios. Siguiendo las normas de la interpretación de los métodos Biplot, cabe destacar que:

- Los ítems de la dimensión de Utilidad Presente (morado) se relacionan de manera directa con el Interés (azul) de los alumnos, y la ansiedad (rojo) con la Utilidad Presente (morado).
- Se observa una correlación inversa entre la dimensión de enfoque superficial (rosa) y profundo (naranja).
- Del análisis de la correlación entre las dimensiones de ambos cuestionarios, se sigue que el aprendizaje profundo se correlaciona de forma directa con el Interés de los alumnos y el Estudio Superficial de manera directa con Ansiedad y Utilidad Presente.
- Tras la docencia recibida, las relaciones entre los ítems de actitud y utilidad profesional por un lado, y ansiedad y utilidad presente por otro, son mayores.

Estos resultados corroboran la relación de los constructos afectivos, valorativos y de aprendizaje, hecho estudiado con un enfoque diferente en (Elias & Sánchez-Gelabert, 2014). El enfoque profundo y el enfoque superficial se reflejan como enfoques opuestos, más que complementarios, en consonancia con Geraldo, del Rincón y del Rincón (2011).

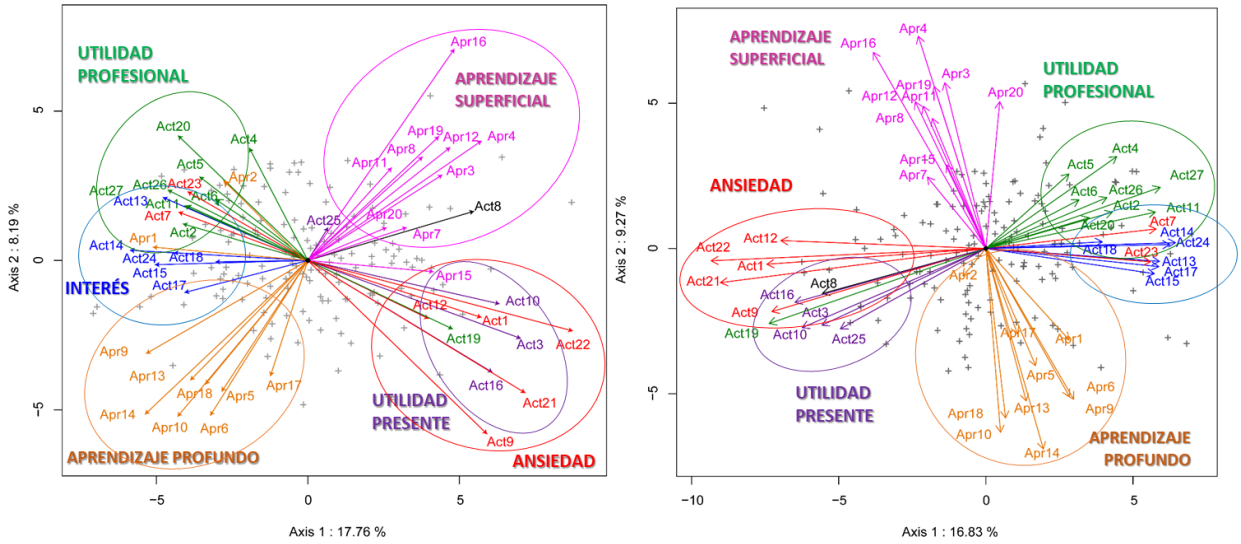


***Figura 13**. Plano factorial 1-2 obtenido en el análisis HJ-Biplot (izquierda: HJ-Biplot para los ítems de actitud pretest y aprendizaje; derecha: HJ-Biplot para los ítems de actitud postest y aprendizaje)*

Las contribuciones relativas del factor al elemento se encuentran en la Tabla S1 del Anexo 1.

El análisis de clúster sobre las coordenadas HJ-Biplot diferencia los perfiles multivariantes de cuatro grupos de estudiantes con comportamientos actitudinales y de aprendizaje diferentes (Figura 14). Se caracterizan por:

- **C1.** Estudiantes con altas puntuaciones en Interés, Utilidad Profesional y aprendizaje profundo antes de la docencia (panel 1, izquierda) y estudiantes con altas puntuaciones en Interés, Utilidad Profesional, pero aumentan su puntuación en estudio Superficial tras la docencia recibida (panel 1, derecha).
- **C2.** Estudiantes con puntuaciones altas en estudio superficial, ansiedad y utilidad presente antes de la docencia (panel 2, izquierda) y grupo de estudiantes con puntuaciones altas en ansiedad y utilidad presente, pero cuya puntuación en aprendizaje superficial disminuye y aumenta en aprendizaje profundo tras la impartición de la materia (panel 2, derecha).
- **C3.** Estudiantes con puntuaciones medias-altas en Interés y Utilidad Profesional, que aprenden de manera superficial y presentan bajos niveles de ansiedad (panel 3, izquierda). Grupo de alumnos que aumenta

su puntuación en estudio de manera superficial, con bajo interés y ansiedad (panel 3, derecha).

- **C4.** Estudiantes con altos niveles de ansiedad, que creen en la utilidad presente de la asignatura y que aprenden de manera profunda antes de la docencia (panel 4, izquierda) y tras la docencia, siguen creyendo en la utilidad de la asignatura para su día a día y aprenden de manera profunda, pero disminuye la ansiedad que presentan (panel 4, derecha).

Tras la docencia, se observa una disminución de los estudiantes que adoptan un enfoque profundo, que puede deberse a que la metodología impulsada por el Espacio Europeo de Educación Superior, favorecedora del aprendizaje profundo a través de la carga de trabajo y el estilo de evaluación, puede incrementar el enfoque superficial (Groves, 2005). Así como se aprecia que los estudiantes decantados por un aprendizaje profundo deben su elección a metas profesionales. Los resultados evidencian a su vez como la actitud del alumnado universitario no es estática, y puede estar condicionada por factores contextuales (García, Duarte, Rivera, Villalba, & Capacho, 2017; Ullah, 2016). Este estudio proporciona información a los docentes de la asignatura Didáctica General sobre la relación entre actitudes, concepciones y acciones (enfoques), que pueden emplearse en la propuesta de estrategias que den lugar a una comunidad estudiantil mejor formada en aspectos didácticos para el diseño y planificación de las secuencias de intervención socioeducativa, como puede ser la priorización de competencias asociadas al desarrollo de la profesión.

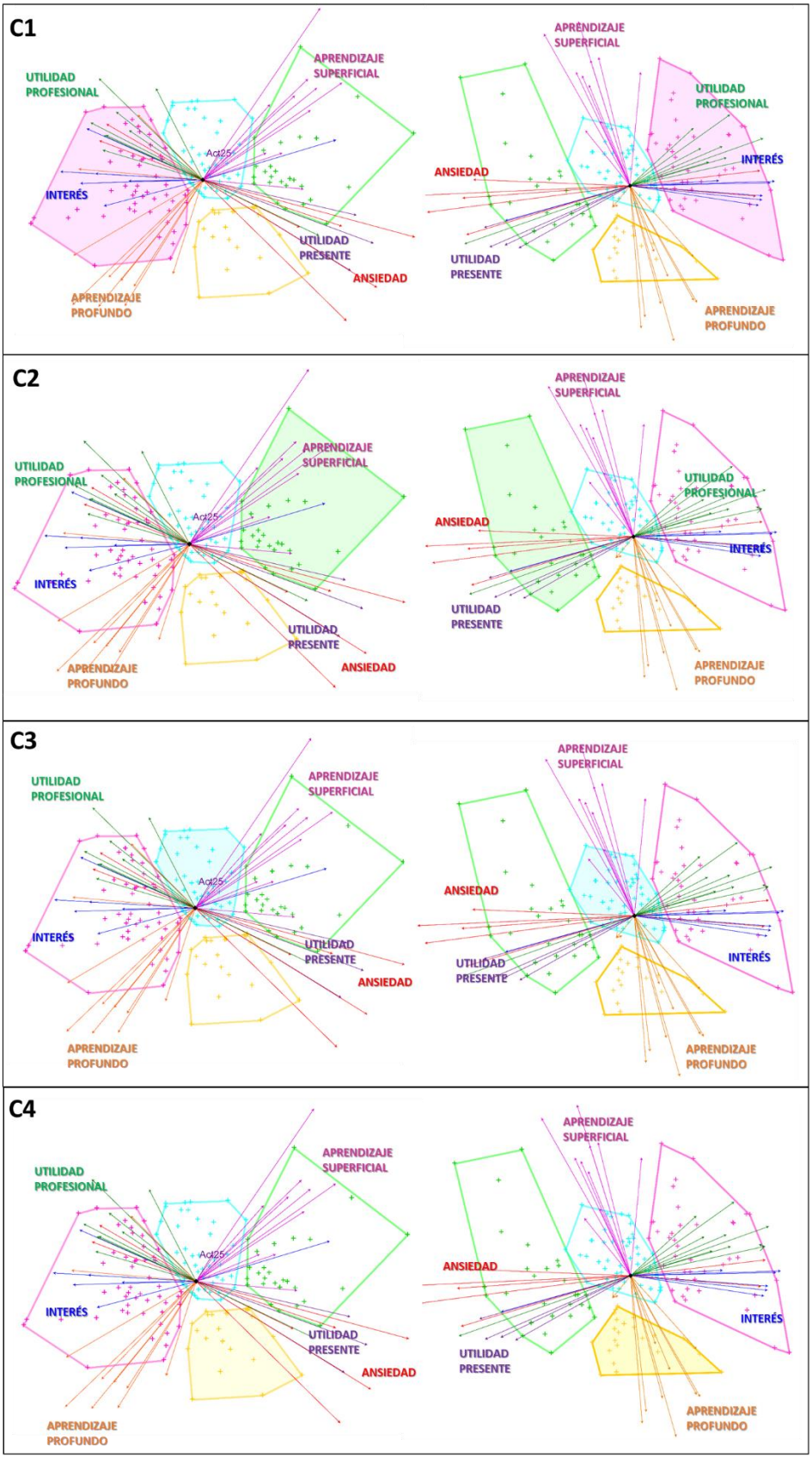


***Figura 14**. Identificación de cuatro perfiles multivariantes de estudiantes universitarios según su actitud y enfoque de aprendizaje. Análisis de clúster jerárquico de Ward sobre las coordenadas HJ-Biplot obtenidas sobre datos recogidos antes de la docencia (izquierda) y tras ella (derecha)*

Un resumen de estos resultados se presentó en el **I Congreso Virtual de Investigación e Innovación Educativa CIVINEDU 2017**, que tuvo lugar en octubre de 2017, con el trabajo "*Medición de la Actitud e Influencia del Tipo de Aprendizaje en el Estudio de la Didáctica*". Recientemente, han sido publicados en forma de artículo "*Attitude and Learning Approaches in the Study of General Didactics. A Multivariate Analysis*" en la revista **Psicodidáctica** (véase Figura 15 y Anexo 3).

Revista de Psicodidáctica, 2019, 24 (2), 154–162

psicodidáctica

Revista de Psicodidáctica

www.elsevier.es/psicod

Original

Attitude and Learning Approaches in the Study of General Didactics. A Multivariate Analysis☆

Nerea González-García[a,*], Ana B. Sánchez-García[b], Ana B. Nieto-Librero[a], and M. Purificación Galindo-Villardón[a]

[a] Departamento de Estadística, Universidad de Salamanca, Instituto de Investigación Biomédica de Salamanca (IBSAL), Salamanca, Spain
[b] INICO, Universidad de Salamanca, Salamanca, Spain

ARTICLE INFO

Article history:
Received 18 May 2018
Accepted 6 February 2019
Available online 14 April 2019

Keywords:
General Didactics
Attitude
Learning approaches
HJ-Biplot

ABSTRACT

The influence of different approaches and attitude towards learning in the General Didactics is studied in few investigations. The main objective of this work is to describe in a multivariate way the relationships between attitude towards General Didactics and learning approaches of students majoring in Social Learning of the University of Salamanca. The Measurement of Attitude towards General Didactics and the Revised Study Process Questionnaire two factor (R-SPQ-2F) questionnaires were used to gather the information. The analysis of the relationship between attitude and learning approaches is performed using the HJ-Biplot. This multivariate statistical technique allows the simultaneous representation of students, attitudes and learning approaches. This methodology, combined with hierarchical clustering method, reveals the existence of four types of students: C1, those characterised by high marks on interest, professional usefulness and deep study of the subject; C2, those that display high anxiety and high marks in superficial study; C3, students that show average interest, low anxiety, superficial study and believe in the professional usefulness of Didactics; and C4, students with high levels of anxiety that study the subject in depth. These results point the existence of a relationship between attitudes and learning approaches and can be used to improve the performance and offerings of educational teams, achieving more efficient strategies that lead to a better educated student community.

***Figura 15**. Artículo publicado en la revista Psicodidáctica (JCR 2018: 2,1 Q2; SJR 2018: 0,928 Q2)*

CAPÍTULO 2

ANÁLISIS DE DATOS DE DOS VÍAS: MÉTODOS SPARSE

2.1 Problema de mínimos cuadrados penalizado

Los métodos de mínimos cuadrados penalizados se definen a partir del problema OLS clásico, en el que se añade algún tipo de restricción sobre los parámetros a estimar. El problema de optimización restringido para la función de mínimos cuadrados se define como:

$$\widehat{\boldsymbol{b}} = \min_{\boldsymbol{b}} \|\boldsymbol{a} - \boldsymbol{b}\|_F^2$$

$$s.a. P(\boldsymbol{b}) \leq t$$

donde $P(\boldsymbol{b})$ es una cierta función de restricción. El problema puede ser reescrito a partir de los multiplicadores de Lagrange como un problema de penalización:

$$\widehat{\boldsymbol{b}} = \min_{\boldsymbol{b}} \|\boldsymbol{a} - \boldsymbol{b}\|_F^2 + \lambda(P(\boldsymbol{b}) - t)$$

La adición del término de penalización en un modelo de optimización definido se controla mediante un parámetro no negativo λ, conocido como parámetro de regularización, que es el encargado de controlar la cantidad de penalización introducida en el modelo. Cuanto más alto sea su valor, mayor será la penalización introducida en el modelo. Si este parámetro es nulo, nos encontramos con el modelo original de la técnica. Lógicamente, existe una relación doble entre λ y t. Estos parámetros serán seleccionados buscando un equilibrio entre la varianza absorbida y el error introducido en el modelo.

2.1.1 Tipos de penalización

Las técnicas de penalización son introducidas en los modelos de optimización con el fin único de reestructurar modelos mal condicionados, debido a la dificultad en la interpretación de los resultados o a la no existencia o no unicidad de la solución, como es el caso en $J \gg I$. La retención de un subconjunto importante de las variables originales produce modelos más interpretables, con menores errores de estimación que los modelos de características completas (Hastie et al., 2009). Las funciones de penalización más utilizadas son las correspondientes a la norma ℓ_p (o Lp) de un vector $\boldsymbol{\beta} \in \mathbb{R}^J$, que se define como:

$$\ell_p = \|\boldsymbol{\beta}\|_p = \sum_{j=1}^{J} p_\lambda(|\boldsymbol{\beta}_j|) = \left(\sum_{j=1}^{J} |\boldsymbol{\beta}_j|^p\right)^{1/p} \quad (2.1)$$

Las técnicas de penalización de la norma del vector de cargas más utilizadas, surgidas en el contexto de la regresión son las normas ℓ_0, ℓ_1 y ℓ_2, definidas a continuación.

En la literatura, el concepto de nulidad de cargas de los métodos sparse se relaciona de manera directa con la llamada norma ℓ_0, que hace referencia a la cardinalidad del vector $\boldsymbol{\beta}$; es decir, al número de sus coeficientes no nulos. Se define como:

$$\ell_0 = \|\boldsymbol{\beta}\|_0 = \lim_{p\to 0} \|\boldsymbol{\beta}\|_p = \lim_{p\to 0} \sum_{j=1}^{J} |\boldsymbol{\beta}_j|^p$$

El caso en que $p = 1$ se introduce posteriormente, puesto que su definición se recoge en años posteriores en la literatura.

Penalización Ridge.

Cuando $p = 2$ se define la norma L2 o norma Ridge (Hoerl & Kennard, 1970). Su función es penalizar la suma de los elementos al cuadrado del vector de cargas; esto es, añade una restricción sobre la norma L2 del vector de coeficientes.

$$L2 = \|\boldsymbol{\beta}\|_2 = \left(\sum_{j=1}^{J} |\boldsymbol{\beta}_j|^2\right)^{1/2}$$

Introduciendo esta restricción en la regresión por mínimos cuadrados ordinarios, se da lugar a la denominada regresión Ridge para la que el problema de optimización puede expresarse como:

$$\widehat{\boldsymbol{\beta}}_{RIDGE} = \min_{\boldsymbol{\beta}} \left\|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\right\|_F^2 = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_F^2$$

$$s.a. \|\boldsymbol{\beta}\|_2^2 \leq t$$

Utilizando los multiplicadores de Lagrange, esta ecuación puede escribirse:

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_F^2 + \lambda(\|\boldsymbol{\beta}\|_2^2 - t)$$

Tanto λ como t son constantes y, por ello, no ejercen ningún tipo de impacto sobre la solución. Por este motivo, podemos reescribir la expresión anterior como:

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_F^2 + \lambda\|\boldsymbol{\beta}\|_2^2$$

Desarrollando la norma de Frobenius, en notación matricial es equivalente a:

$$\widehat{\boldsymbol{\beta}}_{RIDGE} = \min_{\boldsymbol{\beta}} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^T \boldsymbol{\beta}$$

donde $\lambda > 0$ es el parámetro de regularización de la cantidad de penalización incluida en el modelo y $\lambda\boldsymbol{\beta}^T\boldsymbol{\beta}$ el término de penalización. Nótese que la relación entre λ y t es una relación de uno a uno entre ellos. Cuanto mayor sea el parámetro de regularización, mayor será la restricción generada (Khatavkar, 2007). Al igual que en el problema OLS, para encontrar el punto óptimo solución del problema es necesario diferenciar la expresión anterior e igualar a 0:

$$\frac{\partial}{\partial \widehat{\boldsymbol{\beta}}} = 0$$

$$\frac{\partial}{\partial \widehat{\boldsymbol{\beta}}} = -2\boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) + 2\lambda\widehat{\boldsymbol{\beta}} = \boldsymbol{0}$$

$$-\boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) + \lambda\widehat{\boldsymbol{\beta}} = \boldsymbol{0}$$

$$\boldsymbol{X}^T\boldsymbol{Y} - \boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \lambda\widehat{\boldsymbol{\beta}} = \boldsymbol{0}$$

$$\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}} + \lambda\widehat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{Y}$$

$$[\boldsymbol{X}^T\boldsymbol{X} + \lambda]\widehat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{Y}$$

$$\widehat{\boldsymbol{\beta}}_{RIDGE} = [\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}]^{-\boldsymbol{1}}\boldsymbol{X}^T\boldsymbol{Y}$$

A partir de la solución $\boldsymbol{\beta}^{OLS}$, la solución $\widehat{\boldsymbol{\beta}}_{RIDGE}$ en matrices de diseño ortonormales se reescribe como:

$$\widehat{\boldsymbol{\beta}}_{RIDGE} = \frac{1}{1+\lambda}\boldsymbol{\beta}^{OLS}$$

Por todo ello, la solución del estimador Ridge viene dada por: $\widehat{\boldsymbol{\beta}}^{ridge} = [\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}]^{-\boldsymbol{1}}\boldsymbol{X}^T\boldsymbol{Y}$. El estimador Ridge es único para $\lambda > 0$. Esto es debido a que al añadir el término λ a la diagonal de $\boldsymbol{X}^T\boldsymbol{X}$ (es decir, calcular $\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}$) la matriz

resultante es no singular. Que la matriz sea no singular quiere decir que su determinante es distinto de 0 y, por ello, $[\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}]^{-1}$ existe y es única. Esta solución sirve para ilustrar la característica esencial de Ridge: la contracción de los coeficientes. Ridge tiene el efecto de contraer las estimaciones de los coeficientes hacia cero introduciendo sesgo, pero reduciendo la varianza de la estimación. Además, Ridge tiene la propiedad favorable de asignar coeficientes similares a variables correlacionadas. A pesar de que Ridge contrae los coeficientes de un vector hacia 0, no los hace exactamente nulos, pues la solución se dará en el punto de intersección de los contornos elípticos de las estimaciones por mínimos cuadrados ordinarios con la región o bola definida por la norma Ridge $\mathfrak{B}^{\ell_2} = \{\boldsymbol{x}/\|\boldsymbol{x}\|_2^2 \leq t\}$ (Figura 16).

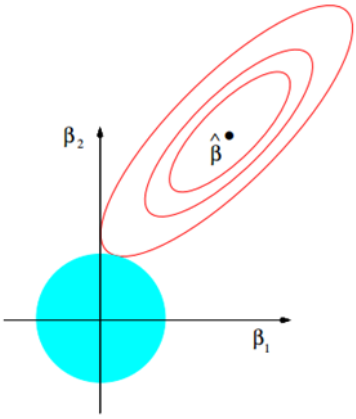


Figura 16. *Interpretación geométrica de la solución Ridge en el punto de intersección de las dos regiones a las que debe pertenecer (Hastie et al., 2009)*

Por último, en cuanto al posible valor del parámetro de regularización λ puede decirse que cuando $\lambda = 0$ el problema de optimización penalizado queda resumido en un problema OLS, de cuya solución ya se ha hablado anteriormente. El parámetro $\lambda \in [\lambda_{min}, \lambda_{max}]$ con λ_{max} aquel valor que haga que todos los valores sean contraídos a 0. Para la elección de este existen técnicas de selección de los parámetros de regularización, que consideran el error introducido en el modelo. La validación cruzada y los criterios de información como el AIC (*Akaike Information Criterion*) (Akaike, 1974) y el criterio de información bayesiano (*Bayesian Information Criterion*, BIC) (Kass & Raftery, 2012) son algunos de los más utilizados.

Penalización Lasso.

En el año 1996, aparece por primera vez en la literatura la norma $L1$. Esta norma, para $p = 1$, conocida como norma Lasso (*Least Absolute Shrinkage and Selection Operator*) (Tibshirani, 1996), es una de las más utilizadas en los métodos de análisis que trabajan con grandes volúmenes de datos. Se encarga de penalizar la suma de los valores absolutos de los coeficientes de un vector $\boldsymbol{\beta}$.

$$L1 = \|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{J} |\boldsymbol{\beta}_j|$$

Este cambio en la función de penalización es sutil, pero tiene un gran impacto en el estimador resultante. La penalización Lasso surgió en el ámbito de los modelos de regresión y, a diferencia de Ridge, es una técnica de penalización de estimación de parámetros y selección de variables automática, puesto que contrae los coeficientes hacia 0, hasta llegar a hacer alguno de ellos exactamente nulo.

El problema de restricción asociado a Lasso viene dado por la función de pérdida:

$$\widehat{\boldsymbol{\beta}}_{LASSO} = \min_{\boldsymbol{\beta}} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_F^2$$

$$s.a. \|\boldsymbol{\beta}\|_1 \leq t$$

con t el radio de la bola $\mathfrak{B}^{\ell_1} = \{\boldsymbol{x}/\|\boldsymbol{x}\|_1 \leq t\}$ encargado de restringir los coeficientes del vector considerado. El problema restringido puede escribirse equivalentemente como el problema penalizado, con $\lambda > 0$:

$$\widehat{\boldsymbol{\beta}}_{LASSO} = \min_{\boldsymbol{\beta}} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_F^2 + \lambda\|\boldsymbol{\beta}\|_1$$

Por un procedimiento similar al caso de Ridge, Tibshirani (1996) demostró que la solución óptima $\widehat{\boldsymbol{\beta}}_{LASSO}$ al problema de optimización de minimización de la función de pérdida viene dada por el operador *soft-thresholding*:

$$\widehat{\boldsymbol{\beta}}_{LASSO} = S_\lambda(\boldsymbol{\beta}^{OLS}) = S(\boldsymbol{\beta}^{OLS}, \lambda) = \text{sign}(\boldsymbol{\beta}^{OLS})(|\boldsymbol{\beta}^{OLS}| - \lambda)_+$$
$$= \begin{cases} \boldsymbol{\beta}^{OLS} + \lambda & si\ \boldsymbol{\beta}^{OLS} < \lambda \\ 0 & si\ \boldsymbol{\beta}^{OLS} \in [-\lambda, \lambda] \\ \boldsymbol{\beta}^{OLS} - \lambda & si\ \boldsymbol{\beta}^{OLS} > \lambda \end{cases}$$

La restricción Lasso es la más utilizada en la literatura puesto que se convierte en una técnica de selección automática de variables, muy útil al trabajar con bases de datos de altas dimensiones. Su interpretación geométrica se observa en la Figura 17. La solución del problema de optimización restringido tendrá lugar en el punto de intersección de los contornos elípticos de las estimaciones por mínimos cuadrados ordinarios con la región de restricción de la penalización impuesta. La diferencia en la región de restricción entre Ridge y Lasso supone un cambio importante. Mientras que la solución Lasso puede darse en alguno de los vértices del polígono regular situados en los ejes de definición, donde alguno de los coeficientes es nulo, en Ridge esto no es así, puesto que nunca se dará la solución en dichos ejes.

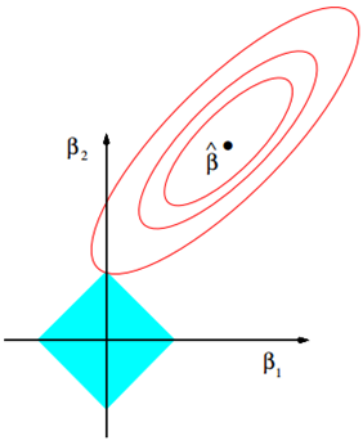


Figura 17. *Interpretación geométrica de Lasso en el caso bidimensional. Fuente: (Hastie et al., 2009)*

Nuevamente, el parámetro de regularización puede ser escogido a partir de algún criterio de información como el BIC, como aquel parámetro que lo minimice o mediante validación cruzada. El valor máximo permitido para λ será aquel que haga que todos los coeficientes sean nulos.

A pesar de sus múltiples ventajas, Lasso presenta algunos inconvenientes que hay que tener en cuenta y que han dado lugar al desarrollo matemático de otros tipos de regularizaciones. Para mejorar estas propiedades desfavorables,

surgen extensiones y/o combinaciones de las penalizaciones clásicas (Figura 18).

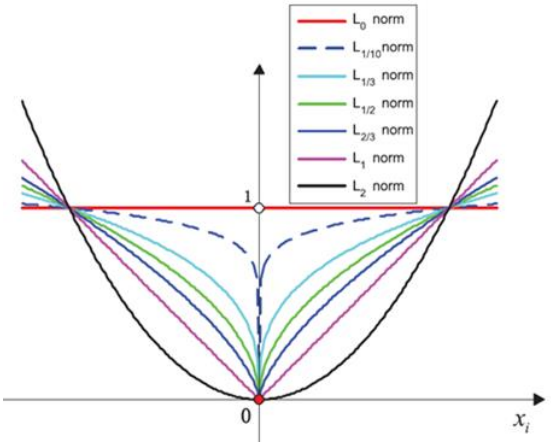


Figura 18. *Restricción sometida a los problemas de optimización en función de la norma del vector de cargas penalizada* (Zhang, Member, Xu, & Member, 2015).

Penalización Adaptive Lasso.

Lasso no verifica ser un procedimiento oracle (Zou, 2006). Si una solución $\widehat{\boldsymbol{\beta}}$ se ha calculado mediante un proceso oracle, esta verifica las siguientes propiedades(Fan & Runze, 2001; Khatavkar, 2007):

- Correcta selección de variables con una alta probabilidad.
- Siendo $\sum^*$ la matriz de covarianzas del modelo verdadero, $\sqrt{I}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \rightarrow N(0, \sum^*)$

En otras palabras, Lasso realiza una selección de variables inconsistente porque no selecciona el conjunto correcto de variables con una probabilidad que converge a 1, permitiendo que aparezcan características redundantes en el modelo estimado. Para manejar este inconveniente Zou (2006) presenta *adaptive* Lasso, una versión ponderada de Lasso que verifica las propiedades de ser un procedimiento oracle.

La función objetivo para adaptive Lasso viene dada por:

$$\widehat{\boldsymbol{\beta}}_{adLASSO} = \min_{\boldsymbol{\beta}} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_F^2 + \lambda \sum_j w_j |\beta_j|$$

con $\boldsymbol{w} = (w_1, \ldots, w_J)$ el vector de pesos para cada coeficiente del vector y que habitualmente se define a partir de la solución del problema OLS asociado como:

$$w_j = |\boldsymbol{\beta}^{OLS}|^{-\gamma}$$

donde $\gamma = 2$ en la práctica como sugieren los autores de adaptive Lasso (Zou, 2006). La solución al problema adaptive Lasso y al problema de restricción Lasso están íntimamente relacionadas, puesto que el valor óptimo $\widehat{\boldsymbol{\beta}}_{adLASSO}$ puede obtenerse a partir de la matriz de pesos $\boldsymbol{W}$ de coeficientes de pesos para cada uno de los vectores a penalizar:

$$\widehat{\boldsymbol{\beta}}_{adLASSO} = \boldsymbol{W}^{-1}\widehat{\boldsymbol{\beta}}_{LASSO}$$

Existen otras extensiones de la penalización Lasso para aquellos casos en que las variables presenten estructura de grupos, como *group* Lasso (Yuan & Lin, 2006).

Penalización Elastic net.

Otra de las desventajas de Lasso, es que tiende a seleccionar una variable de entre un grupo de correlacionadas. La esencia de los métodos multivariantes radica en aprovechar la relación entre las variables para explicar los patrones de los datos. Esto supone una inconsistencia de Lasso a nivel práctico en diversas disciplinas, como en el análisis de la expresión génica de microarrays, donde es importante tener en cuenta la actividad de compromiso de los genes en múltiples mecanismos biológicos (Hore et al., 2016; Wang et al., 2015), o en el análisis psicométrico de cuestionarios en psicología, donde cada una de las construcciones latentes está compuesta por un conjunto de ítems (Barahona et al., 2018; Vega-Hernández et al., 2018). Como solución a esto surge la penalización Elastic net (Zou & Hastie, 2005) (Figura 19). Se define como una combinación de las penalizaciones Lasso y Ridge:

$$\ell_1 + \ell_2 = \alpha\|\boldsymbol{\beta}\|_2^2 + (1-\alpha)\,\|\boldsymbol{\beta}\|_1$$

de modo que, cuando $\alpha = 0$ se trata de la penalización Lasso y Ridge cuando $\alpha = 1$. Esta penalización surge para hacerse eco de las ventajas de Ridge y Lasso y solventar sus desventajas. Se convierte en un método de selección de variables automática al incluir en su modelo la penalización Lasso y gracias a

Ridge asocia coeficientes similares a las variables pertenecientes a un grupo de correlacionadas.

El problema de restricción de un vector al espacio Elastic net viene dado por:

$$\widehat{\boldsymbol{\beta}}_{ENET} = \min_{\boldsymbol{\beta}} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_F^2$$

$$s.a. \alpha\|\boldsymbol{\beta}\|_2^2 + (1-\alpha)\,\|\boldsymbol{\beta}\|_1 \leq t$$

con t el radio de la bola de la región Elastic net $\mathfrak{B}^{\ell_1+\ell_2} = \{\boldsymbol{x}/\alpha\|\boldsymbol{x}\|_2^2 + (1-\alpha)\,\|\boldsymbol{x}\|_1 \leq t\}$.

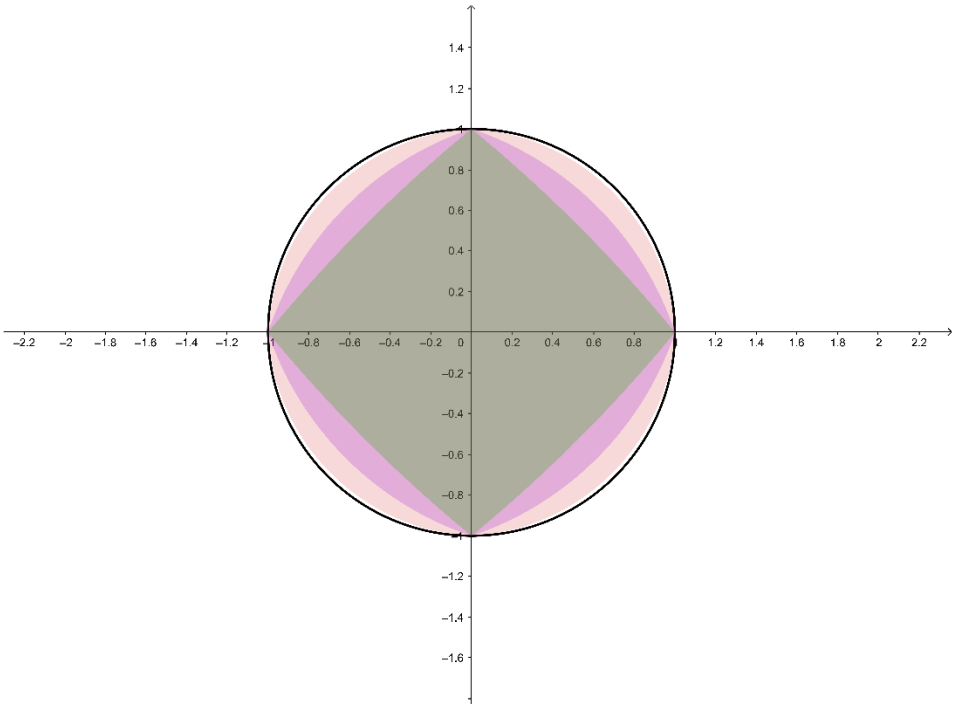


***Figura 19**. Interpretación geométrica bidimensional de Elastic net. En verde Elastic net para $\alpha = 0{,}1$, en rosa para $\alpha = 0{,}5$ y naranja $\alpha = 0{,}9$. En negro aparece la región $\mathfrak{B}^{\ell_2}$*

Cabe mencionar que la combinación de Elastic net y adaptive Lasso ha dado lugar a la penalización adaptive Elastic net que se calcula como una combinación de Ridge y Lasso ponderado (Zou & Zhang, 2009).

Penalizaciones no convexas.

Las restricciones introducidas hasta ahora son penalizaciones convexas. Existen otro tipo de restricciones, no convexas, que pueden ser añadidas al problema de optimización. Algunas de las más útiles en la práctica son SCAD (*Smoothly Clipped Absolute Deviation*) (Fan & Li, 2001) y MCP (*Minimax Concave Penalty*) (Zhang, 2010), pero no serán objeto de investigación en este trabajo.

2.2 Métodos sparse de factorización matricial

La penalización sparse también ha sido añadida en el campo de las metodologías de descomposición matricial.

2.2.1 Descomposición matricial penalizada PMD

Sea $\boldsymbol{X}$ una matriz de dimensión IxJ de rango $R \leq min(I,J)$ y la SVD de $\boldsymbol{X}$:

$$\boldsymbol{X} = \boldsymbol{UDV^T}$$

verificando $\boldsymbol{U^TU} = \boldsymbol{I}, \boldsymbol{V^TV} = \boldsymbol{I}$ y $\boldsymbol{D}$ una matriz diagonal que almacena $\boldsymbol{\lambda_1} \geq \boldsymbol{\lambda_2} \geq \ldots \geq \boldsymbol{\lambda}_R > \boldsymbol{0}$ y donde las matrices $\boldsymbol{U}, \boldsymbol{D}, \boldsymbol{V}$ son las matrices que minimizan la normal al cuadrado de Frobenius:

$$\underset{d,\boldsymbol{u},\boldsymbol{v}}{\operatorname{argmin}} \frac{1}{2} \|\boldsymbol{X} - d_r \boldsymbol{u}_r \boldsymbol{v}_r{}^T\|_F^2$$

$$\text{s.a. } \{\boldsymbol{u}_r{}^T \boldsymbol{u}_r = \boldsymbol{v}_r{}^T \boldsymbol{v}_r = 1, \boldsymbol{u}_r{}^T \boldsymbol{u}_{r'} = \boldsymbol{v}_r{}^T \boldsymbol{v}_{r'} = 0 \quad \forall r \neq r'.$$

Partiendo de la aproximación de rango 1:

$$\underset{d,\boldsymbol{u},\boldsymbol{v}}{\operatorname{argmin}} \frac{1}{2} \|\boldsymbol{X} - d\boldsymbol{u}\boldsymbol{v}^T\|_F^2$$

$$\text{s.a. } \|\boldsymbol{u}\|_F^2 = 1, \|\boldsymbol{v}\|_F^2 = 1$$

Witten, Tibshirani y Hastie (2009) desarrollan PMD (*Penalized Matrix Decomposition*), un método de descomposición matricial penalizado que generaliza el problema de descomposición en valores singulares, mediante la adición de restricciones sobre los vectores de las matrices $\boldsymbol{U}$ y $\boldsymbol{V}$:

$$\underset{\boldsymbol{u},\boldsymbol{v}}{\operatorname{argmin}} \frac{1}{2}\|\boldsymbol{X} - d\boldsymbol{u}\boldsymbol{v}^T\|_F^2$$
$$\text{s.a. } \|\boldsymbol{u}\|_F^2 = 1, \|\boldsymbol{v}\|_F^2 = 1, P_1(\boldsymbol{u}) \leq c_1, P_2(\boldsymbol{v}) \leq c_2 \tag{2.2}$$

Las funciones P_1 y P_2 son funciones de penalización convexas que pueden tomar variedad de formas según el objetivo que se pretenda conseguir con la restricción añadida. Witten, Tibshirani y Hastie (2009) proponen el uso de Lasso ($P_1 = \sum_j |\boldsymbol{u_j}|$) y Fused-Lasso ($P_1 = \sum_j |\boldsymbol{u_j}| + \lambda \sum_j |\boldsymbol{u_j} - \boldsymbol{u_{j-1}}|$).

Por otro lado, dadas $\boldsymbol{U}$ una matriz de dimensión IxQ y $\boldsymbol{V}$ una matriz de dimensión JxQ:

$$\frac{1}{2}\|\boldsymbol{X} - \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\boldsymbol{T}}\|_F^2 = \frac{1}{2}\|\boldsymbol{X}\|_F^2 - \frac{1}{2}\sum_{q=1}^{Q} d_q \boldsymbol{u}_q^T \boldsymbol{X} \boldsymbol{v}_q + \frac{1}{2}\sum_{q=1}^{Q} d_q^2 \tag{2.3}$$

Reescribiendo (2.3) como un problema de optimización a maximizar a partir de (6):

$$\max_{\boldsymbol{u},\boldsymbol{v}} \boldsymbol{u}^T \boldsymbol{X} \boldsymbol{v}$$
$$\text{s.a. } \|\boldsymbol{u}\|_F^2 = 1, \|\boldsymbol{v}\|_F^2 = 1, P_1(\boldsymbol{u}) \leq c_1, P_2(\boldsymbol{v}) \leq c_2 \tag{2.4}$$

Este problema es no convexo para $\boldsymbol{u}$ y $\boldsymbol{v}$ debido a la igualdad en la penalización L2 y por eso, esta restricción debe ser relajada a:

$$\max_{\boldsymbol{u},\boldsymbol{v}} \boldsymbol{u}^T \boldsymbol{X} \boldsymbol{v}$$
$$\text{s.a. } \|\boldsymbol{u}\|_F^2 \leq 1, \|\boldsymbol{v}\|_F^2 \leq 1, P_1(\boldsymbol{u}) \leq c_1, P_2(\boldsymbol{v}) \leq c_2 \tag{2.5}$$

Para resolver este problema se plantea un método iterativo de dos pasos biconvexo en el que se fijará $\boldsymbol{v}$ para calcular $\boldsymbol{u}$ como solución al problema de optimización que ahora si es convexo:

$$\max_{\boldsymbol{u},\boldsymbol{v}} \boldsymbol{u}^T \boldsymbol{X} \boldsymbol{v}$$
$$\text{s.a. } \|\boldsymbol{u}\|_F^2 \leq 1, P_1(\boldsymbol{u}) \leq c_1$$

En el caso de que la penalización P_1 sea la restricción Lasso, el vector solución $\boldsymbol{u}$ vendrá dado por el operador *soft-thresholding* y la normalización del vector:

$$\boldsymbol{u} = \frac{\boldsymbol{S}_{\Delta}(\boldsymbol{Xv})}{\|\boldsymbol{S}_{\Delta}(\boldsymbol{Xv})\|_2}$$

para un cierto Δ tal que $\|\boldsymbol{u}\|_1 \leq c_1$. Posteriormente dado $\boldsymbol{u}$, se obtendrá la solución $\boldsymbol{v}$ al problema:

$$\max_{\boldsymbol{u},\boldsymbol{v}} \boldsymbol{u}^T\boldsymbol{Xv}$$

$$\text{s.a. } \|\boldsymbol{v}\|_F^2 \leq 1, P_2(\boldsymbol{v}) \leq c_2$$

cuya solución vendrá de la misma forma que en el caso anterior. El algoritmo que resuelve este problema se resume en la Tabla 9.

Tabla 9. *Pseudocódigo para la implementación de PMD para un solo factor*

Algoritmo: PMD	
Entrada:	$\boldsymbol{X} \in \mathbb{R}^{IxJ}$, rango Q, $\varepsilon \approx 0$
Salida:	$\boldsymbol{u} \in \mathbb{R}^I, d \in \mathbb{R}, \boldsymbol{v} \in \mathbb{R}^J$
2:	Inicializar $\boldsymbol{v}$ aleatoriamente con norma L2 igual a 1
	Mientras no se cumpla el criterio de convergencia **hacer**:
	$\boldsymbol{u} = \max(\boldsymbol{u}^T\boldsymbol{Xv})$ s.a. $P_1(\boldsymbol{u}) \leq c_1$ y $\|\boldsymbol{u}\|_F^2 = 1$
	$\boldsymbol{v} = \max(\boldsymbol{u}^T\boldsymbol{Xv})$ s. a. $P_2(\boldsymbol{v}) \leq c_2$ y $\|\boldsymbol{v}\|_F^2 = 1$
	$d = \boldsymbol{u}^T\boldsymbol{Xv}$
9:	**Fin**

El algoritmo anterior está diseñado para el cálculo de un solo factor sparse, pero puede ser fácilmente extendido al cálculo de múltiples factores aplicando nuevamente el algoritmo sobre la matriz deflactada para $q = 1, \ldots, Q$:

$$\boldsymbol{X_{q+1}} \leftarrow \boldsymbol{X_q} - \boldsymbol{u}_q d_q \boldsymbol{v}_q^T$$

2.2.2 Sparse NMF

La NMF es un método popular de clustering desde el punto de vista de los métodos de factorización matricial. Su gran cualidad es la generación de Q dimensiones latentes no negativas; esto es, formadas por la combinación de coeficientes positivos. La NMF plantea aproximar una matriz de I observaciones

y J variables mediante Q factores latentes: $\boldsymbol{X} \approx \boldsymbol{HW^T}$, con $\boldsymbol{H} \in \mathbb{R}_+^{I\,x\,Q}$ y $\boldsymbol{W} \in \mathbb{R}_+^{JxQ}$, a partir del problema de minimización de la función de pérdida:

$$\min_{\boldsymbol{H},\boldsymbol{W}>0} \|\boldsymbol{X} - \boldsymbol{HW^T}\|_F^2$$

La utilidad de esto se desenvuelve en que la NMF produce una representación de los datos sparse. Esto la convierte automáticamente en un método que proporciona resultados sparse (nulos), puesto que para definir las matrices de la descomposición y que la aproximación pueda darse, algunos de los coeficientes pueden tener que ser cero. Sin embargo, dado que la sparsity se produce de manera natural y no como un objetivo de la factorización, no se puede controlar la cantidad de coeficientes nulos que se originan e incluso puede que no se generen factores sparse.

Para tener un mayor poder sobre esto, algunos autores como Hoyer (2004) y Roux, Weninger y Hershey (2015) proponen añadir la restricción sparse al problema de la NMF para que esta sea controlada. Hoyer propone minimizar la distancia euclídea entre $\boldsymbol{X}$ y $\boldsymbol{HW^T}$ añadiendo una restricción de sparsity fija en ambas matrices (Hoyer, 2004). La descomposición sparse NMF (sNMF) se refiere al problema de encontrar las matrices $\boldsymbol{H}$ y $\boldsymbol{W}$ tales que $\boldsymbol{X} \approx \boldsymbol{HW^T}$, sujetas a las restricciones:

$$\boldsymbol{X}, \boldsymbol{W}, \boldsymbol{H} \geq \boldsymbol{0};\ \boldsymbol{\sigma}(\boldsymbol{h}_j) = \boldsymbol{\sigma}_{\boldsymbol{H}};\ \boldsymbol{\sigma}(\boldsymbol{w}_j) = \boldsymbol{\sigma}_{\boldsymbol{W}}$$

donde $\boldsymbol{h}_j$, $\boldsymbol{w}_j$ son los vectores columna de las matrices $\boldsymbol{H}$ y $\boldsymbol{W}$ respectivamente y $\boldsymbol{0} \leq \boldsymbol{\sigma}_{\boldsymbol{H}}, \boldsymbol{\sigma}_{\boldsymbol{W}} \leq \boldsymbol{1}$ escalares que controlan la sparsity de los vectores mencionados con $\boldsymbol{\sigma}(\boldsymbol{x}) = (\sqrt{\mathrm{j}} - \|\boldsymbol{x}\|_1/\|\boldsymbol{x}\|_2)/(\sqrt{j} - 1)$. Este es uno de los trabajos más importantes en el ámbito del sparse NMF. Este problema es resuelto por Hoyer mediante un algoritmo del gradiente descendente para llegar a la solución del problema mediante la definición de una serie de reglas multiplicativas (NMFSC, *Nonnegative Matrix Factorization with Sparseness Constraints*), aunque más tarde, Heiler y Schnörr (2006) y Potluru et al., (2013) proponen dos algoritmos diferentes para su programación; este último basado en el algoritmo de las coordenadas descendentes en bloque. La unicidad de la

solución a este problema fue estudiada en (Theis, Stadlthanner, & Tanaka, 2005)**.**

En el año 2008, Kim y Park (2008) desarrollan la sparse NMF, incluyendo sobre el problema de optimización la penalización L1. Para ello, Kim y Park proponen dos formulaciones diferentes de la NMF que generen resultados sparse. Por un lado, proponen imponer la penalización sparse en la matriz de componentes $\boldsymbol{H}$:

$$\min_{\boldsymbol{H},\boldsymbol{W}>0} \|\boldsymbol{X} - \boldsymbol{H}\boldsymbol{W}^{\boldsymbol{T}}\|_F^2 + \mu\|\boldsymbol{W}\|_F^2 + \beta \sum_{q=1}^{Q} \left\|\boldsymbol{h}_j\right\|_1^2$$

con $\boldsymbol{h}_j$ vectores columna de $\boldsymbol{H}$**,** $\mu \geq 0$ medida que controla el crecimiento de $\boldsymbol{W}$ y β el parámetro de regularización que controla la balanza entre la precisión de la aproximación y la restricción sparse sobre la norma L1 de $\boldsymbol{H}$. Por otro lado, equivalentemente, proponen agregar la restricción sobre la norma L1 de la matriz de cargas $\boldsymbol{W}$**:**

$$\min_{\boldsymbol{H},\boldsymbol{W}>0} \|\boldsymbol{X} - \boldsymbol{H}\boldsymbol{W}^{\boldsymbol{T}}\|_F^2 + \mu\|\boldsymbol{H}\|_F^2 + \beta \sum_{s=1}^{Q} \|\boldsymbol{w}_s\|_1^2$$

con $\boldsymbol{w}_s$ vectores columna de $\boldsymbol{W}$**,** $\mu \geq 0$ medida que controla el crecimiento de $\boldsymbol{H}$ y β el parámetro de regularización que controla la balanza entre la precisión de la aproximación y la restricción sparse sobre la norma L1 de $\boldsymbol{H}$.

Los métodos más frecuentes se centran en restringir o penalizar la norma L1 de una de las matrices $\boldsymbol{H}$ o $\boldsymbol{W}$ (Hoyer, 2002, 2004). Sin embargo, unos años más tarde surgen extensiones del sparse NMF que tratan de restringir otro tipo de normas del vector de cargas. Qian, Jia, Zhou y Robles-Kelly (2011) proponen utilizar la restricción sobre la norma L1/2 (Figura 20), siendo el sparse NMF la solución al problema:

$$\min_{\boldsymbol{H},\boldsymbol{W}>0} \|\boldsymbol{X} - \boldsymbol{H}\boldsymbol{W}^{\boldsymbol{T}}\|_F^2 + \lambda\|\boldsymbol{W}\|_{1/2}$$

mediante un algoritmo iterativo. La norma L1/2 se define como: $\|\boldsymbol{W}\|_{1/2} = \sum_{j=1}^{J} \boldsymbol{w}_j^{1/2}$.

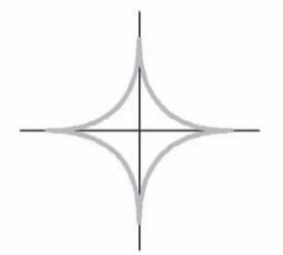


Figura 20 Representación gráfica en dos dimensiones de la región $L_{1/2}$ o $L_{0,5}$

Un año más tarde, Peharz y Pernkopf (2012) proponen añadir una restricción sobre la norma L0 de los vectores de puntuaciones factoriales y de cargas. Recuérdese que la norma L0 hace referencia a la cardinalidad del vector, restringiendo el número de coeficientes no nulos de este. Así, plantean dos posibles problemas sparse restringidos:

$$\min_{\boldsymbol{H},\boldsymbol{W}>0} \|\boldsymbol{X} - \boldsymbol{H}\boldsymbol{W}^{\boldsymbol{T}}\|_F^2$$
$$s.a. \left\|\boldsymbol{h}_j\right\|_1 \leq N$$

$$\min_{\boldsymbol{H},\boldsymbol{W}>0} \|\boldsymbol{X} - \boldsymbol{H}\boldsymbol{W}^{\boldsymbol{T}}\|_F^2$$
$$s.a. \|\boldsymbol{w}_s\|_1 \leq M$$

donde $\forall j, s,\ N, M \in \mathbb{N}$ y representan el número máximo de coeficientes no nulos de cada vector.

2.3 Técnicas sparse de reducción de la dimensión

En la literatura, son muchas las técnicas que incorporan la penalización sparse en su definición: sparse LDA (Moghaddam, Weiss, & Avidan, 2007; Shao, Wang, Deng, & Wang, 2011), sparse PLS (Colombani et al., 2012; Lê Cao, Rossouw, Robert-Granié, & Besse, 2008), modelos de regresión sparse (Algamal & Lee, 2015; Filzmoser, Gschwandtner, & Todorov, 2012; Hesterberg, Choi, Meier, & Fraley, 2008; Meier, 2008), sparse CCA (Gao, Ma, & Zhou, 2017; Lykou & Whittaker, 2010), sparse FA (Engelhardt & Stephens, 2010),… En este capítulo nos centraremos en dos de las técnicas de reducción de la dimensión

más utilizadas: PCA y métodos Biplot y su extensión a los métodos de selección de variables. La penalización sparse limitará el número de variables en el modelo, facilitando ampliamente la visualización e interpretación de los resultados.

2.3.1 Análisis de Componentes Principales Sparse

La problemática principal del PCA es la dificultad en la interpretación de los resultados debido a que cada una de las PCs es combinación lineal de todas las variables de partida. Lo ideal sería que cada componente se calculara como la combinación de un subconjunto de las variables originales (con cargas nulas), pero en la práctica esto no sucede por normal general.

Existen múltiples formas de identificar la contribución de una variable al modelo definido y restringir el número de variables seleccionadas, como la Descomposición CUR o la factorización PMD. A continuación, veremos otra alternativa: análisis de componentes principales Sparse (Sparse PCA). La finalidad del sparse PCA es producir componentes principales modificadas con cargas nulas, $v_j = 0$ (*sparse)*. Es decir, componentes forzadas a ser combinación lineal de un pequeño subconjunto de las variables originales más importantes. Además de plantearse como solución al problema de interpretación de las componentes clásicas, surge como forma de solventar la inconsistencia del PCA para datos de altas dimensiones.

La umbralización ha sido la manera habitual de restringir el valor de las cargas de definición de un modelo factorial, con el objetivo de generar cargas cero, $v_j = 0$. Cada investigador plantea, de manera subjetiva, un umbral por debajo del cual todas las cargas serán consideradas como nulas. Para lograr esto atendiendo a criterios teóricos y siguiendo la idea de contraer el valor de las cargas factoriales, surgen los modelos sparse. Su idea principal es la de acercar a 0 aquellas cargas con magnitudes bajas a través de alguna función de penalización de la norma de los vectores de cargas. Al igual que muchas otras técnicas, existen diferentes formulaciones del problema (Ning-min & Jing, 2015) que se diferencian entre sí de acuerdo con la función objetivo planteada y a las condiciones de penalización de los coeficientes. Además, las dos vertientes

encontradas en su antecesor PCA (subespacio de mejor ajuste a los datos (Pearson, 1901) y maximización de la información de los datos originales absorbida por las componentes (Hotelling, 1933) vuelven a aparecer aquí puesto que puede considerarse como un método de PCA modificado. Los distintos algoritmos y extensiones de este método se agrupan resolviendo el problema desde ambas vías. Las técnicas de penalización centran su atención en las coordenadas de los vectores de carga, tratando de restringirlos esta vez con un criterio teórico, al igual que trataba de hacer la umbralización, aunque esta de manera subjetiva.

El primer método de Sparse PCA apareció en la literatura en el año 2003 de manos de Jolliffe, Trendafilov y Uddin bajo el nombre de SCoTLASS (*Simplified Component Technique for Least Absolute Shrinkage and Selection*) (Jolliffe et al., 2003), basándose en las ideas de restricción de las cargas a un subconjunto de enteros (Anaya-Izquierdo et al., 2011; Hausman, 1982; Vines, 2000). Planteado el problema de optimización desde el punto de vista de maximizar la varianza de las componentes principales, su principal objetivo es restringir las cargas del PCA clásico imponiendo la penalización Lasso a estas. El problema de optimización restringido que plantean es:

$$\max_{\substack{\|\boldsymbol{v}\|_2=1 \; y \; \|\boldsymbol{v}\|_1 \le \tau \\ \boldsymbol{v} \perp \{\boldsymbol{v}_1,\dots,\boldsymbol{v}_{j-1}\}}} \boldsymbol{v}^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{v} \tag{2.6}$$

con $\boldsymbol{v}_1, \dots, \boldsymbol{v}_Q$ los vectores asociados a las cargas de las Q componentes sparse retenidas para algún parámetro τ de ajuste de la regularización Lasso incluida en el modelo.

Más tarde se desarrolló el DSPCA (d'Aspremont, El Ghaoui, Jordan, & Lanckriet, 2007). Desde el punto de vista de maximización de la varianza y la penalización Lasso, su objetivo principal es proponer una relajación convexa del problema original (debido a la no convexidad de las penalizaciones $l_q, q = 1, \dots, \infty$) mediante programación semidefinida. Los métodos GPower (GPower0, GPower1) de Journée y Nesterov (2010) proponen el cálculo de las Sparse PCs (SPCs) secuencialmente, o en bloques de una en una, secuencialmente, mediante un problema de maximización de una función convexa inducida a distintos tipos de penalizaciones.

Desde el punto de vista de minimización del error de reconstrucción dos de las formulaciones más influyentes son el SPCA de Zou, Hastie y Tibshirani (2006) y el rSVD-sPCA de (Shen & Huang, 2008). El SPCA se plantea el PCA como un problema de optimización del tipo regresión, facilitando la modificación del algoritmo mediante la adición de la penalización Elastic net (Zou & Hastie, 2005). Para un vector dado $b \in \mathbb{R}^J$, el vector solución penalizado a que contiene las cargas de definición de cada componente sparse:

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

La solución a da para el i –ésimo vector las cargas de las componentes sparse. La principal desventaja de este método es que las cargas no verifican ser ortogonales y el cálculo de las SPCs se realiza en bloque, dificultando la selección de los parámetros de regularización debido a las correlaciones entre ellos.

El rSVD-sPCA plantea el problema desde el punto de minimización del error como un problema de optimización penalizada, a través de la SVD.

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

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. Se trata de un método consistente en la estimación de los parámetros del modelo, que facilita su uso al calcular cada una de las SCPs secuencialmente.

Por otro lado, generalizando las definiciones del SCoTLASS, SPCA y rSVD-sPCA, Witten, Tibshirani, & Hastie (2009) extienden el uso de PMD al cálculo del sparse PCA e incluso al análisis sparse de correlación canónica.

Más reciente en la literatura Qi, Luo y Zhao plantean un procedimiento similar a SCoTLASS (Qi, Luo, & Zhao, 2013), desde el punto de vista de minimización del error y combinando las penalizaciones l_1 y l_2, de manera que este conjunto de restricciones sea estrictamente convexo. Qi, Luo, y Zhao (2013) construyen componentes sparse usando la suma ponderada de las normas ℓ_1 y ℓ_2, así: $\|v\|_\lambda^2 = (1-\lambda)\|\boldsymbol{v}\|_2^2 + \lambda\|\boldsymbol{v}\|_1^2$. Así, el $j-ésimo$ vector de cargas de las componentes sparse es la solución del problema de optimización:

$$\min_{\|v\|_2=1} \frac{\boldsymbol{v}^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{v}}{\|\boldsymbol{v}\|^2_{\lambda_i}}$$
$$\boldsymbol{v} \perp \{\boldsymbol{v}_1, \dots, \boldsymbol{v}_{j-1}\}$$

donde $\boldsymbol{v}_1, \dots, \boldsymbol{v}_{j-1}$ 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 $0 \leq \lambda < 1$. A pesar de que las cargas no son ortonormales, es decir, $A^T A \neq I_r$, las SCPs obtenidas son no correlacionadas.

Una explicación más detallada de estos métodos, puede encontrarse en algunas de las revisiones publicadas más relevantes: (Ning-min & Jing, 2015; Trendafilov, 2014; Zhang et al., 2015).

2.3.2 Métodos Biplot sparse

Como se ha visto en secciones anteriores, los métodos Biplot permiten graficar una matriz multivariante en un espacio de menor dimensión con pérdida de información mínima a partir de matrices de marcadores fila y columnas representados en el mismo subespacio. Cuando los datos a analizar presentan algún tipo de estructura los métodos Biplot no la tienen en cuenta.

La incorporación de la penalización sparse de selección de variables automática a los métodos Biplot es muy reciente y apenas existe evidencia de trabajos que la incorporen. A continuación se presentan los dos métodos disponibles en la literatura para generar selección de variables en el ámbito de los métodos Biplot: CDBiplot y sparse HJ-Biplot (Elastic net HJ-Biplot).

Clustering Disjoint Biplot

Cuando se quieren detectar patrones de similitud entre individuos con estructura de grupos, generalmente se suelen utilizar técnicas de reducción de la dimensión complementados con técnicas de clasificación a posteriori, como los métodos de clúster jerárquicos y no jerárquicos. Sin embargo, este tipo de métodos no incorpora al problema de optimización la resolución de ambos problemas (reducción de la dimensión y análisis de clúster) y no son del todo óptimos pues no se buscan las direcciones de máxima separación entre grupos. Para solventar esta problemática, algunos autores proponen combinar las

técnicas de reducción de la dimensión y de análisis de cluster (Vichi & Kiers, 2001).

En este entorno surge el *Clustering Disjoint HJ-Biplot (CDBiplot)* (Nieto-Librero, Sierra, Vicente-Galindo, Ruíz-Barzola, & Galindo-Villardón, 2017), una extensión del Clustering Disjoint PCA (Vichi & Saporta, 2009) a los métodos Biplot, que propone simultanear la clasificación de objetos y la reducción de la dimensión mediante un algoritmo de mínimos cuadrados alternados. El objetivo del método es encontrar la mejor clasificación de las observaciones en un subespacio de dimensión reducida a la vez que mejorar la interpretación de los resultados del HJ-Biplot mediante la construcción de ejes factoriales disjuntos. Esto se consigue de manera que cada variable original solo pueda contribuir a la formación de una de las dimensiones latentes.

El modelo CDBiplot plantea el siguiente problema de optimización:

$$min\left\|\boldsymbol{X} - \boldsymbol{U}\bar{\boldsymbol{A}}\boldsymbol{\Lambda}^{-1}\boldsymbol{B}^{\boldsymbol{T}}\right\|_F^2$$

donde $\boldsymbol{X}$ es la matriz original con I objetos en filas y J variables en columnas, $\boldsymbol{U}$ es la matriz que define la localización de los I objetos en los P clusters considerados, $\bar{\boldsymbol{A}}$ es la matriz de centroides en el espacio de las Q componentes disjuntas, $\boldsymbol{\Lambda}$ es la matriz diagonal de valores propios de la SVD de la matriz $\boldsymbol{Z}$ que contiene información de los centroides correspondientes a cada objeto proyectados en el espacio de los P clústers y $\boldsymbol{B}$ es la matriz de coordenadas del HJ-Biplot de las variables de partida sobre las componentes disjuntas.

Para la resolución del problema de minimización del error residual Nieto-Librero et al. (2017) proponen un algoritmo en el sentido de los mínimos cuadrados alternados (ALS). De manera resumida, cada iteración del algoritmo se centra en la consecución de dos objetivos: agrupar las I observaciones de partida mediante el algoritmo de clasificación K-Means y encontrar el subespacio vectorial de dimensión reducida, aplicando el HJ-Biplot sobre la matriz de centroides resultantes de manera que las J variables de inicio sólo puedan contribuir a la formación de una única componente. Un esquema del CDBiplot puede verse en la Figura 21.

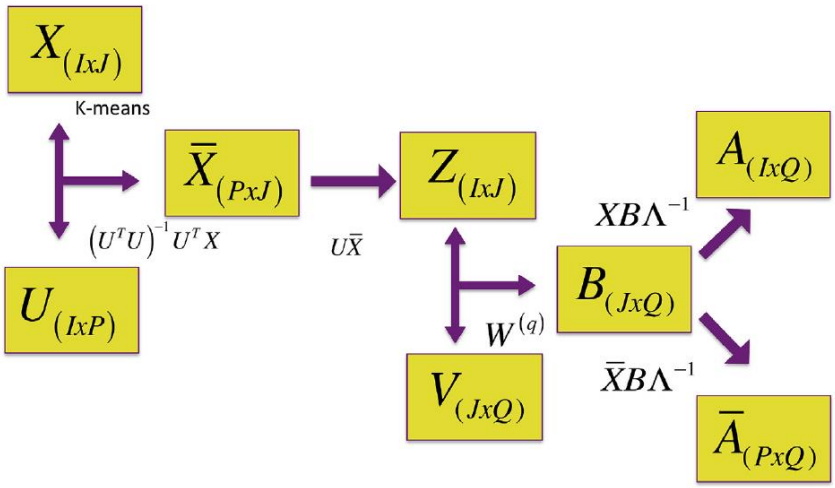


***Figura 21**. Esquema del algoritmo del CDBiplot. Fuente:* (Nieto-Librero et al., 2017)

Los autores desarrollan la función CDBiplot en R, incorporada a la librería *BiplotbootGUI* para su aplicación.

Sparse HJ-Biplot

Recientemente (Cubilla-Montilla, Galindo-Villardón, Nieto-Librero, Vicente, & García-Sánchez, 2019) proponen incorporar la penalización Elastic net al método HJ-Biplot para producir componentes sparse. El método trata de optimizar:

$$min\left\|\boldsymbol{X}-\boldsymbol{A}\boldsymbol{D}^{-1}\boldsymbol{B}^{T}\right\|_F^2+\lambda_2\sum_{j=1}^{J}\boldsymbol{\beta}_j^2+\lambda_1\sum_{j=1}^{J}\left|\boldsymbol{\beta}_j\right|$$

con $\lambda_1,\lambda_2>0$ parámetros de regularización que controlan la cantidad de penalización sparse incluida en el modelo. La penalización es impuesta sobre las cargas de las componentes principales. Estas técnicas están implementadas en el paquete SparseBiplots disponible en R (Cubilla-Montilla, Torres, Nieto-Librero, & Villardon, 2019).

Los dos métodos presentados presentan el inconveniente de que la generación de soluciones sparse o sparse disjuntas se produce en detrimento de la ortogonalidad de las componentes latentes. Ambos métodos generan componentes sparse correlacionadas, compartiendo información entre ellas.

2.4 Contribuciones al análisis de escalas psicométricas: una aplicación en psicología

Las técnicas factoriales de reducción de la dimensión de la estadística multivariante son uno de los recursos más utilizados en el ámbito de las ciencias sociales. Esto es debido a que en este tipo de disciplinas (psicología, educación,…) las variables de interés en el estudio no son las evaluadas directamente sobre un conjunto de individuos, como ocurre en las ciencias experimentales, sino que lo son los constructos no tangibles latentes en los datos, que existen a causa de la relación de las variables inicialmente medidas. Este tipo de factores latentes son cuantificados a través de test psicológicos, cuyas propiedades psicométricas hay que verificar antes de un estudio, como se ha mostrado en la sección 0. La forma clásica de estudiar la validez factorial de un constructo latente es utilizar un FA exploratorio, CFA y/o en multitud de ocasiones el PCA (Alarcón, 2006; García, Herrero, & León, 2007; Silva, Taveira, Marques, & Gouveia, 2015; Vellone, Barbaranelli, Lee, & Riegel, 2015). Ahora bien, este tipo de procedimientos son muy útiles cuando las dimensiones latentes presentan alta variabilidad o cuando el número de variables observables no es muy alto, pero fracasan estrepitosamente cuando los datos presentan efecto techo/suelo en cuanto a variabilidad se refiere o cuando el número de variables observables es muy alto como ocurre, por ejemplo, en los modernos estudios de datos de altas dimensiones. Proponemos aquí una alternativa que resolvería ambos problemas, basada en el Sparse PCA, mostrándose como una solución óptima en casos en que la absorción de inercia de los factores es baja (véase la Aplicación 2.1) o mejorando la capacidad informativa de los datos, generando soluciones factoriales más fácilmente interpretables (véanse la sección 1 y la Aplicación 2.2).

2.4.1 Sparse PCA como herramienta alternativa de análisis de validez factorial en situaciones de baja absorción de varianza.

Burnout. La sociedad actual está marcada por un ritmo de vida acelerado que propicia efectos negativos en la calidad de vida de la población. A nivel laboral, la carga de trabajo y las exigencias en cuanto a calidad y producción; a nivel emocional; la vulnerabilidad y el manejo de las emociones, y/o cualquier otra situación que provoque frustración personal, han puesto el foco de atención en la prevención de uno de los riesgos psicosociales con más importancia de los últimos años: burnout. El Burnout (término que puede traducirse como “estar quemado”) es un síndrome relativamente reciente que se dio a conocer de la mano de Freudenberger (1974) y está relacionado con el ámbito laboral al entenderse como una respuesta prolongada al estrés laboral crónico. En el año 1993, la psicóloga Christina Maslach, de la Universidad de Berkeley (California), lo definió como “un síndrome psicológico de agotamiento emocional, despersonalización y logros personales reducidos que pueden ocurrir entre individuos que trabajan con otras personas de alguna forma”. Esta situación, provocada por el trabajo, se traduce en una pérdida progresiva de energía que acaba desmotivando al profesional en su función, disminuyendo el desempeño del trabajo y falta de compromiso. Conlleva graves consecuencias físicas y psicológicas, como el agotamiento o la depresión cuando el fenómeno se somatiza y ha sido relacionado con otros fenómenos como la insatisfacción laboral o estrés general (Khamisa, Oldenburg, Peltzer, & Ilic, 2015) y ansiedad (Wild et al., 2014).

Las profesiones relacionadas con el mundo sanitario, servicios sociales, educación o administración pública suelen ser las que más incidencia reflejan (Hsieh, 2012)(Chao, Mccallion, & Nickle, 2011)(Loera, Converso, & Viotti, 2014). En el ámbito de las ciencias de la salud, la literatura incluye numerosos estudios acerca de esta patología: enfermería, medicina, odontología, … En el mundo educativo, se han realizado estudios en profesores de educación primaria, secundaria y estudios superiores universitarios. No obstante, el fenómeno ha sido también descrito en otro tipo de profesiones o incluso fuera del ámbito laboral (Carlin & Garcés de los Fayos, 2010) (Yavuz & Doğan, 2014).

Maslach Burnout Inventory. Existen diversos instrumentos creados para explorar y medir el burnout (véase Tabla 10). De entre todos ellos, el *Maslach Burnout Inventory* (MBI) (Maslach & Jackson, 1981) es el instrumento más utilizado internacionalmente para evaluar los niveles del síndrome Burnout (Shirom & Melamed, 2006). Por este motivo, ha sido traducido y validado en múltiples países y utilizado para explorar el burnout en diversas disciplinas (Ayyala, Ahmed, Ruzal-Shapiro, & Taylor, 2019; Bria, Spânu, Băban, & Dumitraşcu, 2014; Gracia et al., 2019; Loera et al., 2014; López et al., 2014; Vesty, Sridharan, Northcott, & Dellaportas, 2018; West, Dyrbye, & Shanafelt, 2018). Dado que las distintas versiones del MBI han sido utilizadas en múltiples disciplinas este ha tenido que ser adaptado, dando lugar a distintas versiones del mismo:

1) *MBI-Human Services Survey* (MBI-HSS, 22 ítems) (Maslach & Jackson, 1981). Es la versión clásica del MBI, diseñado para los profesionales de los servicios humanos.
2) *MBI-Educators Survey* (MBI-ES, 22 ítems) (Maslach, Jackson, Leiter, & Schaufeli, 1986). Esta versión está destinada a profesionales de educación, adaptada del MBI-HSS y que mantiene su estructura factorial.
3) *MBI-General Survey* (MBI-GS, 16 ítems) (Schaufeli, Leiter, Maslach & Jackson, 1996). Es la versión genérica dirigida a aquellos profesionales fuera del ámbito de los servicios humanos. Fue adaptada al español por Gil-Monte (2002).
4) *MBI-Students Survey* (MBI-SS, 15 ítems). Adaptada del MBI-GS, se construyó de manera específica para medir de manera exclusiva burnout en estudiantes universitarios. Existe una versión adaptada al español por Schaufeli, Martinez, Pinto, Salanova y Bakker (2002).

***Tabla 10**. Distintos instrumentos para evaluar el Burnout*

Cuestionario	Siglas	Referencia	Dimensiones latentes
Maslach Burnout Inventory	MBI	(Maslach & Jackson, 1981)	3
Copenhagen Burnout Inventory	CBI	(Kristensen, Borritz, Villadsen, & Christensen, 2005)	2
Oldenburg Burnout Inventory	OLBI	(Halbesleben & Demerouti, 2005)	2
Shirom-Melamed Burnout Measure	SMBM	(Shirom & Melamed, 2006)	
Bergen Burnout Inventory	BBI	(Feldt et al., 2014)	3

En este estudio se hizo uso del MBI-HSS, formado por 22 ítems valorados en escala tipo Likert (desde “nunca”, hasta “siempre”), que miden la frecuencia con que cada sujeto experimenta cada una de las situaciones descritas asociadas a su interacción con el trabajo. Como apuntan Maslach y Jackson (1981), los 22 ítems se agrupan en torno a tres factores: i) agotamiento emocional, sentimientos de cansancio afectivo (AE, 9 ítems, p.e. “*Me siento emocionalmente agotado/a por mi trabajo*”); ii) realización profesional, tendencia a una valoración personal negativa que puede afectar a la relación con las personas a las que atienden (AP, 8 ítems, p.e. “*Me siento muy activo/a*”) y iii) despersonalización, tendencia a adoptar una actitud fría, indiferente y distante hacia los demás, tratando a estos como si fueran objetos (DP, 5 ítems, p.e. “*Me he vuelto más insensible con la gente desde que ejerzo esta profesión*”). Puntuaciones altas en DP, AG y bajas en AP denotan la presencia de burnout.

La estructura de la escala de tres factores ha sido validada en múltiples investigaciones (Faye-Dumanget, Carré, Le Borgne, & Boudoukha, 2017)(Samaranayake & Seneviratne, 2012). Sin embargo, algunos autores señalan la cuestionable validez de la estructura factorial del mismo, con una ambigüedad patente en lo que respecta al número de dimensiones latentes. Algunos autores han propuesto soluciones bifactoriales (Kalliath, O’Driscoll, Gillespie, & Bluedorn, 2000)(Mészáros, Ádám, Szabó, Szigeti, & Urbán, 2014) o

incluso otros señalan la existencia de una estructura 4-dimensional (Chao et al., 2011).

Hasta dónde llega nuestro conocimiento, el MBI-HSS ha sido aplicado para medir burnout sobre distintas poblaciones; sin embargo, no existen muchas investigaciones que lo analicen en el ámbito farmacéutico.

Objetivo general. Comparar las diferencias entre técnicas clásicas y penalizadas de reducción de la dimensión en el análisis exploratorio de la validez factorial del cuestionario MBI en una muestra con baja absorción de varianza de profesionales farmacéuticos.

Participantes. Se recogieron las respuestas al cuestionario MBI-HSS de 51 farmacéuticos de Castilla y León (España), mediante muestreo no probabilístico de conveniencia. La participación en el estudio fue voluntaria. Los cuestionarios se administraron y recogieron en un sobre cerrado para garantizar la confidencialidad. Los datos se trataron de forma anónima. La muestra está formada mayoritariamente por mujeres (80,4%). El 58,8% de la muestra tenía entre 51-60 años, seguido de un 31,4% de personas con 41-50 años, 5,9% con más de 60 años y 3,9% entre 31-40 años.

Análisis de datos. La fiabilidad de cada una de las dimensiones del cuestionario fue analizada mediante el índice alfa de Cronbach. Se realizó una primera inspección de la estructura factorial a partir de la medida de adecuación muestral de KMO y la matriz de correlaciones entre ítems. La estructura factorial del cuestionario se evaluó comparando los resultados del análisis factorial con rotación Varimax, del análisis de componentes principales con Varimax y del análisis de componentes principales sparse. La elección de la rotación Varimax, que produce factores ortogonales, se debe al supuesto del MBI que supone independencia entre los factores.

Los datos se analizaron con el software libre R (Team, 2019). El análisis factorial se llevó a cabo mediante la función *factanal* de la librería *stats*, el análisis de componentes principales usando la función *princomp* de la librería genérica de R y el análisis de componentes principales sparse se ejecutó mediante la función *spca* de la librería *elasticnet*.

Resultados. La consistencia interna de las tres dimensiones refleja fuertes índices de fiabilidad para las subescalas de AE ($\alpha = 0{,}9$) y RP ($\alpha = 0{,}8$). Sin embargo, la fiabilidad estimada para la subescala de la DP fue baja ($\alpha = 0{,}1$). La prueba de esfericidad de Bartlett resultó altamente significativa (p=0,000), indicando una relación entre ítems y factores. A pesar de esto, el valor obtenido del KMO, aunque aceptable, no fue alto (KMO=0,683). La Figura 22 recoge la matriz de correlaciones entre ítems. Se observa una correlación directa entre ítems del AE y entre ítems del RP respectivamente (a excepción del ítem 4 "*Fácilmente comprendo cómo se sienten los pacientes*"). Además, se observa una relación inversa entre ítems del AE y del RP. Por el contrario, las correlaciones entre ítems de la subescala DP presentan correlaciones muy bajas, cercanas a 0.

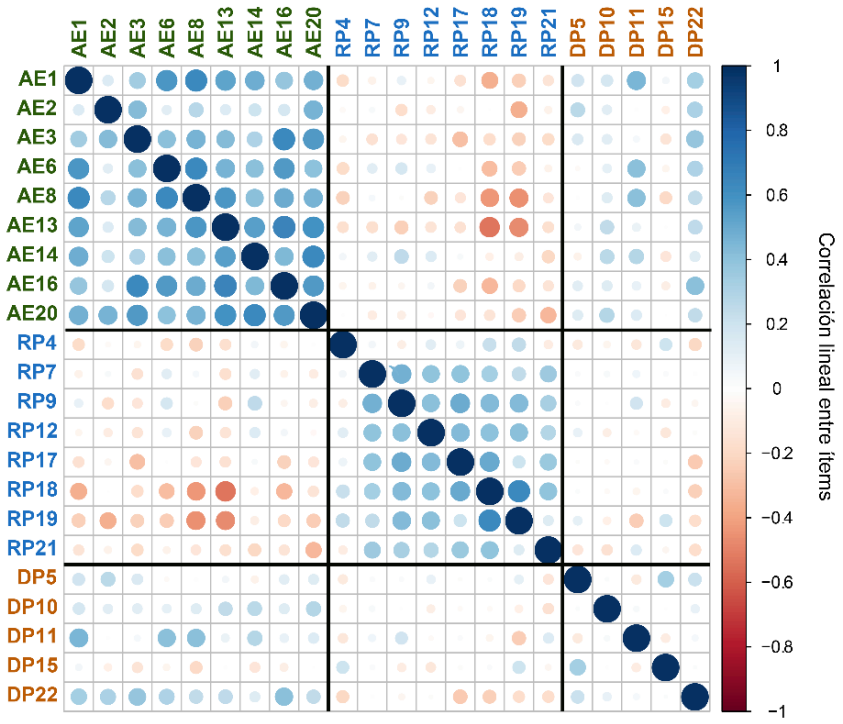


***Figura 22**. Matriz de correlaciones entre ítems de las 3 dimensiones del MBI (verde: agotamiento; azul: realización personal; naranja: despersonalización)*

Sobre la matriz de respuestas de los usuarios se lleva a cabo: i) EFA con rotación Varimax; ii) PCA con rotación Varimax y iii) SPCA con restricción sobre norma Elastic net de los vectores de carga (Tabla 11). En todos los análisis se retuvieron tres variables latentes y se supusieron como nulas aquellos valores

menores a 0,1 en valor absoluto para facilitar la interpretación. En el caso del EFA con rotación Varimax, el primer eje factorial absorbió un 14,9% de la varianza, el segundo un 13,8% y el tercero un 12,8%. Los ejes retenidos en el PCA explicaron un 26,91%, 17,76% y 9,54% de la varianza total. Las componentes principales sparse explicaron un 22,87%, 13,23% y 6,3% de la variabilidad. A pesar de que las dimensiones del EFA y del PCA explican un mayor porcentaje de la información original, la configuración de las saturaciones y las cargas respectivamente no reproducen la estructura teórica esperada.

1) *Estructura factorial del EFA.* Los ítems del AE contribuyen ampliamente a la formación de los tres ejes factoriales, mientras que los ítems del RP contribuyen sustancialmente al eje factorial 2, a excepción de los ítems RP-18 (*Me siento estimulado/a después de trabajar en contacto con los clientes*), RP-19 (*He conseguido muchas cosas útiles en mi profesión*) y RP-21 (*En mi trabajo trato los problemas que se me presentan con mucha calma*) a la formación de los ejes factoriales 2 y 1 respectivamente. Ahora bien, si se considerasen nulas aquellas saturaciones con valores absolutos por debajo de 0,3 (Tabla 12) podemos identificar la contribución de los ítems de AE al primer factor y los de RP al segundo. Sin embargo, los ítems de la DP no se agrupan en torno al tercer factor; esto es, no apuntan a la existencia de dicha subescala. El tercer factor se refiere a una combinación de AE y RP. De hecho, tan solo dos de los cinco ítems de la subescala de la DP presentan cargas no supuestas nulas.
2) *Estructura factorial del PCA.* En primer lugar es importante destacar los bajos coeficientes de todas las cargas a nivel general. No aparece ninguna variable especialmente relevante en la formación de las componentes, pues la mayoría tienen valores absolutos entre 0,2 y 0,3, a excepción del ítem RP-9 (*Creo que estoy influyendo positivamente, con mi trabajo, en la vida de los demás*) con una carga de 0,5. La primera componente principal se calcula como una combinación lineal de los ítems del AE y del RP y del ítem DP-11 (*Me preocupa el hecho de que este trabajo me esté endureciendo emocionalmente*), aunque este último con una contribución baja. De

manera similar ocurre con los ítems que más contribuyen a la formación de la segunda componente principal: principalmente ítems del RP, seguidos de ítems del AE y del ítem 11. Por último, como es lógico, parte de los ítems de la DP cargan en la tercera componente principal, que hasta ahora habían presentado coeficientes nulos en las otras dos variables latentes (a excepción del ítem DP-11). Al aumentar el límite del umbral de los coeficientes de los tres vectores de carga, a diferencia de los resultados del FA, no se mejora la interpretación de las componentes conforme a la estructura factorial esperada. Los ítems del AE contribuyen a las tres componentes, exceptuando el MBI-2, que no es explicado por ninguna de las nuevas dimensiones. Algo similar ocurre con los ítems del RP y tan solo los ítems DP-11 y DP-15 presentan cargas no nulas, contribuyendo a la formación tanto del eje 2 como del eje 3.

3) *Estructura factorial del Sparse PCA.* Se realizó un análisis de componentes principales sparse según el método spca con la penalización Elastic net, con el objetivo de seleccionar de manera automática las variables que con más fuerza cargan en cada una de las componentes. Así, se fuerza a que cada una de las componentes latentes sparse sea combinación de solo una parte de las variables originales. La matriz de cargas factoriales se muestra en la Tabla 11. En la primera dimensión se agruparon los ítems que miden AE, en la segunda dimensión contribuyen los ítems de la subescala RP, y por último en la tercera dimensión se agrupan los ítems de la DP. A diferencia de lo que ocurría con los métodos anteriores, el SPCA identifica los ítems de la despersonalización contribuyendo en la tercera componente sparse. Mejora la interpretación de la matriz de configuración en detrimento de la variabilidad captada a nivel global, además de confirmar la estructura de las tres dimensiones.

Las técnicas clásicas para el análisis de la validez factorial de escalas no son herramientas óptimas para datos en los que se da el efecto suelo (o efecto techo) en cuanto a variabilidad se refiere. Los farmacéuticos presentan bajos

niveles de despersonalización (efecto suelo) por lo que el PCA clásico fracasa a la hora de reproducir la estructura factorial. En este trabajo el Sparse PCA ha mostrado ser la estrategia más efectiva, postulándose como una alternativa útil cuando no se generan modelos interpretables con el FA o PCA. Soluciones basadas en la penalización de los vectores de carga son superiores a aquellas basadas en la variabilidad de los factores de carga.

***Tabla 11**. Matriz de saturaciones (FA con rotación Varimax y umbralización) y cargas (PCA, con rotación Varimax y umbralización, y SPCA) (AE=Agotamiento Emocional, RP=Realización Profesional, DP=Despersonalización)*

	Ítems	FA			PCA			SPCA		
		F1	F2	F3	PC1	PC2	PC3	SPC1	SPC2	SPC3
AE	AE-1	0,42		0,61	-0,24	0,16		-0,38		
	AE-2	0,47	-0,11					-0,10		
	AE-3	0,61	-0,18	0,20	-0,19			-0,23		
	AE-6	0,36		0,71	-0,31	0,33		-0,36		
	AE-8	0,39	-0,16	0,71	-0,36	0,18	0,14	-0,41		
	AE-13	0,55	-0,33	0,46	-0,32			-0,25		
	AE-14	0,64	0,19	0,33	-0,22	0,29	-0,31	-0,34		
	AE-16	0,60	-0,13	0,39	-0,26	0,14	0,15	-0,36		
	AE-20	0,90		0,14	-0,22	0,10	-0,24	-0,35		
RP	RP-4		0,10	-0,34	0,12		-0,56			
	RP-7		0,57	0,13		0,21			0,37	
	RP-9		0,78	0,21	0,17	0,50			0,58	
	RP-12		0,59		0,17	0,33	-0,17		0,44	
	RP-17	-0,11	0,62		0,22	0,33			0,33	
	RP-18		0,73	-0,48	0,34	0,21	-0,16		0,43	
	RP-19	-0,12	0,58	-0,39	0,27	0,12	-0,30	0,13	0,18	
	RP-21	-0,30	0,44		0,24	0,27	0,47		0,14	
DP	DP-5									0,81
	DP-10	0,28					0,11			
	DP-11		0,15	0,52	-0,12	0,21	0,21			-0,31
	DP-15			-0,17			-0,22			0,49
	DP-22	0,30	-0,15	0,24				-0,20		0,10
Varianza explicada			41,8%			54,20%			40,5%	

***Tabla 12**. Matriz de saturaciones (FA con rotación Varimax y umbralización de 0,3) y cargas (PCA, con rotación Varimax y umbralización de 0,15, y SPCA) (AE=Agotamiento Emocional, AP=Realización Profesional, DP=Despersonalización)*

	Ítems	FA			PCA			SPCA		
		F1	F2	F3	PC1	PC2	PC3	SPC1	SPC2	SPC3
AE	AE-1	0,42		0,61	-0,24	0,16		-0,38		
	AE-2	0,47						-0,10		
	AE-3	0,61		0,20	-0,19			-0,23		
	AE-6	0,36		0,71	-0,31	0,33		-0,36		
	AE-8	0,39		0,71	-0,36	0,18		-0,41		
	AE-13	0,55	-0,33	0,46	-0,32			-0,25		
	AE-14	0,64		0,33	-0,22	0,29	-0,31	-0,34		
	AE-16	0,60		0,39	-0,26		0,15	-0,36		
	AE-20	0,90		0,14	-0,22		-0,24	-0,35		
RP	RP-4			-0,34	,		-0,56			
	RP-7		0,57			0,21			0,37	
	RP-9		0,78		0,17	0,50			0,58	
	RP-12		0,59		0,17	0,33	-0,17		0,44	
	RP-17		0,62		0,22	0,33			0,33	
	RP-18		0,73	-0,48	0,34	0,21	-0,16		0,43	
	RP-19		0,58	-0,39	0,27		-0,30	0,13	0,18	
	RP-21	-0,30	0,44		0,24	0,27	0,47		0,14	
DP	DP-5									0,81
	DP-10									
	DP-11			0,52		0,21	0,21			-0,31
	DP-15						-0,22			0,49
	DP-22	0,30						-0,20		0,10
Varianza explicada			41,8%			54,20%			40,5%	

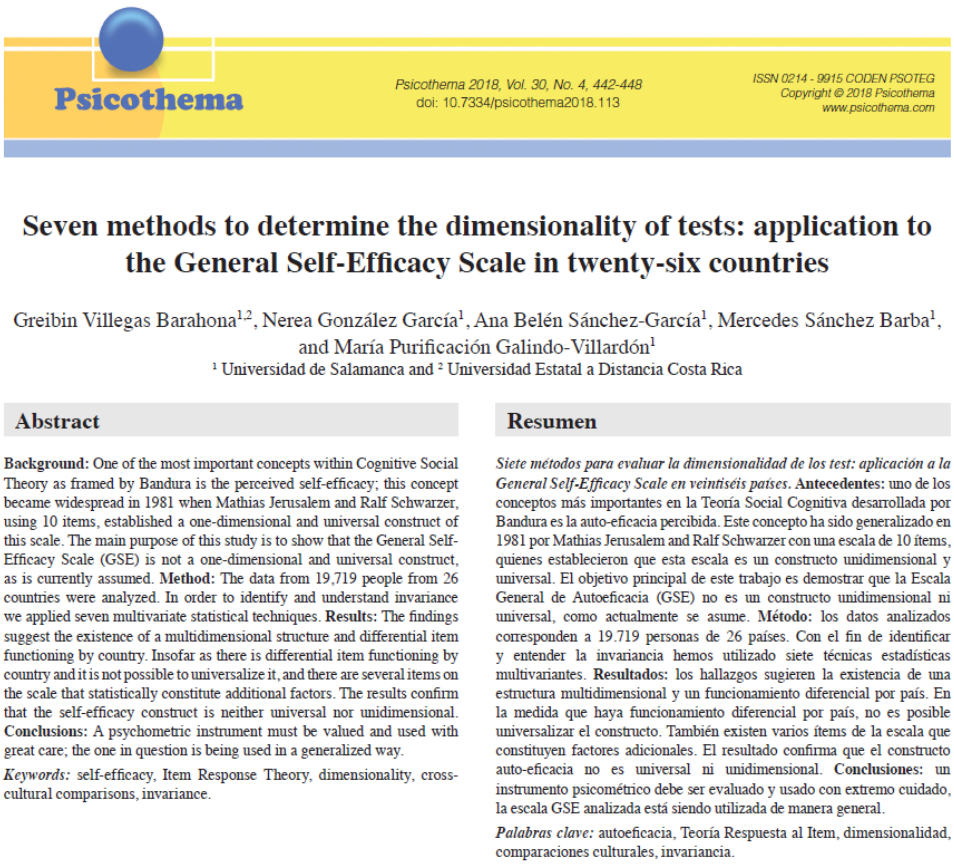
Estos resultados fueron expuestos en el **III Congreso Nacional de Psicología**, celebrado en Oviedo en julio de 2017, con una ponencia titulada "*Y si las técnicas factoriales clásicas no funcionan... ¿qué? Sparse PCA como alternativa a la metodología tradicional*". Me gustaría destacar que, a pesar de que finalmente no pudiéramos asistir, la presentación de este trabajo hizo que recibiéramos la invitación del decano de la facultad de psicología y del rector de la Universidad Nacional Federico Villarreal en Lima, Perú, a la **II Jornada de Psicometría y Evaluación Psicológica.** Se presentó también una

comunicación oral en el **XXVI Congreso Internacional de Psicología INFAD**, celebrado en Salamanca en junio de 2019, bajo el título "*Sparse ACP como herramienta alternativa de análisis de validez factorial en situaciones de baja absorción de varianza*".

Con un objetivo diferente se presentó una comparativa teórica del Sparse PCA frente a la Descomposición CUR, ambas diseñadas con el objetivo central de seleccionar variables automáticamente aunque desde distintos puntos de vista, en el **XXVI Simposio Internacional de Estadística 2016,** bajo el título "*Sparse PCA vs Descomposición CUR. Una comparación práctica a través del síndrome de Burnout*".

2.4.2 General Self-Efficacy Scale

A su vez, el Sparse PCA mostró su utilidad en este campo con una aplicación diferente, realizada en contribución con (Villegas Barahona, 2018), dentro del marco teórico de su tesis doctoral y que fue publicado en un artículo que forma parte de ella (Figura 23). En dicho estudio, Villegas propone un listado de siete técnicas multivariantes para examinar la multidimensionalidad de los test psicológicos (Barahona et al., 2018). Se revisaron metodológicamente técnicas clásicas como el FA, PCA, STATIS dual, análisis factorial confirmatorio multi-grupo (*Multi-group Confirmatory Factor Analysis*, MGCFA), teoría de respuesta al ítem (*Item Response Theory*, IRT) y funcionamiento diferencial del ítem (*Differential Item Functioning*, DIF), y junto a ellas se postula la utilidad del SPCA de Zou, Hastie y Tibshirani (2006) como técnica exploratoria para el estudio de la validez factorial.



Psicothema

Psicothema 2018, Vol. 30, No. 4, 442-448
doi: 10.7334/psicothema2018.113

ISSN 0214 - 9915 CODEN PSOTEG

www.psicothema.com

Seven methods to determine the dimensionality of tests: application to the General Self-Efficacy Scale in twenty-six countries

Greibin Villegas Barahona[1,2], Nerea González García[1], Ana Belén Sánchez-García[1], Mercedes Sánchez Barba[1], and María Purificación Galindo-Villardón[1]

[1] Universidad de Salamanca and [2] Universidad Estatal a Distancia Costa Rica

Abstract

Background: One of the most important concepts within Cognitive Social Theory as framed by Bandura is the perceived self-efficacy; this concept became widespread in 1981 when Mathias Jerusalem and Ralf Schwarzer, using 10 items, established a one-dimensional and universal construct of this scale. The main purpose of this study is to show that the General Self-Efficacy Scale (GSE) is not a one-dimensional and universal construct, as is currently assumed. **Method:** The data from 19,719 people from 26 countries were analyzed. In order to identify and understand invariance we applied seven multivariate statistical techniques. **Results:** The findings suggest the existence of a multidimensional structure and differential item functioning by country. Insofar as there is differential item functioning by country and it is not possible to universalize it, and there are several items on the scale that statistically constitute additional factors. The results confirm that the self-efficacy construct is neither universal nor unidimensional. **Conclusions:** A psychometric instrument must be valued and used with great care; the one in question is being used in a generalized way.

Keywords: self-efficacy, Item Response Theory, dimensionality, cross-cultural comparisons, invariance.

Resumen

Siete métodos para evaluar la dimensionalidad de los test: aplicación a la General Self-Efficacy Scale en veintiséis países. **Antecedentes:** uno de los conceptos más importantes en la Teoría Social Cognitiva desarrollada por Bandura es la auto-eficacia percibida. Este concepto ha sido generalizado en 1981 por Mathias Jerusalem and Ralf Schwarzer con una escala de 10 ítems, quienes establecieron que esta escala es un constructo unidimensional y universal. El objetivo principal de este trabajo es demostrar que la Escala General de Autoeficacia (GSE) no es un constructo unidimensional ni universal, como actualmente se asume. **Método:** los datos analizados corresponden a 19.719 personas de 26 países. Con el fin de identificar y entender la invariancia hemos utilizado siete técnicas estadísticas multivariantes. **Resultados:** los hallazgos sugieren la existencia de una estructura multidimensional y un funcionamiento diferencial por país. En la medida que haya funcionamiento diferencial por país, no es posible universalizar el constructo. También existen varios ítems de la escala que constituyen factores adicionales. El resultado confirma que el constructo auto-eficacia no es universal ni unidimensional. **Conclusiones:** un instrumento psicométrico debe ser evaluado y usado con extremo cuidado, la escala GSE analizada está siendo utilizada de manera general.

Palabras clave: autoeficacia, Teoría Respuesta al Ítem, dimensionalidad, comparaciones culturales, invariancia.

***Figura 23**. Publicación de la contribución realizada en el marco de la tesis doctoral de Villegas (2018) en la revista Psicothema (JCR 2018: 1,551 Q2; SJR 2018: 0,641 Q2). Aplicación del Sparse PCA como método para determinar la dimensionalidad de los test.*

Ambas investigaciones ponen de manifiesto la contribución del Sparse PCA como técnica desarrollada bajo la necesidad de hacer frente a la alta dimensionalidad a nuevos campos de aplicación, como puede ser el análisis de las propiedades de cuestionarios en Psicología y Educación, entre otras.

2.5 Contribuciones al análisis de datos genómicos

2.5.1 Análisis de los factores genéticos de los meningiomas: predicción de recidivas

Los meningiomas son los tumores cerebrales más frecuentes del sistema nervioso central en adultos. Son tumores habitualmente benignos de crecimiento lento que disponen de un tratamiento curativo eficaz: la resección quirúrgica. Sin embargo, en torno a un 20% de los casos presentan un comportamiento agresivo que da lugar a recurrencias del tumor. El modo habitual de proceder en la clasificación de los meningiomas es mediante sus características histopatológicas; sin embargo, existen tumores diagnosticados del mismo grado con distinto riesgo de recurrencia (reaparición del tumor), que puede ser debida a su gran heterogeneidad. La nueva clasificación de la OMS acepta la incorporación de parámetros genéticos al protocolo de diagnóstico. Por este motivo, el objetivo de este estudio es, mediante técnicas de selección de variables, identificar alteraciones en factores genéticos que ayuden a dilucidar las causas que conducen a la formación del meningioma, su progresión a un grado agresivo y la aparición de recidivas.

Para ello se analizó la expresión génica de un total de 38 muestras de meningioma, de la serie GSE43290 (HG-U133A Affymetrix Human Genome U133A Array) del repositorio GEO. De las 38 muestras analizadas, que presentaban cuatro patrones genéticos diferentes (complejo, diploide, monosomía-22, deleción), 11 presentaron aparición de recidiva. Los datos se normalizaron según el método RMA para que los valores de expresión genética de los microarrays fuesen comparables entre sí, eliminando el efecto de los errores sistemáticos acumulados debidos a la fluorescencia, impresión y experimento biológico. Este proceso de normalización constó de tres pasos: corrección de fondo (estimar y eliminar la intensidad del ruido de fondo), normalización (misma variación para todas las sondas) y sumarización (conversión de las sondas a genes).

Para poder realizar análisis posteriores de clasificación, se llevó a cabo la búsqueda de los genes más relevantes. La reducción de la dimensión, selección de variables e identificación de patrones se llevó a cabo mediante las siguientes etapas (Figura 24):

1. **Descomposición CUR.** Se realiza una primera selección de los genes que más información aportan mediante los valores de influencia generados a partir de la descomposición CUR. Esta permite identificar los genes que más información aportan.
2. **Regresión logística con penalización Elastic net con remuestreo**. Una vez seleccionados los genes más contribuyentes en cuanto a variabilidad se refiere mediante CUR, se buscó seleccionar aquellos que explicasen la aparición de recidiva en las muestras consideradas. Esta selección se realizó utilizando remuestreo Bootstrap, para analizar la estabilidad de la solución, haciendo inferencia mediante intervalos de confianza.
3. **Identificación de patrones de comportamiento comunes a distintos genes**. Una vez obtenida la matriz de datos final de genes relevantes por su aportación al modelo para la explicación de la aparición de recidiva, los datos se analizaron mediante el HJ-Biplot para identificar patrones de muestras en base al nivel de expresión presentada en los distintos genes. Se completó el estudio con el análisis clúster mediante el método Ward sobre las coordenadas del HJ-Biplot.

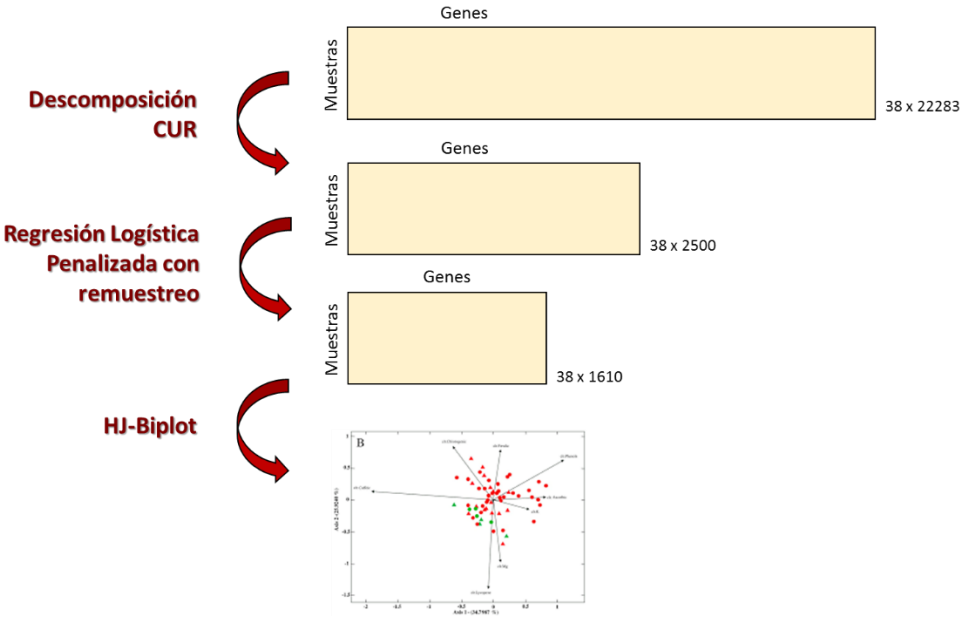


***Figura 24**. Etapas del análisis de selección de genes importantes en la aparición de recidivas*

La población de estudio está formada por un total de 38 muestras de 12 hombres y 26 mujeres con meningiomas (Tabla 13), con una edad media de 60 ⌖ 15,4 años. La edad mínima presentada fue de 23 años y la máxima de 84. Un 34,2% de las muestras tienen un patrón citogenético complejo, un 39,5% un patrón de pérdida diploide, un 15,8% de monosomía-22 y un 10,5% deleción. Sufrieron aparición de recidiva un 53,8% de los complejos, un 13,3% de los diploides, un 16,7% de la monosomía-22 y un 25% de las deleciones. Cabe destacar, que una de las muestras con patrón complejo sufrió una recidiva de grado III y otra de grado II, y una de las muestras diploides sufrió una recidiva de grado II. El resto de las recidivas sufridas fueron de grado I.

***Tabla 13**. Características basales clínicas y perfil citogenético de las 38 muestras en estudio*

Características Basales Clínicas	**(n=38)**
Edad (años)	60 ⌖ 15,4 (23,84)
Hombres, *n* (%)	12, (31,6%)
Mujeres, *n* (%)	26, (68,4 %)
Perfil Citogenético	
Complejo (%)	13, (34,2%)
Recidiva	
No	6, (46,2%)
Sí	7 (53,8%)
Grado I	5 (71,4%)
Grado II	1 (14,3%)
Grado III	1 (14,3%)
Diploide (%)	15, (39,5%)
Recidiva	
No	13, (86,7%)
Sí	2, (13,3%)
Grado I	1 (50%)
Grado II	1 (50%)
Grado III	-
Monosomía-22 (%)	6, (15,8%)
Recidiva	
No	5, (83,3%)
Sí	1, (16,7%)
Grado I	1 (100%)
Grado II	-
Grado III	-
Deleción	4, (10,5%)
Recidiva	
No	3, (75%)
Sí	1, (25%)
Grado I	1 (100%)
Grado II	-
Grado III	-

El estudio de la variabilidad inicial de las muestras se realizó mediante histogramas de las funciones de densidad y boxplot (Figura 25), donde se observa cómo las muestras no siguen el patrón de distribución normal simétrico y dando lugar a la desviación de las muestras de un comportamiento general.

Estos resultados evidencian la necesidad de normalización de los datos como corrección de la variabilidad existente. La variabilidad de las muestras debida los experimentos técnicos y biológicos fue suprimida usando la normalización RMA.

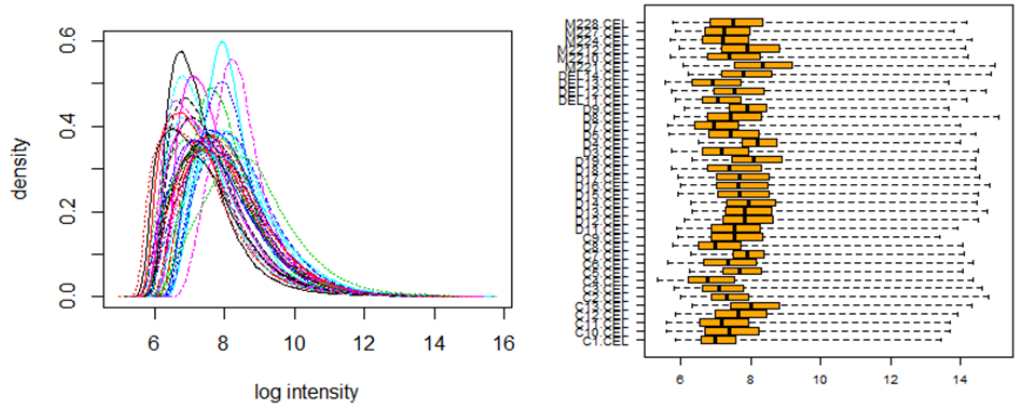


Figura 25. *Exploración inicial de los datos. A la izquierda los histogramas de la función de densidad logarítmica de cada muestra. A la derecha, boxplot de las 38 muestras con la distribución de sus niveles de expresión*

Reducción de la dimensión y selección de genes.

El estudio de los datos genéticos, que constituyen el perfil genético de los individuos, puede beneficiarse no sólo de una selección previa de genes importantes, sino también de una reducción de la dimensión posterior mediante factores latentes que reproduzcan la información original en un espacio de menor dimensión y complejidad en términos de interpretabilidad.

Descomposición CUR

El uso de una combinación adecuada de genes en el modelo es esencial para para obtener una alta precisión y exactitud en los resultados. Por ello, una vez pre-procesada la matriz de datos, se aplica la descomposición CUR (Mahoney & Drineas, 2009) como técnica de selección de genes relevantes. Esta realiza una reducción de la dimensionalidad mediante la aproximación de la matriz inicial, sin necesidad de buscar una estructura factorial. La selección de los genes se realiza mediante la definición de los factores de influencia, conocidos como *leverage scores*, de cada una de las sondas de genes en el modelo global. Los leverage representan la contribución de información que proporcionan a la variabilidad total. Permiten realizar una selección de las sondas

iniciales, almacenar los datos originales de los genes seleccionados en la matriz C y aproximar la matriz de datos original como:

$$X \approx CUR$$

La Figura 26 recoge los valores de los leverage calculados para cada una de las 22283 sondas iniciales. Interpretando el leverage como la cantidad de información que cada sonda aporta de la información total del modelo, se observa como la gran mayoría de ellos aportan una cantidad mínima de información en el modelo. El valor de los leverage es utilizado como criterio de selección de los genes de información relevante. El número óptimo de genes a seleccionar se decide, a partir de la Figura 26, localizando en la curva el punto en el que el decrecimiento se empieza a suavizar: 2.500 sondas.

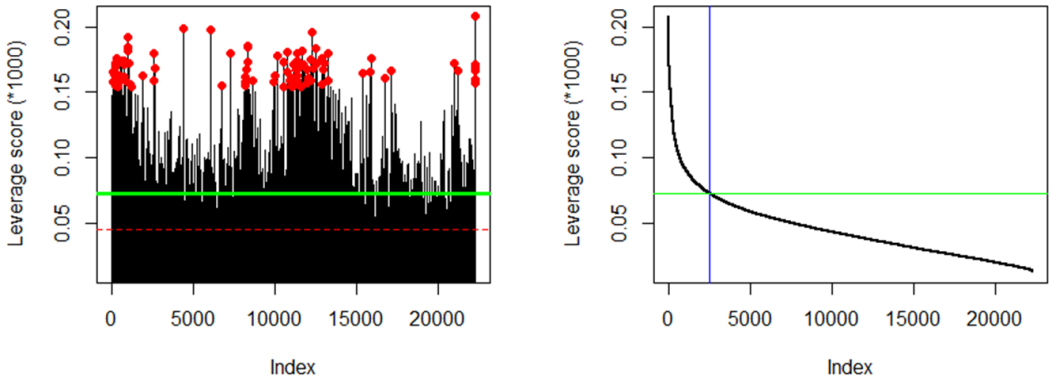


Figura 26. *Gráfico de leverage para cada uno de los 22283 genes. El gráfico de la izquierda representa el valor de la contribución de información de cada gen. A la derecha, el scree plot de la influencia de los genes ordenados de manera descendente.*

Reducida la dimensión de la matriz original, es necesario asegurarse de que la selección no introduce un error cuadrático medio elevado en el modelo y que las 2.500 variables seleccionadas son capaces de reproducir los datos iniciales con un error mínimo, cercano a 0. Es lógico que, a mayor número de genes seleccionados, el error cuadrático medio cometido en la aproximación por CUR será menor. En la Figura 27 se observa cómo a partir de 50 variables seleccionadas el error cometido en la aproximación de la matriz original es

mínimo y prácticamente nulo. Así, se confirma la bondad de los 2.500 genes seleccionados.

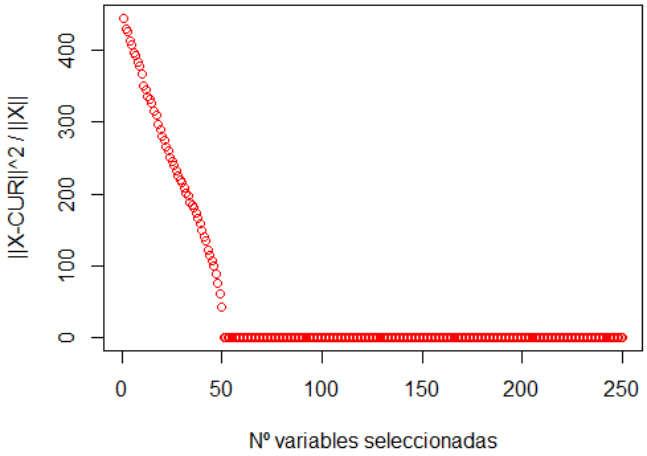


Figura 27. *Error cuadrático medio de la aproximación de la matriz original mediante la descomposición CUR en base al número de genes seleccionados*

Regresión logística penalizada con remuestreo Bootstrap

La introducción de un gran número de variables en un estudio (muchas de ellas no necesarias), puede suponer un sobreajuste del modelo teórico a definir. En general es necesario restringirlas, extrayendo las más relevantes. Esto proporcionará una reducción de la complejidad de los modelos, así como de la potencia computacional necesaria para llevarlos a cabo, facilitando la interpretación de los resultados. Existen múltiples maneras de identificar la contribución de cada gen al modelo y restringir el número de sondas seleccionadas, como la Descomposición CUR. A continuación, se restringirá el número de características a aquellas que sean más relevantes para la variable respuesta que queremos predecir: la aparición de recidivas tras un tumor cerebral.

Se plantea el modelo de regresión logística penalizada (Friedman, Hastie, & Tibshirani, 2010; Simon, Friedman, & Hastie, 2014) añadiendo al problema de mínimos cuadrados la penalización Elastic net (Zou & Hastie, 2005). La función objetivo que se plantea para el modelo de regresión logística penalizada con respuesta binomial es:

$$\min_{(\beta_0,\beta)\in\mathbb{R}^{p+1}} -\left[\frac{1}{N}\sum_{i=1}^{N} y_i(\beta_0 + x_i^T\beta) - log(1 + e^{(\beta_0 + x_i^T\beta)}) + \lambda[(1-\alpha)\|\beta\|_F^2/2 + \alpha\|\beta\|_1]\right]$$

En el caso de datos en los que el número de variables es muy superior al número de muestras, como es lo ocurrido en este caso, la regresión logística es inconsistente. Elastic net ayuda a aliviar estos problemas, así como seleccionar las variables a su vez. Surge como combinación de las penalizaciones Ridge (Hoerl & Kennard, 1970) y Lasso (Tibshirani, 1996) y combina las propiedades favorables de ambas. La penalización introducida en el modelo está controlada por un parámetro de regularización. Cuanto más alto sea su valor, mayor será la penalización introducida en el modelo. Si este parámetro es nulo, nos encontramos con el modelo original de regresión logística. El parámetro de regularización se seleccionará de manera que el error que introduzca en el modelo sea mínimo (Figura 28).

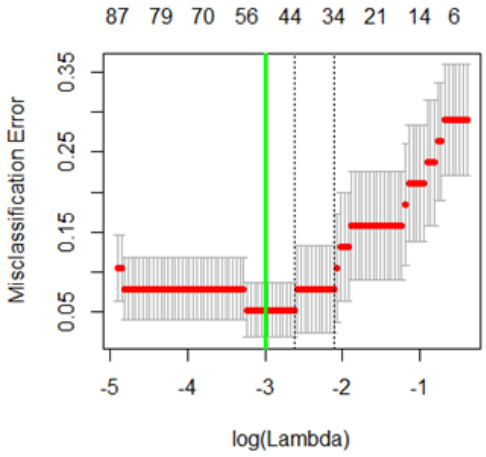


***Figura 28**. Selección del parámetro de regularización en el modelo de regresión logística penalizada*

Sobre la matriz de 38 muestras y 2.500 genes seleccionados por CUR se construye un modelo de regresión logística penalizada, con el objetivo de seleccionar aquellas sondas significativas a la hora de explicar la aparición de recidivas. El parámetro de regularización se escogerá de entre aquellos que supongan un menor error cuadrático medio del modelo definido. Además, se

hará inferencia estabilizando los resultados mediante remuestreo bootstrap, con 5.000 repeticiones y se seleccionarás los genes significativos en relación con la aparición de recidivas. La matriz final obtenida está formada por 1610 genes.

Identificación de Patrones Genéticos

El HJ-Biplot se estima mediante la Descomposición en Valores Singulares con una transformación inicial de doble centrado sobre los datos seleccionados a partir de la descomposición CUR y la regresión logística penalizada. Reteniendo tres dimensiones sobre la matriz de datos de 1610 sondas genéticas, se consigue representar los datos en una dimensión reducida que representa un 44,7% de la información total (Tabla 14). Esto es, transformando los datos de un espacio de 1600 variables a uno de 3 latentes se consigue representar prácticamente la mitad de la información total.

***Tabla 14**. HJ-Biplot con doble centrado. Varianza absorbida por los 3 ejes retenidos*

Ejes	Valores Propios	% Varianza	% Varianza Acumulada
Eje 1	5039,898	24,689	24,689
Eje 2	2265,617	11,099	35,787
Eje 3	1823,791	8,934	44,722

La representación de individuos en el espacio tridimensional generado por tres componentes principales se recoge en la Figura 29 y Figura 30. En rojo se presentan las muestras de individuos con aparición de recidiva. En ambas figuras se observa como el eje 1 es un gradiente de información que separa aquellos pacientes que han sufrido recidivas de mayor grado. Además, se observa la asociación de muestras que proceden del mismo paciente, como es el caso de C9-C2 (paciente 3), D19-D13 (paciente 1), C10-C11 (paciente 5) y C6-C13 (paciente 4). El plano 1-2 (Figura 29) muestra la separación de pacientes sin recidiva o recidiva de grado I de los pacientes con recidiva de mayor grado, por sus patrones de comportamiento en su expresión genética (que se analizará posteriormente). El plano 1-3 (Figura 30) muestra una separación más clara de los pacientes con o sin recidiva.

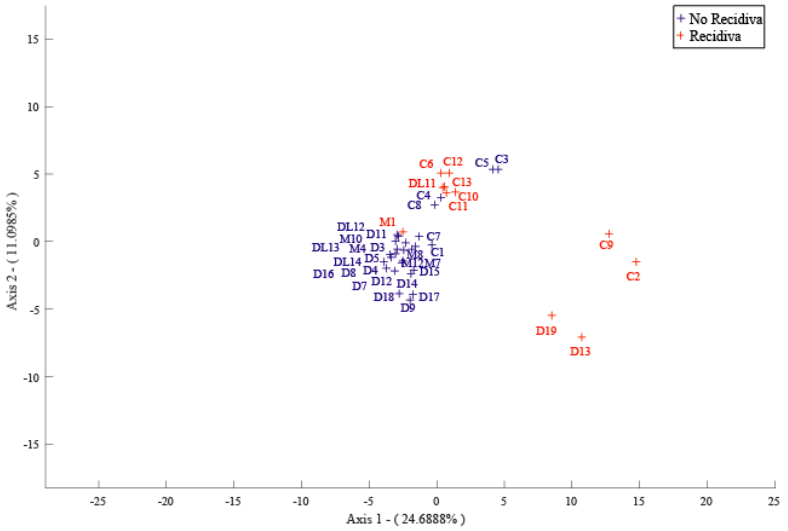


***Figura 29**. Representación de muestras de pacientes en el plano 1-2 generado por el HJ-Biplot. En rojo, pacientes con recidiva. En azul, pacientes sin reproducción de meningioma*

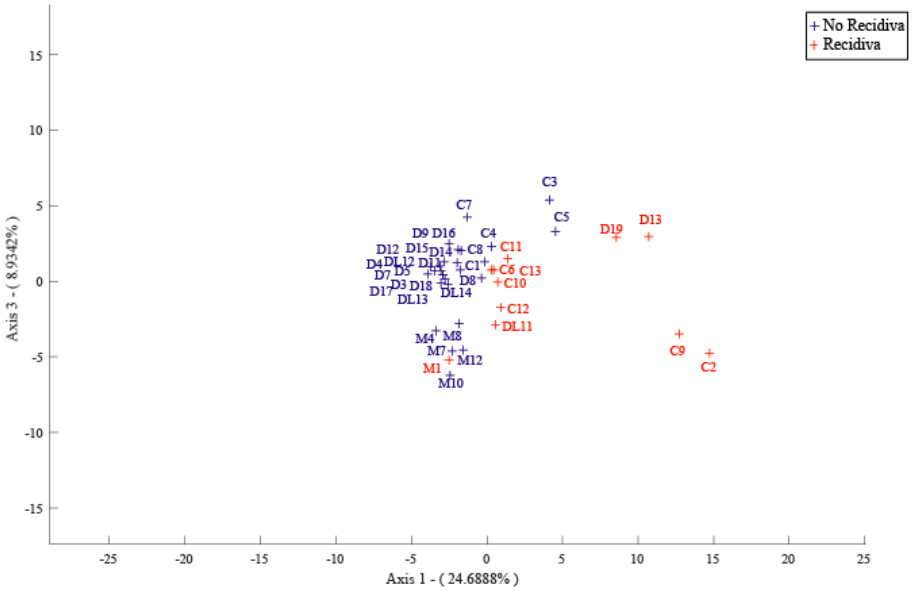


***Figura 30**. Representación de muestras de pacientes en el plano 1-3 generado por el HJ-Biplot. En rojo, pacientes con recidiva. En azul, pacientes sin reproducción de meningioma*

Aunque los resultados del Biplot suelen emplearse para la búsqueda de patrones de comportamiento su finalidad no es esa puesto que es un método de reducción de la dimensión y no de clasificación. Por ello, se aplica sobre las coordenadas obtenidas del HJ-Biplot un análisis de clúster jerárquico, mediante el método de Ward, con 5 clusters. El dendograma de la Figura 31 y la Tabla 15 recogen la agrupación de muestras de pacientes. Se contempla la asociación de las muestras en torno a 5 grupos de comportamiento similar: monosomías y

deleciones/diploides constituyendo un solo grupo de no recidivas, un grupo de recidivas de grado 1, entre las que se asociaban las muestras con patrón citogenético complejo, y finalmente dos grupos asociados a las recidivas de mayor grado: diploides y complejos respectivamente. Cabe destacar el hecho de que deleciones y diploides aparecen en un mismo grupo por su grado de homogeneidad biológico.

Tabla 15 *Clasificación de las muestras de pacientes en 5 clusters obtenidos mediante el método de Ward de clúster jerárquico sobre las coordenadas del Biplot. En negrita aparecen las muestras de pacientes con recidiva*

Interpretación		Grupos	Individuos
Recidiva Grado II-III	*Diploide*	**Cluster 1**	**D19, D13**
	Complejo	**Cluster 2**	**C9, C2**
Recidiva Grado I		**Cluster 3**	C5, C3, **DEL11, C12**, C8, C4, **C11, C6, C13, C10**
No recidiva	*Monosomía-22*	**Cluster 4**	M228, M224, M2210, M227, M2212, **M221**
	Deleción/Diploide	**Cluster 5**	D18, D9, D17, D16, D15, D14, C7, DEL14, DEL13, D8, D5, D7, D4, D12, C1, D11, DEL12, D3

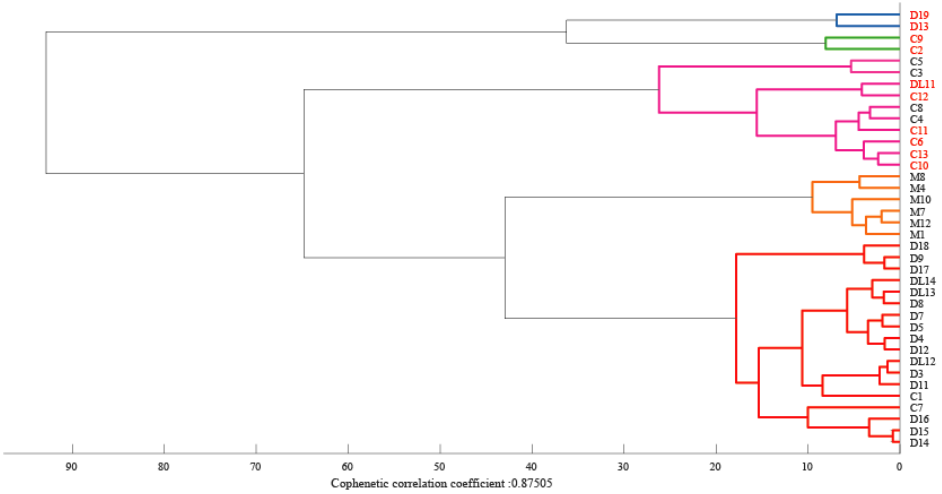


Figura 31. *Dendograma de clasificación. Método de Ward con 5 clusters a partir de las coordenadas del HJ-Biplot*

La representación de los individuos con las coordenadas Biplot según la clasificación obtenida por el análisis de clúster asocia esta agrupación en el espacio tridimensional (Figura 32, Figura 33). El eje 1 es un gradiente de alta

capacidad informativa. Tanto en el plano 1-2 como en el 1-3 sirve para caracterizar el grado de recidiva observado en los pacientes. En el primero de ellos, el plano 1-2, las muestras se asocian de manera que mayor positividad en el eje se traduce en mayor grado de recidiva. Se perciben 3 grupos claramente diferenciados: no recidiva, recidiva de grado 1 y recidiva de grado mayor. El eje dos se convierte en la dirección discriminante entre las muestras que padecen recidivas de grado I y el resto de las muestras. El plano 1-3 amplia la información anterior, agregando la capacidad informativa del eje 3, que es capaz de separar el cluster de muestras que no presentan recidiva por sus características, diferenciando monosomías de deleciones/diploides.

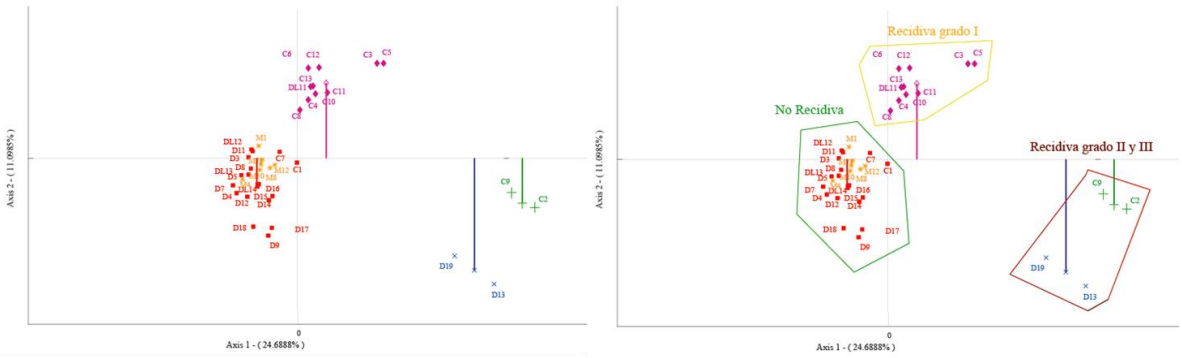


***Figura 32**. Representación de muestras de pacientes en el plano 1-2 generado por el HJ-Biplot (izquierda). Representación de muestras de pacientes en el plano 1-2 caracterizado por su grado de recidiva (derecha)*

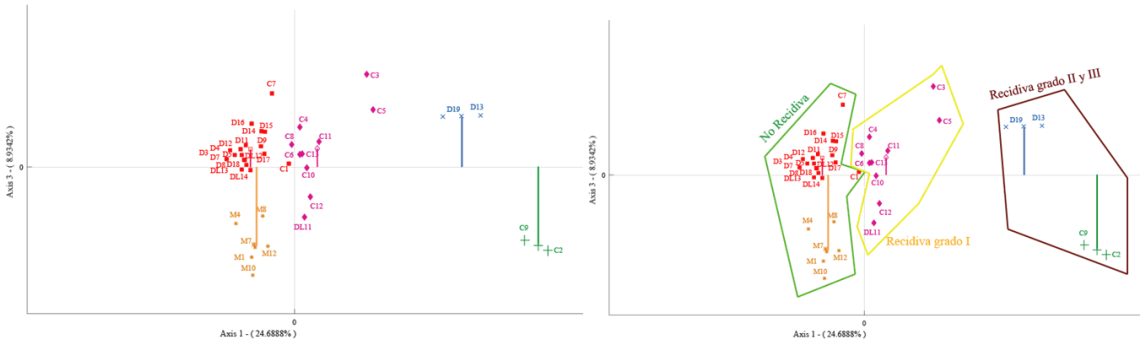


***Figura 33**. Representación de muestras de pacientes en el plano 1-3 generado por el HJ-Biplot (izquierda). Representación de muestras de pacientes en el plano 1-3 caracterizado por su grado de recidiva (derecha)*

Ahora bien, el hecho de realizar la clasificación sobre las coordenadas del Biplot supone la ventaja de poder asociar los patrones a las características que los definen; esto es, con la representación Biplot puede definirse la causa de la

formación de los 5 grupos, al conocer gráficamente con la expresión genética de qué genes están asociados y su correlación con las diferentes muestras (Figura 34, Figura 35). Se representan en el plano tridimensional aquellos genes cuya contribución a los ejes es superior a 600.

La representación de los planos 1-2 y 1-3 permite diferenciar los clústeres de recidivas de mayor grado (verde, azul), del clúster de recidivas de grado 1 y complejos (rosa) y del grupo de monosomías (naranja) y deleciones/diploides (rojo). Puede verse en el plano 1-2 cómo los clústeres de recidivas de mayor grado se caracterizan por valores altos en la expresión de genes como *H3F3A*, *CADVL, MT1G, MLEC* y *CBFB*, frente a valores bajos en las sondas del eje 1 izquierda, como *SLC26A2, LEPR, ADIRF, CLU, TMEM30B*. El clúster de recidivas de grado 1 se diferencia del resto de grupos por valores altos de expresión en genes de plano 2 como *RCN1, NBL1* y *CDH1*, y bajos de *FZD7*, *REEP1* y *RPL3*. Además, el cluster de complejos con recidiva de grado mayor que 1 se diferencia del cluster de recidivas diploides de mayor grado por su alto valor en la expresión genética de *RCN1*, asemejándose así con el resto de las muestras de patrón complejo y recidiva de grado I.

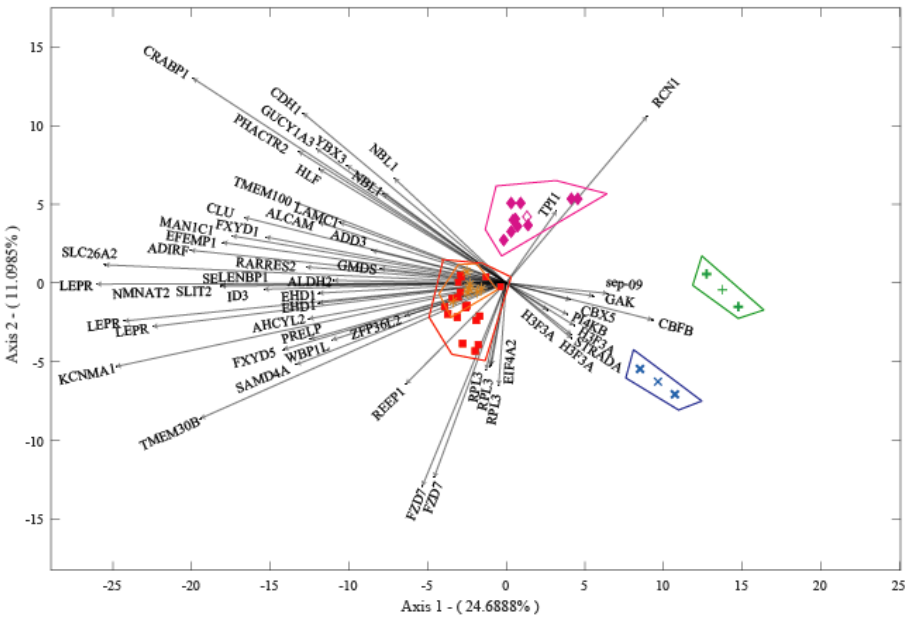


***Figura 34**. HJ-Biplot Plano 1 (24,7%) – 2 (11,1%)*

El plano 1-3 permite ver la evolución de los distintos clusters en base a su expresión genética a lo largo del eje 1, caracterizando a su vez las muestras de no recidivas en función de su perfil citogenético. Se observa un decrecimiento de la expresión genética en genes como *LEPR, SLITZ, CLU, FGI2* asociada a los distintos individuos, de manera que valores más altos en la expresión de estos genes se asocia a deleciones y monosomías, valores medios a recidivas de grado 1 y valores bajos a recidivas de mayor grado. Recíprocamente, los valores más altos en la expresión *CBFB, CANX, PPT1* y *GAK* los presentan los pacientes con recidivas de mayor grado y réplicas de recidivas, y los valores bajos en esta expresión los pacientes sin recidivas (deleciones y monosomías).

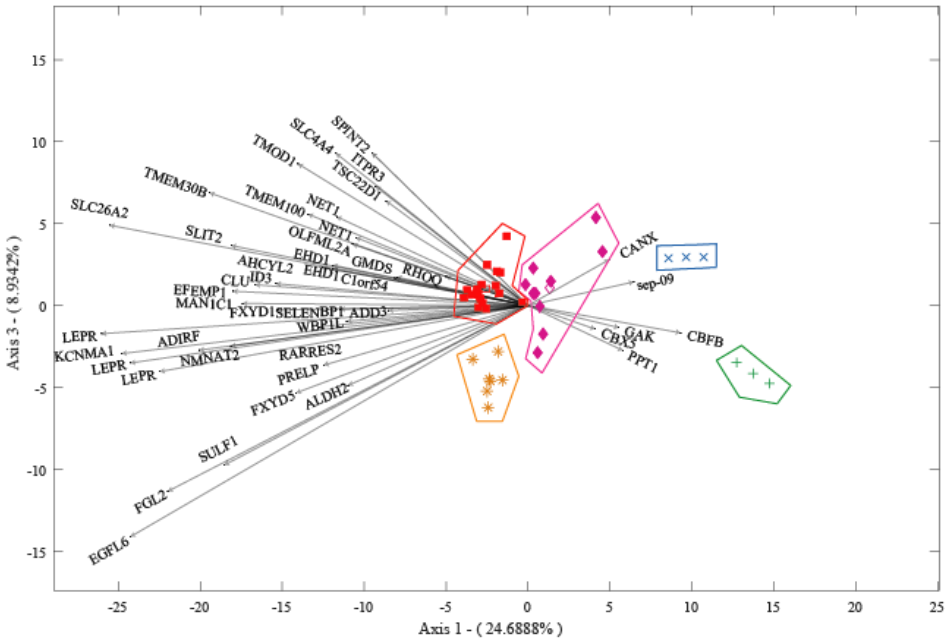


***Figura 35**. HJ-Biplot Plano 1 (24,7%) – 3 (8,9%)*

Mediante una técnica de análisis novedosa, adaptada al estudio de matrices de altas dimensiones, se han conseguido establecer patrones en la heterogeneidad de los meningiomas y aparición de recidivas. En genética, como en muchas otras disciplinas, se espera que sólo una pequeña parte de los genes estén implicados en los procesos biológicos de estudio y son necesarias técnicas que los detecten. Para ello, la descomposición CUR ha permitido realizar una primera selección de las 2.500 sondas que más información aportan en el modelo real. Por otro lado, debido a que las bases de datos utilizadas son de

altas dimensiones, y las técnicas clásicas son inconsistentes para su estudio, es necesario el uso de métodos tradicionales modificados mediante técnicas de penalización. De entre las 2.500 sondas seleccionadas, la regresión logística penalizada ha facilitado la identificación de 1610 sondas implicadas en el proceso de aparición de recidivas. Estos resultados han sido estabilizados mediante el remuestreo bootstrap sobre los resultados de los coeficientes de regresión, con 5.000 repeticiones. Finalmente, la unión del HJ-Biplot como método de reducción de la dimensión coordinado con el Análisis de Clúster jerárquico ha permitido identificar patrones en las muestras de estudio. El método Ward sobre las coordenadas del método Biplot ha dado lugar a la detección de 5 grupos de comportamiento similar: monosomías y deleciones/diploides constituyendo un solo grupo de no recidivas, un grupo de recidivas de grado 1, entre las que se asociaban las muestras con patrón complejo, y finalmente dos grupos asociados a las recidivas de mayor grado: diploides y complejos respectivamente. El HJ-Biplot favorece la determinación de las causas del agrupamiento de los 5 colectivos. Mediante su interpretación gráfica, se define la expresión genética que da sentido a los grupos establecidos en el análisis de clasificación.

Parte de estos resultados fueron presentados en la conferencia "**XVI Spanish Biometric Conference CEB**", que tuvo lugar en septiembre de 2017 en Sevilla.

2.5.2 Contribución al diagnóstico histológico de gliomas astrocíticos difusos mediante su perfil multivariante

Los gliomas son los tumores cerebrales más comunes y en concreto los tumores astrocíticos, que se desarrollan a partir de células gliales, representan el 80% de las neoplasias cerebrales diagnosticadas. Según el criterio de la Organización Mundial de la Salud (OMS), los gliomas astrocíticos se clasifican en tumores no difusos y difusos clasificados en cuatro grados de malignidad (de grado I a IV). El grupo de gliomas astrocíticos difusos incluye astrocitomas grado II (DA, OMS II), astrocitoma anaplásico grado III (AA, OMS III) y glioblastoma multiforme grado IV (GBM, OMS IV). Los GBM son la forma más agresiva y maligna de astrocitomas, con una alta tasa de mortalidad con tiempos de supervivencia inferiores a 5 años en la mayoría de los casos.

Estos subtipos de tumores astrocíticos difusos son extremadamente heterogéneos y su pronóstico y tratamiento son muy diversos. El diagnóstico patológico es el objetivo estándar del examen clínico práctico, aunque las variaciones intratumorales e interobservadores complican la clasificación histológica de los tumores. Por lo tanto, el reconocimiento de las alteraciones genéticas en los subtipos de glioma astrocítico difuso podría contribuir a refinar su diagnóstico histológico actual. El objetivo del presente era examinar los perfiles de expresión génica asociados a tres subtipos de gliomas astrocíticos difusos (DA, AA, GBM) para identificar las alteraciones genéticas de los GBM, que los diferencian de los astrocitomas de grado inferior.

Se recopilaron 176 muestras de tumores astrocíticos (129 GBM y 47 DA y AA) propias de 7 bases de datos diferentes de GEO: una cohorte propia del laboratorio de la Dra. Tabernero con el que se realizó este estudio (GSE43289) y 6 series adicionales (GSE4290, GSE15824, GSE9171, GSE13041, GSE2817, GSE29796). Para identificar bases de datos potenciales se planteó una estrategia de búsqueda en el GEO que contuviera los descriptores en ciencias de la salud en inglés (MeSH) “Astrocytoma”, “Gene expression” y “Humans”. Se identificaron un total de 1330 estudios entre los años 2000 y 2017. Se consideraron como válidas aquellas series que median perfil de expresión de genes por arrays, recogidas sobre la plataforma HGU133Plus2. De las 19 series válidas resultantes se consideraron un total de 7 para el estudio de la identificación de genes. Las 176 muestras recogidas histológicamente fueron clasificadas, siguiendo el criterio de la WHO, como: DAs (n=19, 11%), AAs (n=28, 16%) y GBMs (n=129, 73%).

Normalización de los datos y análisis estadístico

La matriz de expresión génica que se analizó contenía la información de 44.723 sondas de genes, respectivas a 21.336 genes (el resto de sondas no fueron consideradas en el estudio por ser sondas control o no disponer de información sobre ellas). Los datos de expresión fueron normalizados mediante la medida RMA y se eliminó la variabilidad debida a la pertenencia de las muestras a series distintas mediante el procedimiento ComBat, disponible en R en la librería *sva*. Para mostrar la eliminación de la variabilidad entre series se realizó un PCA clásico. En los gráficos de puntuaciones factoriales (Figura 36)

se observan las diferencias antes y tras haber eliminado la variabilidad debida a las series.

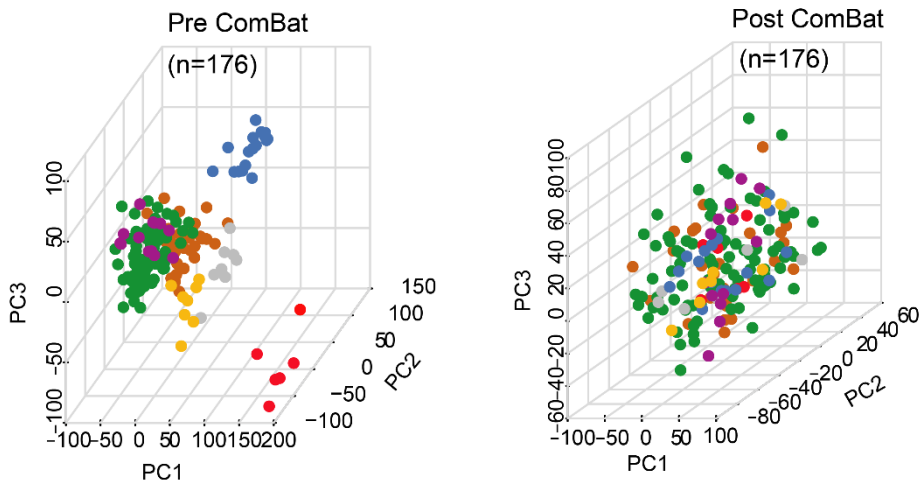


***Figura 36**. Puntuaciones factoriales obtenidas en el PCA sobre la matriz de datos original antes y tras aplicar el método de ComBat para eliminar la variabilidad específica de cada una de las series del GEO*

Selección de los genes más relevantes

Se seleccionaron los genes más relevantes en la posible identificación de las tres subclases de tumores a partir de dos pasos secuenciales:

1) Se seleccionaron las 1.000 sondas de genes que mostraban mayor variabilidad en términos multivariantes a partir de los leverage de la descomposición CUR, que miden la influencia de las variables en términos de varianza.
2) Se realizó un estudio de las diferencias entre pares de subgrupos de gliomas (DA vs AA, AA vs GBM, DA vs AA) mediante los valores *fold change* (FC) corregidos por FDR (*false discovery rate*) para determinar sondas de genes expresadas diferencialmente entre los gliomas. Por otro lado, se aplicó la regresión ordinal penalizada con el fin de detectar los genes envueltos en la discriminación de los tres subconjuntos de tumores. Las comparaciones por pares mostraron más de 500 sondas de genes con expresión génica significativamente diferente entre los tres grupos analizados (Figura 37). No se encontraron sondas con diferencias significativas entre DAs y AAs, de donde se sigue que podrían ser considerados el mismo tipo molecular de tumores. La comparación entre DAs y GBMs evidenció 409 probes significativas y la comparación entre

AAs y GBMs 406 significativas. Los genes con mayores diferencias de expresión en la comparación DA-GBM fueron *ETNPPL, SFRP2, SH3GL2* (más expresados en DAs) y *CHI3L1, COL1A1, COL3A1, POSTN, NNMT, PTX3, IGF2BP3, COL1A2, XIST, MALSU1, HS3ST3B1, IGFBP2, SHOX2, TOP2A, VEGFA* (más expresados en GBMs). En la diferenciación entre AAs y GBMs, *ETNPPL* fue el único gen con mayores expresiones en AAs, frente a *CHI3L1, COL1A1, COL3A1, POSTN, NNMT, PTX3,* más expresados en GBMs. La regresión ordinal penalizada mostró 14 sondas significativas en la explicación de las tres clases tumorales simultáneamente (Figura 37).

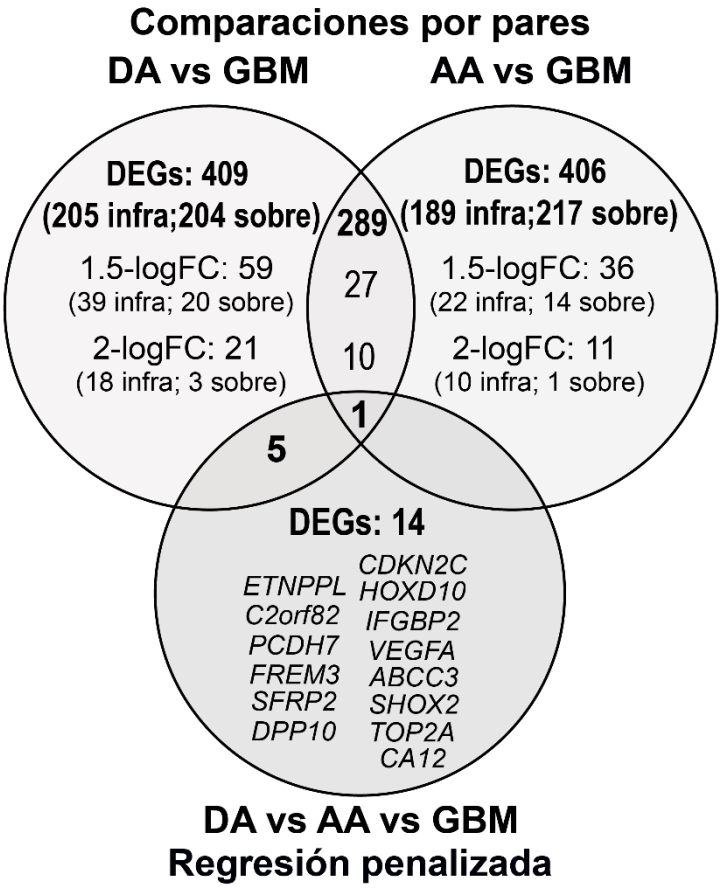


Figura 37*. Sondas de genes identificadas como significativas en las comparaciones por pares y en la regresión logística penalizada, en la distinción de los tres subconjuntos de muestras astrocíticas difusas.*

En total, se identificaron y seleccionaron 26 genes expresados diferencialmente usando ambos enfoques (Tabla 16).

***Tabla 16**. Expresión génica de los 26 genes expresados diferencialmente entre los tres subtipos de gliomas analizados*

Gene		Chromosomal location	Gene expression values DA (n=19)	AA (n=28)	GBM (n=129)
XIST	X inactive specific transcript	Xq13.2	6,2 (1,4)	9,3 (3,5)	8,5 (3,5)
ETNPPL	Ethanolamine-phosphate phospho-lyase	4q25	9,9 (1,6)	9,4 (2,4)	7,2 (1,7)
SFRP2	Secreted frizzled related protein 2	4q31.3	9,5 (1,5)	8,2 (2,2)	7,1 (1,5)
PCDH7	Protocadherin 7	4p15	9,3 (0,9)	9,1 (1,3)	7,6 (1,6)
SH3GL2	SH3 domain containing GRB2 like2	9p22	9,0 (1,1)	8,4 (1,6)	7,0 (1,5)
SNORC	Secondary ossification center associated regulator of chondrocyte	2q37.1	7,4 (1,3)	7,3 (1,5)	6,1 (1,0)
DPP10	Dipeptidyl peptidase like 10	2q14.1	7,4 (1,2)	7,2 (1,2)	5,7 (1,2)
FREM3	FRAS1 related extracellular matrix 3	4q31.21	5,7 (1,4)	4,7 (1,2)	4,2 (0,6)
CHI3L1	Chitinase 3 like 1	1q32.1	9,1 (2,5)	9,7 (2,8)	11,9 (1,8)
COL1A1	Collagen type I alpha 1 chain	17q21.33	8,0 (2,4)	7,9 (2,2)	10,1 (1,7)
IGFBP2	Insulin like growth factor binding protein2	2q35	8,2 (1,4)	8,8 (1,9)	10,7 (1,2)
COL3A1	Collagen type III alpha 1 chain	2q31	7,3 (2,4)	7,5 (2,1)	9,7 (1,6)
CDKN2C	Cyclin dependent kinase inhibitor 2C	1p32	8,5 (0,8)	9,3 (1,2)	9,9 (0,9)
COL1A2	Collagen type I alpha 2 chain	7q22.1	7,6 (2,4)	7,8 (2,0)	9,8 (1,4)
VEGFA	Vascular endothelial growth factor A	6p12	7,6 (0,9)	8,0 (1,2)	9,7 (1,5)
CA12	Carbonic anhydrase 12	15q22	8,1 (1,2)	8,4 (1,1)	9,6 (1,2)
NNMT	Nicotinamide N-methyltransferase	11q23.1	7,4 (2,0)	7,5 (1,8)	9,6 (1,8)
ABCC3	ATP binding cassette subfamily C member 3	17q22	7,7 (0,8)	8,0 (1,1)	9,0 (1,1)
POSTN	Periostin	13q13.3	6,6 (2,4)	6,8 (2,3)	8,9 (2,4)
TOP2A	Topoisomerase DNA II alpha	17q21.2	6,4 (0,9)	7,3 (1,8)	8,4 (1,4)
PTX3	Pentraxin 3	3q25	6,4 (1,5)	6,3 (1,5)	8,4 (1,7)
HOXD10	Homeobox D10	2q31.1	6,3 (0,4)	6,9 (1,2)	8,0 (1,3)
HS3ST3B1	Heparan sulfate-glucosamine3-sulfotransferase 3B1	17p12	5,6 (1,1)	6,1 (1,5)	7,7 (1,7)
IGF2BP3	Insulin like growth factor 2 mrna binding protein 3	7p11	5,4 (1,0)	6,2 (1,9)	7,7 (1,5)
MALSU1	Mitochondrial assembly of ribosomal large subunit1	7p15.3	5,5 (1,1)	6,2 (1,8)	7,6 (1,5)
SHOX2	Short stature homeobox 2	3q25.32	5,2 (0,8)	6,2 (1,7)	7,4 (1,3)

Patrones multivariantes

En segundo lugar, se llevó a cabo la factorización no negativa de la matriz formada por las 176 muestras tumorales y las 26 sondas significativas identificadas en las comparaciones por pares y en la regresión multivariante. Se identificaron dos grupos de muestras (correspondientes a DAs&AAs y GBMs) (Figura 38).

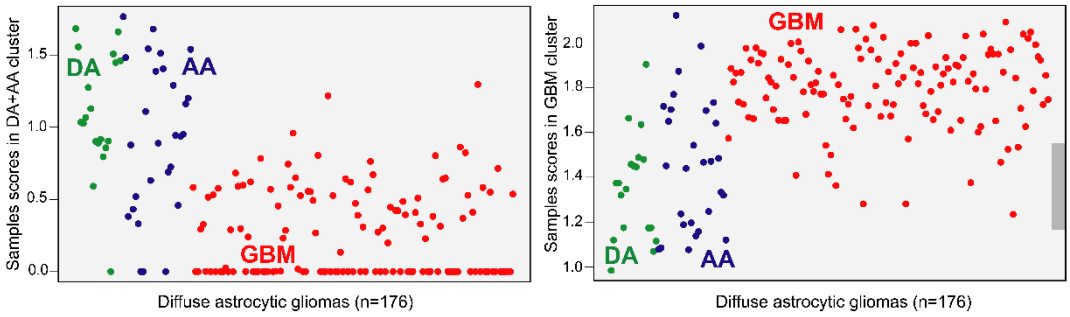


***Figura 38**. Puntuaciones de las muestras sobre los dos clusters retenidos*

La importancia de cada gen a la formación del cluster; es decir, la matriz de cargas factoriales, contiene la información de los genes que generan cada uno de los factores latentes. En la Figura 39 se vislumbran dos grupos de genes claramente diferenciados por sus relaciones. Los genes característicos del grupo de DAs&AAs son *ETNPPL, SH3GL2, PCDH7, SFRP2, DPP10, SNORC, FREM3* seguidos de una baja contribución de *CDKN2C*, *XIST* y *ABCC3.* La lista de genes que principalmente explican el grupo de GBMs incluye *CHI3L1, IGFBP2, COL1A1, VEGFA, COL1A2, CA12, CDKN2C, COL3A1, NNMT, POSTN, ABCC3, TOP2A, PTX3, XIST, HOXD10, IGF2BP3, MALSU1, HS3ST3B1* y *SHOX2. CHI3L1*, *COL1A1*, *VEGFA* y *IGFBP2* son genes ya resaltados como marcadores genéticos asociados a este tipo de tumores cerebrales. Entre los genes novedosos que se reportan aquí destacan *PCDH7, DPP10* y *SNORC.*

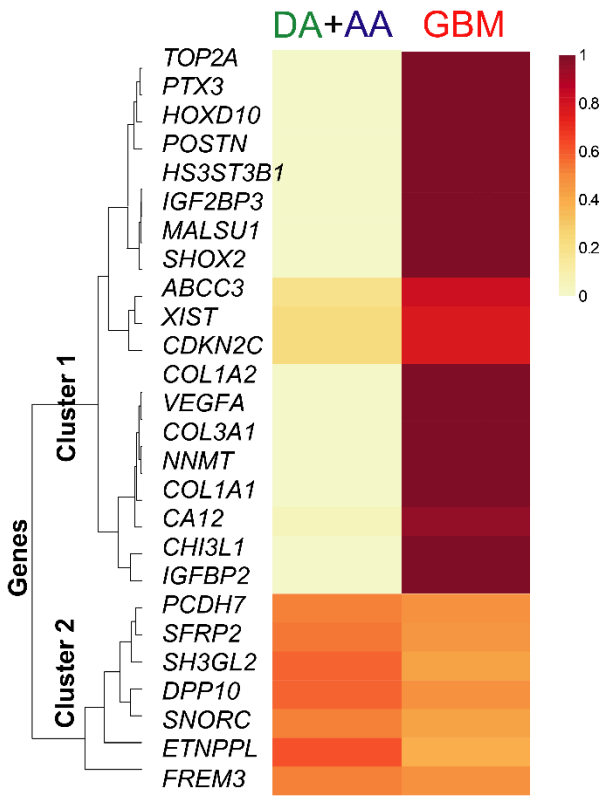


Figura 39. *Contribuciones de los 26 genes seleccionados a la formación de los dos clusters*

Finalmente, se realizó un Biplot canónico como herramienta de visualización de una matriz multivariante con estructura de grupos conocida a priori en un espacio de dos dimensiones, con el objetivo de detectar los genes con máxima capacidad discriminante entre grados histológicos tumorales (Figura 40).

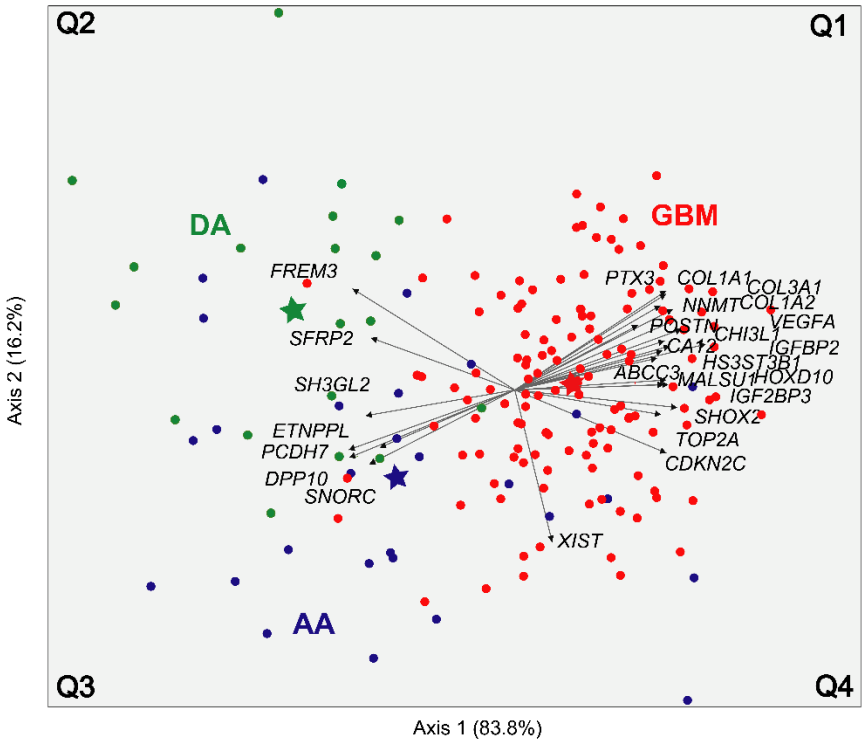


***Figura 40**. Plano factorial 1-2 del Biplot canónico sobre la matriz de 26 genes seleccionados*

La Figura 41 contiene un esquema global del proceso seguido en la investigación. La mayoría de los genes contribuyeron sustancialmente a la formación del eje factorial 1, que proporcionó una diferenciación de GBM de las muestras de DAs&AAs. Esta separación se explica principalmente por *ETNPPL, TOP2A, SHOX2, IGF2BP3, MALSU1, HOXD10, IGFBP2, CHI3L1, VEGFA* y *CHI3L1*. Además, el cluster DAs&AAs mostró valores de expresión más altos de *DPP10, SH3GL2, PCDH7* y *SNORC*. Por el contrario, las GBMs mostraron valores más altos de *IGFBP2, IGF2BP3, SHOX2* y *VEGFA*, entre otros. El eje vertical 2 fue relevante para diferenciar los DAs de los gliomas AAs. Esta dimensión estaba explicada en gran medida por *XIST* y *FREM3* y por genes que contribuyen a la formación de ambos ejes, como *SFRP2, PTX3, COL1A1, COL3A1, COL1A2, NNMT* y *POSTN,* entre otros. La baja expresión de *XIST* en los DAs los separó de los tumores AAs. La expresión *FREM3* y *SFRP2* discriminó a los DAs de los AAs, siendo mayor en los primeros. Similar a los resultados obtenidos con NMF, el Biplot canónico expuso la baja contribución de *CDKN2C* y *XIST* al clúster DAs&AAs. Esto puede ser debido a que algunas muestras de AAs comparten intensidades de expresión similares con GBMs.

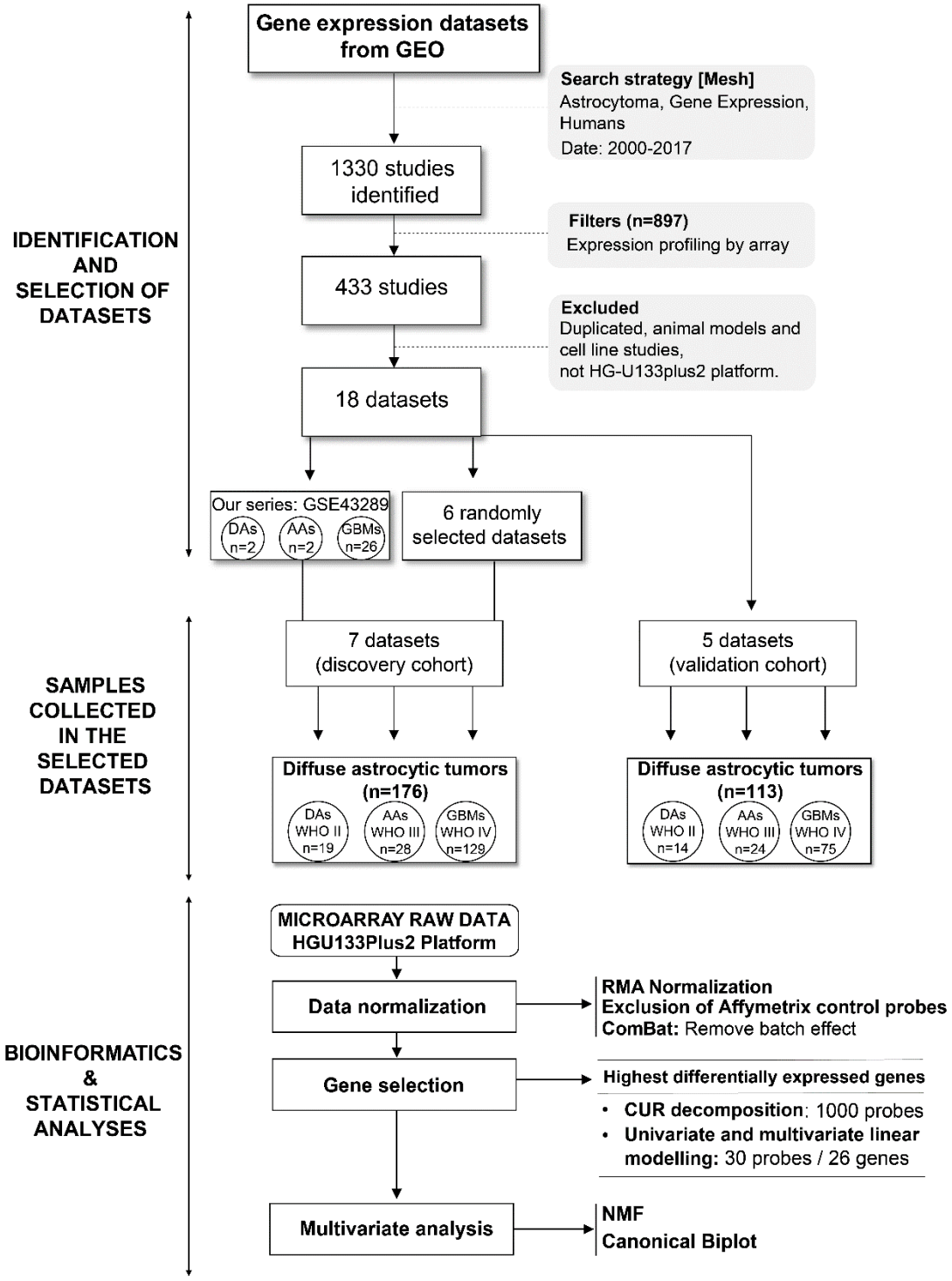


***Figura 41**. Esquema global del proceso llevado a cabo para la determinación de genes específicos en la diferenciación de gliomas difusos astrocíticos*

Este estudio nos permitió determinar un grupo potencial de genes discriminantes para la clasificación difusa de gliomas astrocíticos, ilustrando el valor del análisis multivariante en la caracterización genómica. Nos permitió concluir que el diagnóstico de astrocitomas futuro debería integrar datos

moleculares novedosos, características histológicas y genéticas que se correlacionen con la malignidad y supervivencia del paciente. Se refuerza así la importancia de marcadores genéticos previos y al mismo tiempo se plantea un nuevo listado de genes novedosos para la comprensión de la progresión tumoral astrocítica. *XIST*, *FREM3* y *SFRP2* fueron los genes que mejor discriminaban entre DAs y AAs gliomas, pero con pequeñas diferencias en la expresión de los genes. Podrían ser considerados el mismo subtipo genético. Los GBM malignos se distinguieron principalmente de los DA y AA por la sobreexpresión de los genes *CHI3L1, IGFBP2, VEGFA, COL, NNMT, HOXD10, SHOX2, IGF2BP3* y *MALSU1*. Encontramos genes *ETNPPL, SH3GL2, PCDH7, SFRP2, DPP10, SNORC* y *FREM3* infraexpresados en GBM. La angiogénesis y las funciones celulares inflamatorias se vieron específicamente afectadas en GBM.

Estos resultados forman parte de los proyectos (CB16/12/00400 e ISCIII PI16/0476), financiados por el Instituto de Salud Carlos III. Parte de estos resultados fueron presentados en póster en el “**IV International workshop on proximity data, multivariate analysis and classification”**.

Actualmente, los resultados globales bajo el título “M*ultivariate analysis reveals differentially expressed genes among distinct subtypes of diffuse astrocytic gliomas: diagnostic implications*”, están sometidos en forma de artículo de investigación en la revista *Scientific Reports*.

CAPÍTULO 3

PROYECCIÓN DE UN VECTOR SOBRE UN CONJUNTO DE RESTRICCIONES CONVEXO

La revisión de la literatura ha permitido exponer en el capítulo anterior algunas de las técnicas penalizadas más utilizadas, como el Sparse PCA o Sparse Biplot y otras alternativas dedicadas a la selección de variables como la descomposición CUR o la NMF, más comúnmente conocidas como técnicas sparse. Sin embargo, como se ha dicho anteriormente, la técnica por excelencia de la estadística multivariante es la SVD.

A nivel práctico, su uso está ligado a disciplinas muy diversas: análisis de texto (Elder et al., 2012; Hyung, Lee, & Lee, 2014; Li & Xue, 2018; Thara & Sidharth, 2017), genética (Franceschini et al., 2016; Lisowska et al., 2016; Meuwissen, Indahl, & Ødegård, 2017) neuro-imagen (Juneja et al., 2016; Zhan et al., 2015), medioambiente (Liu, Harley, Bergés, Greve, & Oppenheim, 2015), política (Skillicorn & Leuprecht, 2015; Wijaya, Billah, & Ahn, 2018), historia (Hussein, 2015)…

A nivel teórico, se trata del corazón de muchas de las técnicas multivariantes más utilizadas; tanto de dos vías: PCA (Jolliffe et al., 2016), Análisis Biplot (Gabriel, 1971; Galindo, 1986), Análisis de Correspondencias (Benzécri, 1973; Greenacre, 2017) o Escalamiento Multidimensional (MDS) (Borg & Groenen, 2003), como de su extensión al análisis de matrices multivía: Análisis de Correlación Canónica (CCA, *Canonical Correlation Analysis*) (Hotelling, 1936), Análisis Factorial Múltiple (MFA, *Multiple factor analysis*) (Abdi, Williams, & Valentin, 2013; Escofier & Pagès, 1994), STATIS (L'Hermier des Plantes, 1976) y sus diferentes versiones como DISTATIS (Abdi et al., 2012), JIVE (Lock et al., 2013), CANDECOMP/PARAFAC (Carroll & Chang, 1970; Harshman, 1970), Tucker (Kroonenberg & Leeuw, 1980; Tucker, 1966) , …

A lo largo de los últimos años también se ha visto sometida a cambios en su formulación teórica, modificando el problema de optimización asociado con la adición de restricciones sobre las normas de los vectores singulares. Estos darán

lugar a los llamados pseudo-vectores singulares, adaptados a situaciones del análisis de datos como puede ser el análisis de datos de altas dimensiones.

Ahora bien, muchas de las técnicas sparse presentan la desventaja de que para poder imponer nuevas restricciones sobre los vectores, debe perder alguna condición óptima que ya cumplían previamente. Habitualmente, es la ortogonalidad de dichos vectores la que se vuelve prescindible pero en el caso de la SVD, esta es una propiedad que no debe dejarse de lado. Por eso, recientemente, a principios del año 2019, Guillemot et al. (2019) presentaron *Constrained Singular Value Decomposition (CSVD),* una descomposición en valores singulares que integra tanto la condición sparsity como la ortogonalidad de los vectores singulares a izquierda y derecha. El problema de optimización clásico que resuelve la SVD se sustituye aquí por un problema de optimización convexa penalizada.

En este tipo de modelos, el objetivo es buscar una serie de parámetros, habitualmente almacenados en un vector v, que minimizan una función objetivo convexa modificada sobre la que se añade un término adicional de penalización. Para ello combinará diferentes restricciones implementando cada una de ellas como una proyección en un conjunto convexo. De esta forma, la solución final deberá pertenecer a la intersección de varios conjuntos convexos, que también es convexa.

Se presenta en este capítulo la teoría subyacente a la *proyección euclídea de un vector* sobre distintos espacios convexos y su metodología. Para ello, se introducirán previamente aspectos del álgebra lineal necesarios para su desarrollo. El estudio de la convexidad de conjuntos y funciones es de especial relevancia en la resolución de problemas de optimización, por tratarse de una propiedad deseable, con importantes implicaciones tanto a nivel de búsqueda de soluciones óptimas como de planteamiento de algoritmos de resolución.

En este capítulo se expondrán algunas definiciones preliminares de los espacios convexos, así como propiedades deseables de estos, además de las bases conceptuales y algorítmicas de la proyección de un vector sobre un espacio convexo.

Definición 1. Dados $\boldsymbol{x_1, x_2} \in \mathbb{R}^J$ dos puntos, se llama segmento lineal cerrado $[\boldsymbol{x_1, x_2}]$ al conjunto:

$$[\boldsymbol{x_1, x_2}] \coloneqq \{\boldsymbol{x} \in \mathbb{R}^J \boldsymbol{x_1, x_2} \in (0,1)\}$$

y segmento lineal abierto $]\boldsymbol{x_1, x_2}[$ a:

$$]\boldsymbol{x_1, x_2}[\coloneqq \{\boldsymbol{x} \in \mathbb{R}^J \boldsymbol{x_1 x_2} \in (0,1)\}$$

Definición 2. En un espacio vectorial, un conjunto C es convexo si $\forall \boldsymbol{x, y} \in C$ y $\forall \lambda \in [0,1]$ el segmento lineal cerrado $[\boldsymbol{x, y}]$ que une los dos puntos está totalmente contenido en el conjunto C (Figura 42,Figura 43):

$$\boldsymbol{xy} \in C$$

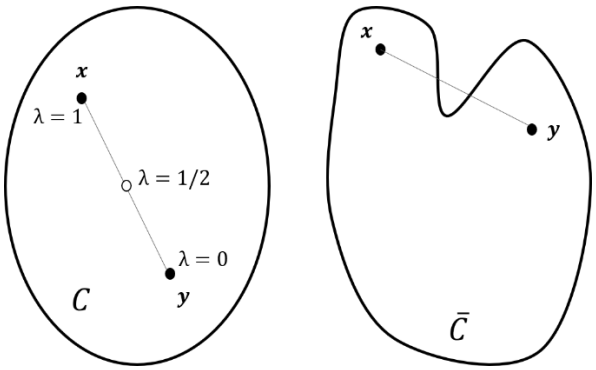


***Figura 42**. Ejemplo de conjunto convexo (izquierda) y no convexo (derecha)*

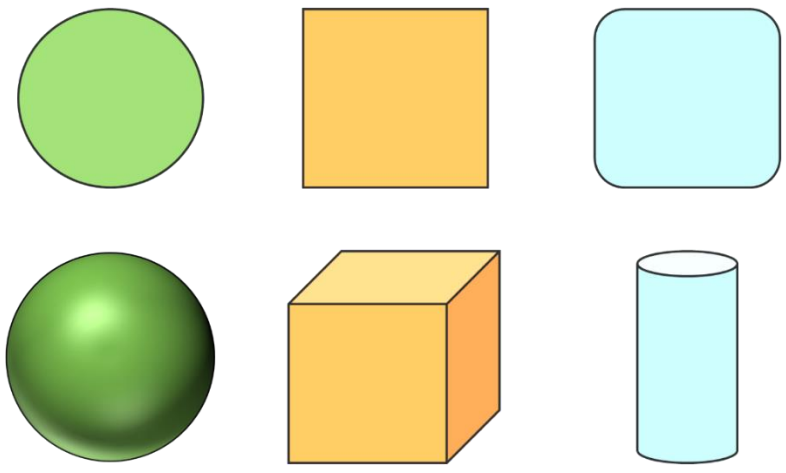


***Figura 43**. Ejemplos de conjuntos convexos*

Proposición 1. La intersección, finita, numerable o no numerable, de conjuntos convexos es un conjunto convexo.

Proposición 2. La combinación lineal de conjuntos convexos es un conjunto convexo.

Definición 3. Sea $C \subseteq \mathcal{H}$ un conjunto convexo y no vacío y sea f una función definida $f: C \to \mathbb{R}$, se dice que f es una función convexa en C si y sólo si $\forall \boldsymbol{x_1}, \boldsymbol{x_2} \in C$ y $\forall \in [0,1]$:

$$f(\lambda \boldsymbol{x_1} + (1-\lambda)\boldsymbol{x_2}) \leq \lambda f(\boldsymbol{x_1}) + (1-\lambda) f(\boldsymbol{x_2})$$

Se dirá que f es una función estrictamente convexa en C si

$$f(\lambda \boldsymbol{x_1} + (1-\lambda)\boldsymbol{x_2}) < \lambda f(\boldsymbol{x_1}) + (1-\lambda) f(\boldsymbol{x_2})$$

Proyección en un conjunto convexo. Propiedades.

Teorema 2. Si $\mathrm{C} \subset \mathcal{H}$ es un espacio finito cerrado no vacío y convexo incluido en el espacio de Hilbert, entonces $\forall \boldsymbol{x} \in \mathbb{R}^J$ existe una única $P_C(\boldsymbol{x}) \in C$ que minimice la distancia entre $\boldsymbol{x}$ e $\boldsymbol{y} \in C$:

$$\|\boldsymbol{x} - \mathrm{P_C}(\boldsymbol{x})\|^2 \leq \|\boldsymbol{x} - \boldsymbol{y}\|^2$$

De otro modo, existe un único vector $\hat{\boldsymbol{x}} = P_C(\boldsymbol{x})$ tal que $\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2 = \min_{\boldsymbol{y} \in C} \|\boldsymbol{x} - \boldsymbol{y}\|^2$ (véase la Figura 44).

Propiedades:

(1) La proyección en un conjunto convexo es idempotente:

$$P_C\big(P_C(\boldsymbol{x})\big) = P_C(\boldsymbol{x})$$

Demostración: Si $x \in C$, entonces es obvio que el mínimo de la distancia entre x y $P_C(x)$ es x:

$$P_C(\boldsymbol{x}) = \boldsymbol{x}$$

Si $P_C(\boldsymbol{x}) \in C$, entonces es obvio que $P_C(P_C(\boldsymbol{x}) \in C) = P_C(\boldsymbol{x})$

(2) $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^J$:

$$\|P_C(\boldsymbol{x}) - P_C(\boldsymbol{y})\|_2 \leq \|\boldsymbol{x} - \boldsymbol{y}\|_2$$

(3) Si el conjunto C es convexo, entonces la proyección de x en C, $P_C(x)$, es única. Esto no es así en caso de conjuntos no convexos.

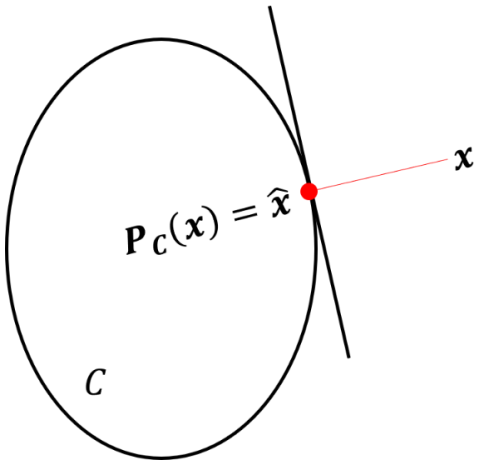


***Figura 44**. Proyección de un vector en dos dimensiones*

En las próximas líneas se presenta la noción básica del algoritmo POCS (*Projections onto convex sets*), utilizado para la proyección de un vector sobre un espacio convexo. Será de gran utilidad en la resolución de los problemas de optimización penalizada que se tratarán en este capítulo para la proyección de un vector sobre las bolas o regiones convexas de restricción Ridge, Lasso o Elastic net.

Algoritmo POCS

En el caso de trabajar con problemas de optimización con más de una restricción convexa, se recurrirá al algoritmo de las proyecciones alternadas o algoritmo POCS, un método iterativo que permite realizar proyecciones sucesivas de un vector $\boldsymbol{x}$ para determinar la solución $\hat{\boldsymbol{x}}$ que cumpla con todas las restricciones convexas establecidas. El algoritmo de las proyecciones alternadas se utiliza para calcular un punto en la intersección de una serie de conjuntos convexos, proyectando en cada uno de los conjuntos de manera secuencial (Boyd & Dattorro, 2003).

El algoritmo POCS se describe en la siguiente proposición:

Proposición 2. Sea i un entero estrictamente positivo y sea I el conjunto $I = \{1, \dots i\}$. Sea $\{C_i\}_{i \in I}$ una familia de subconjuntos convexos no vacíos y cerrados de un espacio de Hilbert $\mathcal{H}$, $\{P_i\}_{i \in I}$ las funciones proyección y $\boldsymbol{x_0} \in \mathcal{H}$. Supóngase que la intersección de los espacios $P_1, \dots, P_m$ es no vacía y sea:

$$\boldsymbol{x_{n+1}} = P_1 \cdots P_m \boldsymbol{x_n} \quad \forall n \in \mathbb{N}$$

entonces existe $(y_1, \dots, y_m) \in C_1 \, x \dots x \, C_m$ tal que

$$\boldsymbol{x_n} \to \boldsymbol{y_1} = P_1 \boldsymbol{y_2},$$

$$P_m \boldsymbol{x_n} \to \boldsymbol{y_m} = P_m \boldsymbol{y_1},$$

$$P_{m-1} P_m \boldsymbol{x_n} \to \boldsymbol{y_{m-1}} = P_{m-1} \boldsymbol{y_m}$$

$$,\dots,$$

$$P_3 \cdots P_m \boldsymbol{x_n} \to \boldsymbol{y_3} = P_3 \boldsymbol{y_4},$$

$$P_2 \cdots P_m \boldsymbol{x_n} \to \boldsymbol{y_2} = P_2 \boldsymbol{y_3}.$$

Para el caso de dos conjuntos convexos C y G tales que $C, G \subset \mathbb{R}^J$, la proposición previa da lugar a lo siguiente. Se denota por P_C y P_G a las proyeciones en C y G respectivamente. El algoritmo parte de un punto $\boldsymbol{x_0} \in C$, que se proyecta de manera alternada sobre C y G:

$$\boldsymbol{y}_k = P_G(\boldsymbol{x_k}), \qquad \boldsymbol{x_{k+1}} = P_C(\boldsymbol{y_k})$$

para $k = 0,1, \dots$ Así, se genera una secuencia de puntos $\boldsymbol{x_k} \in C$ e $\boldsymbol{y}_k \in G$ que, en caso de que $C \cap G \neq \emptyset$ convergen a un punto $\hat{\boldsymbol{x}} \in C \cap G$ (Figura 45)*;* hecho que fue demostrado por Cheney y Goldstein (1959). El algoritmo POCS puede no converger a la solución $\hat{\boldsymbol{x}} \in C \cap G$ en un número finito de iteraciones, pero se verifica que $\text{dist}(\boldsymbol{x_k}, G) \to 0$ y $\text{dist}(\boldsymbol{y_k}, C) \to 0$ (Boyd & Dattorro, 2003). Es trivial por tanto concluir que cuanto mayor sea el número de iteraciones más se aproximará el elemento la intersección de todos los espacios convexos. El algoritmo de proyección sobre más de dos espacios convexos es también conocido como algoritmo de proyección secuencial o cíclico.

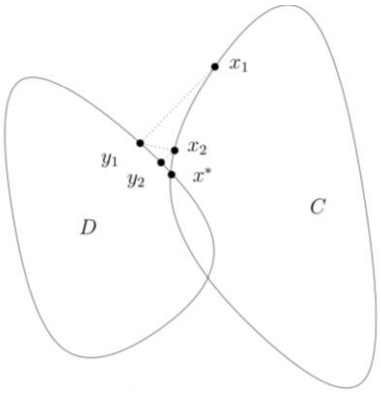


***Figura 45**. Esquema del algoritmo POCS para la proyección de un vector $\boldsymbol{x}$ en la intersección de dos espacios convexos. Fuente: (Boyd & Dattorro, 2003)*

3.1 Métodos de proyección de un vector sobre la bola ℓ_p.

A continuación, se presenta el algoritmo utilizado para resolver los problemas de optimización penalizados, basado en el algoritmo *Spectral-Projected-Gradient* (SPG) (Birgin, Martínez, & Raydan, 2000), utilizado como base en la solución de problemas ℓ_p-regularizados (Berg, Schmidt, Friedlander, & Murphy, 2008; Guillemot et al., 2019; Mairal, Bach, Ponce, & Sapiro, 2009; van den Berg & Friedlander, 2009). Cuando la regularización es incluida en el modelo como una restricción en lugar de como una penalización, el operador solución de la función a optimizar puede obtenerse mediante algoritmos pivote en tiempo lineal. Son este tipo de algoritmos los que se utilizarán en cada uno de los métodos que se muestran de ahora en adelante.

En las próximas secciones se hará referencia a la proyección de un vector sobre la bola de restricción Lasso (Berg et al., 2008; Duchi, Shalev-Shwartz, Singer, & Chandra, 2008), base del desarrollo de CSVD (Guillemot et al., 2019) y se presentará también la metodología propuesta por Mairal, Bach, Ponce y Sapiro (2010) para la proyección de un vector sobre la bola de restricción Elastic net (Figura 46). Esta formulación sentará las bases de la metodología que se propone en esta tesis doctoral, a partir de una formulación equivalente para la proyección de un vector en la bola Elastic net y su extensión para la formulación de C$_{enet}$SVD, principal contribución de este trabajo.

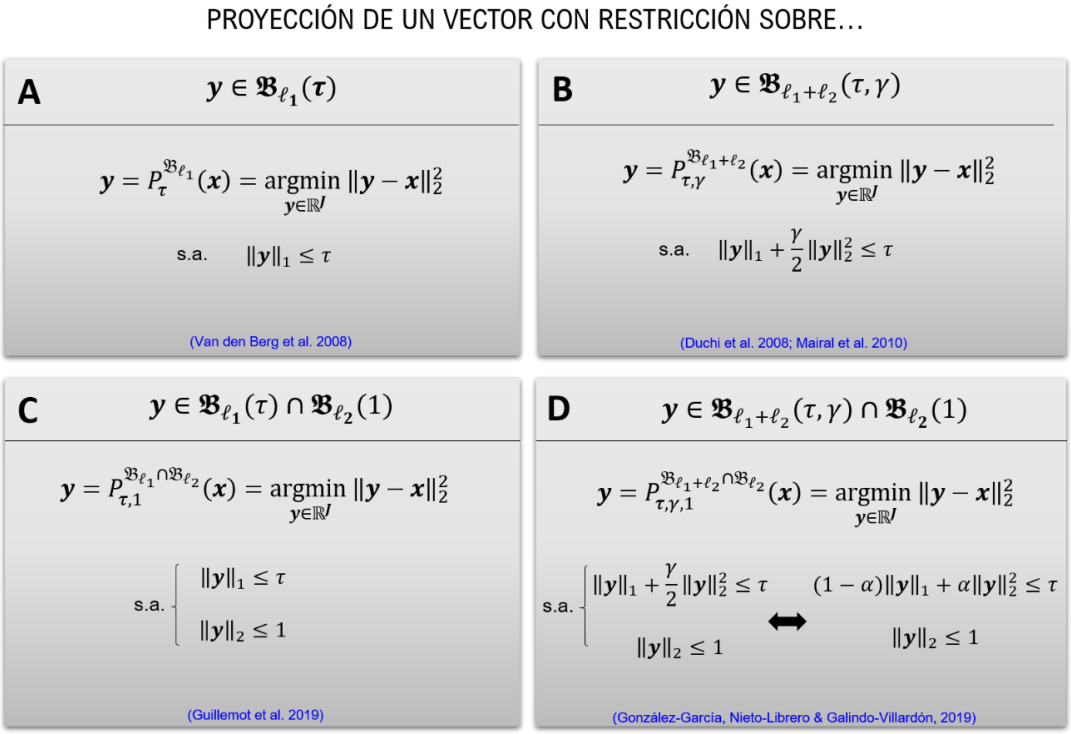


Figura 46. *Proyección de un vector* $\boldsymbol{x}$ *sobre el balón* $\mathfrak{B}_{\ell_1}(\tau)$ *(Lasso) (A), sobre* $\mathfrak{B}_{\ell_1+\ell_2}(\tau,\gamma)$ *(Elastic net) (B), sobre* $\mathfrak{B}_{\ell_1}(\tau) \cap \mathfrak{B}_{\ell_2}(1)$ *(Lasso normalizado) (C) y propuesta teórica de proyección sobre* $\mathfrak{B}_{\ell_1+\ell_2}(\tau) \cap \mathfrak{B}_{\ell_2}(1)$ *(Elastic net normalizado) (D)*

3.1.1 Proyección sobre el espacio $\mathfrak{B}_{\ell_1}$ de restricción Lasso

Berg et al. (2008) presentan un método eficiente de proyección de un vector sobre una restricción en la cota de su norma ℓ_1 que fue previamente desarrollado de manera independiente por Candes & Romberg (2005) y Daubechies, Fornasier y Loris (2008), y que los autores mejoran para su extensión al caso de proyección sobre la restricción *Group* Lasso. Este último no es de interés en este trabajo.

De manera general, el objetivo es encontrar un vector solución a un problema de optimización de mínimos cuadrados clásico, sujeto a una restricción sobre el límite de la norma ℓ_1 de sus coeficientes. De esta forma, el vector solución debe residir en el espacio $\mathfrak{B}_{\ell_1}$. El problema de optimización restringido se escribe:

$$\boldsymbol{y} = P_{\tau}^{\mathfrak{B}_{\ell_1}}(\boldsymbol{x}) = \left\{ \underset{\boldsymbol{y}\in\mathbb{R}^J}{\operatorname{argmin}} \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2 \,;\, s.a.: \|\boldsymbol{y}\|_1 \leq \tau \right\} \tag{3.1}$$

con $\tau > 0$ y $\boldsymbol{c} \in \mathbb{R}^J$ conocidos. La ecuación (3.1) plantea un problema de optimización *restringido (constrained)*, en el que se busca un vector solución $\boldsymbol{x}$**,** con una cota superior sobre su norma ℓ_1, que minimice la función objetivo. Matemáticamente, este problema se puede reformular de manera equivalente como un problema de optimización *penalizada*, en el que se busca una solución que minimice la función objetivo modificada, incluyendo la penalización en el propio problema:

$$\underset{\boldsymbol{y}\in\mathbb{R}^J}{\operatorname{argmin}} \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2 + \lambda\|\boldsymbol{y}\|_1 \tag{3.2}$$

con $\lambda > 0$ un valor real no negativo. Ambas formulaciones son equivalentes en el sentido de que $\forall\tau > 0$ (respectivamente λ) existe un $\lambda > 0$ (respectivamente τ) de manera que las soluciones de (3.1) y (3.2) son la misma. Se ignorará de aquí en adelante la solución trivial $\boldsymbol{y} = \boldsymbol{x}$.

Tibshirani (1996) demostró que la solución al problema de optimización penalizada (3.2) viene dada por la aplicación del operador *soft-thresholding* a cada uno de los coeficientes *j* del vector $\boldsymbol{x}$:

$$S_\lambda(x_j) = S(x_j,\lambda) = \operatorname{sign}(x_j)\left(|x_j| - \lambda\right)_+ = \begin{cases} x_j + \lambda & si\ x_j < \lambda \\ 0 & si\ x_j \in [-\lambda,\lambda] \\ x_j - \lambda & si\ x_j > \lambda \end{cases} \tag{3.3}$$

donde la función $(z)_+$ es la función que asigna el valor máximo entre z y 0, para cualquier escalar z. Si se halla el valor λ que haga los casos (3.1) y (3.2) equivalentes, entonces ambos problemas de optimización podrían resolverse haciendo uso del operador *soft-thresholding* componente a componente del vector. Si $\|\boldsymbol{x}\|_1 > \tau$, esto se traduce en encontrar λ tal que:

$$\|S_\lambda(\boldsymbol{x})\|_1 = \tau \tag{3.4}$$

Por ello, si se define la función $\phi(\lambda) = \|S_\lambda(\boldsymbol{x})\|_1$, encontrar la solución $\boldsymbol{y}$ al problema (3.1) supone desarrollar un método con el que hallar λ (que existe debido a las propiedades de ϕ y de las que se hablará más adelante), tal que:

$$\phi(\lambda) = \|S_\lambda(\boldsymbol{x})\|_1 = \tau \tag{3.5}$$

Para facilitar la comprensión del método propuesto en (Berg et al., 2008), tal y como proponen los autores, se considera a partir de ahora el vector $\tilde{\boldsymbol{x}} = (\tilde{x}_1, \dots, \tilde{x}_J)$ formado por los valores absolutos del vector $\boldsymbol{x}$ ordenados de manera decreciente, elemento a elemento, con $\tilde{x}_{J+1} = 0$.

La función $\phi(\lambda)$ se define como una función:

i) Continua.

ii) Monótona decreciente en λ, desde $\phi(\tilde{x}_{J+1}) = \phi(0) = \|\boldsymbol{x}\|_1$ a $\phi(\tilde{x}_1) = 0$ (Figura 47). Esto se debe a la definición del vector $\tilde{\boldsymbol{x}}$ ordenado de manera decreciente y el operador *soft-thresholding*:

$$\phi(\tilde{x}_{J+1}) = \left\| S_{\tilde{x}_{J+1}}(\boldsymbol{x}) \right\|_1 = \left\| \text{sign}(\boldsymbol{x})(|\boldsymbol{x}| - \tilde{x}_{J+1})_+ \right\|_1$$
$$= \|\text{sign}(\boldsymbol{x})(|\boldsymbol{x}| - 0)_+\|_1 = \|\text{sign}(\boldsymbol{x})|\boldsymbol{x}|\|_1 = \|\boldsymbol{x}\|_1$$

$$\phi(\tilde{x}_1) = \left\| S_{\tilde{x}_1}(\boldsymbol{x}) \right\|_1 = \|\text{sign}(\boldsymbol{x})(|\boldsymbol{x}| - \tilde{x}_1)_+\|_1 = \|\text{signo}(\boldsymbol{x}) \cdot 0\|_1 = 0$$

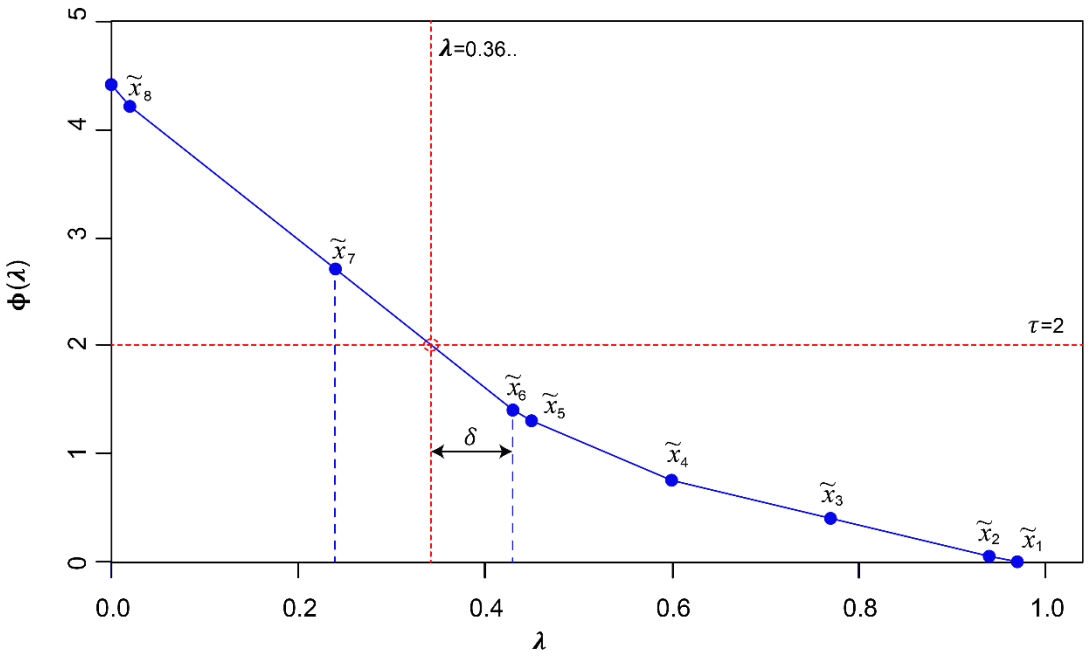


***Figura 47**. Valores de $(\boldsymbol{\lambda}, \boldsymbol{\phi}(\boldsymbol{\lambda}))$ para un vector aleatorio $\boldsymbol{x}$ de longitud 8. Fuente: (Berg et al., 2008)*

Dado que $\phi(\lambda)$ se define como una función decreciente, existe $k \in \mathbb{Z}$ tal que:

$$\phi(\tilde{x}_k) \leq \tau < \phi(\tilde{x}_{k+1}), \quad con\ \tilde{x}_k > \tilde{x}_{k+1} \tag{3.6}$$

(véase Figura 47), donde $\phi(\tilde{x}_j), j = 1, \dots J$, por definición de $\phi(\lambda) = \|S_\lambda(\boldsymbol{x})\|_1$ es:

$$\begin{aligned} \phi(\tilde{x}_j) = \left\|S_{\tilde{x}_j}(\boldsymbol{x})\right\|_1 &= \left\|\text{sign}(\boldsymbol{x})\left(|\boldsymbol{x}| - \tilde{x}_J\right)_+\right\|_1 = \left\|\left(|\boldsymbol{x}| - \tilde{x}_J\right)_+\right\|_1 \\ &= \sum_{i=1}^{J} \max\{0, |x_i| - \tilde{x}_j\} = \sum_{i=1}^{J} \max\{0, \tilde{x}_i - \tilde{x}_j\} \\ &= \sum_{i=1}^{j}\left(\tilde{x}_i - \tilde{x}_j\right) + \sum_{i=j}^{J} 0 = \sum_{i=1}^{j}\left(\tilde{x}_i - \tilde{x}_j\right) = \sum_{i=1}^{j}\tilde{x}_i - \sum_{i=1}^{j}\tilde{x}_j \\ &= \sum_{i=1}^{j}\tilde{x}_i - j \cdot \tilde{x}_j \end{aligned} \tag{3.7}$$

Debe tenerse en cuenta que la función $\text{sign}(\cdot)$ queda anulada por la definición de la norma $\|\cdot\|_1$ y los signos de $\boldsymbol{x}$ sólo serán relevantes en la solución final. Además, por definición de $\tilde{\boldsymbol{x}}$, $\tilde{x}_j \geq 0$.

El objetivo principal es hallar λ tal que $\phi(\lambda) = \|S_\lambda(\boldsymbol{x})\|_1 = \tau$, pero por las propiedades de ϕ se traduce en encontrar k tal que $\phi(\tilde{x}_k) \leq \tau < \phi(\tilde{x}_{k+1})$. Equivalentemente, hay que encontrar δ tal que $\phi(\tilde{x}_k - \delta) = \tau$, con $0 \leq \delta \leq \tilde{x}_k - \tilde{x}_{k+1}$. Para cualquier índice j y $\delta \in [0, \tilde{x}_j - \tilde{x}_{j+1}]$:

$$\begin{aligned} \phi\left(\tilde{x}_j - \delta\right) &= \sum_{i=1}^{j}\left(\tilde{x}_i - (\tilde{x}_j - \delta)\right) = \sum_{i=1}^{j}\tilde{x}_i - \sum_{i=1}^{j}\left(\tilde{x}_j - \delta\right) \\ &= \sum_{i=1}^{j}\tilde{x}_i - j \cdot \tilde{x}_j + j \cdot \delta = \left(\sum_{i=1}^{j}\tilde{x}_i - j \cdot \tilde{x}_j\right) + j \cdot \delta \\ &= \phi\left(\tilde{x}_j\right) + j \cdot \delta \end{aligned} \tag{3.8}$$

Suponiendo k conocida, δ tal que $\phi(\tilde{x}_k - \delta) = \tau$ viene dado por:

$$\begin{cases} \phi(\tilde{x}_k - \delta) = \tau \\ \phi(\tilde{x}_k - \delta) = \phi(\tilde{x}_k) + k \cdot \delta \end{cases} \rightarrow \ \phi(\tilde{x}_k) + k \cdot \delta = \tau \ \rightarrow \ \delta = \frac{\tau - \phi(\tilde{x}_k)}{k} \tag{3.9}$$

Por último, como el fin era encontrar λ tal que $\phi(\lambda) = \tau$,

$$\begin{cases} \phi(\tilde{x}_k - \delta) = \tau \\ \phi(\lambda) = \tau \end{cases} \quad \rightarrow \quad \lambda = \tilde{x}_k - \delta \tag{3.10}$$

Una vez definido λ, el vector solución $\boldsymbol{x}$ viene dado por el operador *soft-thresholding*:

$$\boldsymbol{y} := P_{\tau}^{\mathfrak{B}_{\ell_1}}(\boldsymbol{x}) = S_{\lambda}(\boldsymbol{x}) = \text{sign}(\boldsymbol{x}) \cdot (|\boldsymbol{x}| - \lambda)_+ \tag{3.11}$$

La Figura 48 muestra el diagrama del proceso global esquematizado, aunque este puede resumirse en las siguientes cuatro etapas:

1- Calcular el vector $\tilde{\boldsymbol{x}}$ de los coeficientes de $\boldsymbol{x}$ en valor absoluto (en el algoritmo que se verá a continuación, por propia construcción, no será necesario que estén ordenados de mayor a menor).
2- Encontrar k que cumpla la ecuación (3.6)
3- Calcular δ y λ según (3.9) y (3.10) respectivamente.
4- Una vez obtenida la penalización λ, calcular el vector proyección utilizando el operador *soft-thresholding.*

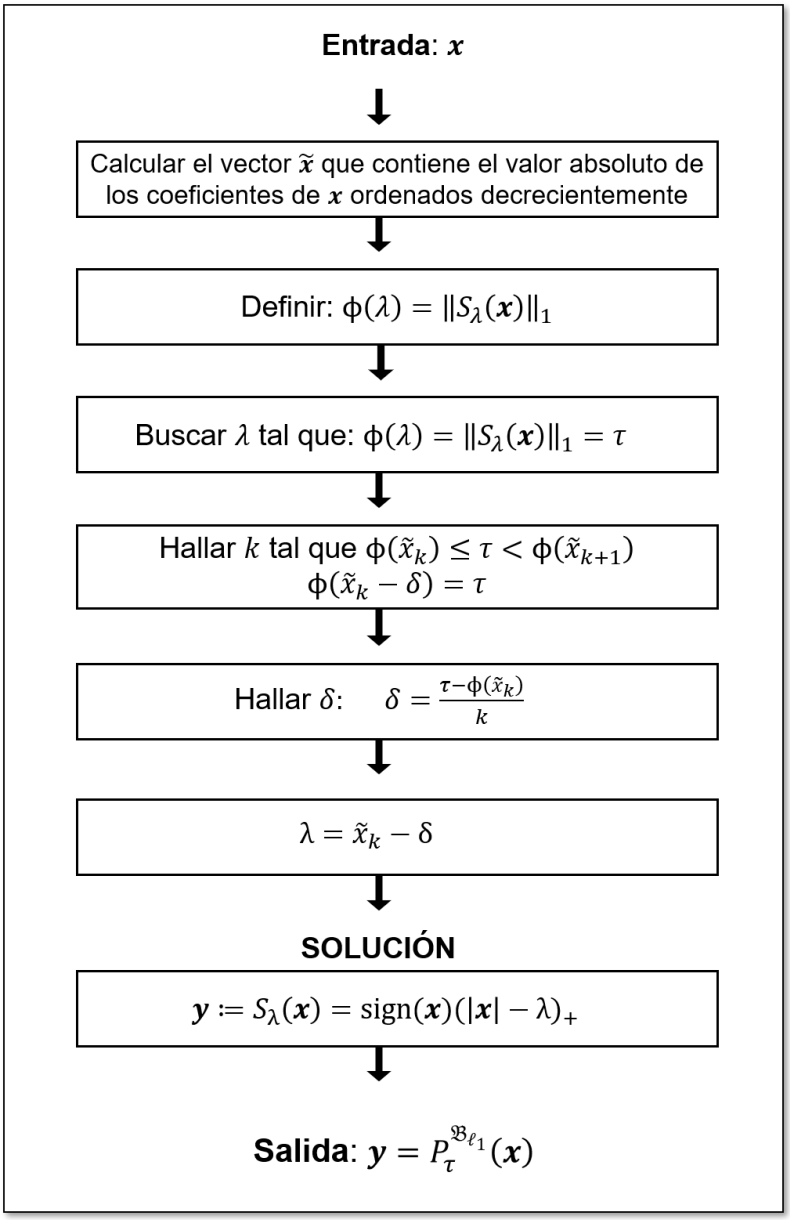


***Figura 48**. Esquema del método propuesto en (Berg et al., 2008) para la proyección sobre la norma $\boldsymbol{\ell_1}$*

Algoritmo de proyección sobre $\mathfrak{B}_{\ell_1}$ en tiempo lineal

Con el fin de desarrollar un algoritmo con un coste computacional lineal, los autores proponen el método de la Tabla 17 para calcular el valor k. Una vez obtenido este, calcular el vector solución a través de los pasos 3 y 4 de la sección anterior es trivial (Berg et al., 2008). El proceso sigue la idea de los algoritmos *divide y vencerás*.

***Tabla 17**. Algoritmo de proyección de un vector sobre la restricción Lasso (Berg et al., 2008)*

Algoritmo Lasso: Proyección sobre el balón $\mathfrak{B}_{\ell_1}$ de radio τ	
Entrada:	$\boldsymbol{x} \in \mathbb{R}^J, \tau \in \mathbb{R}$ con $\tau \in [1, \sqrt{J}]$
Salida	$\boldsymbol{y} = P_{\tau}^{\mathfrak{B}_{\ell_1}}(\boldsymbol{x})$
Inicialización:	$s = 0, p = abs(\boldsymbol{x})$

1: **Si** $\|\boldsymbol{x}\|_1 \leq \tau$ **entonces devuelve** $\boldsymbol{y} = \boldsymbol{x}$ #Ya cumple la restricción

2: **Si no**

3: **Mientras** que $U \neq \emptyset$ hacer:

4: $k = \text{longitud}(p)/2$ #Otra opción: selección aleatoria de k

5: $\tilde{x}_k = \text{mediana}(p)$ #Otra opción: $\tilde{x}_k = p[k]$

6: Particionar p en dos conjuntos disjuntos:

$$H = \{j \in [1, \dots, J] \,/\, |x_j| < |x_k|\}$$

$$L = \{j \in [1, \dots, J] \,/\, |x_j| > |x_k|\}$$

7: $s_{low} = \sum_{i \in L} |p_i| + \tilde{x}_k$

8: $\phi(\tilde{x}_k) = \text{s} + s_{low} - k\tilde{x}_k$

9: **Si** $\phi(\tilde{x}_k) < \tau$ **entonces**

10: Actualización de p: $p = \{p_i\}_{i \in L}$

11: **Si no**

12: $\tilde{x}_{k+1} = \max(\{p_i\}_{i \in H})$

13: $\phi(\tilde{x}_{k+1}) = s + s_{low} - k\tilde{x}_{k+1}$

14: **Si** $\phi(\tilde{x}_{k+1}) < \tau$ **entonces salir**

15: Actualización de p: $p = \{p_i\}_{i \in H}$

16: Actualización de s: $s = s + s_{low}$

17: **Fin Si**

18: **Fin Mientras**

19: #Encontrar la penalización λ:

$$\lambda = \tilde{x}_k - \frac{\tau - \phi(\tilde{x}_k)}{k}$$

20: # Solución: $\boldsymbol{y} = P_{\tau}^{\mathfrak{B}_{\ell_1}}(\boldsymbol{x})$

$$\boldsymbol{y} \coloneqq S_{\lambda}(\boldsymbol{x}) = \text{sign}(\boldsymbol{x})(|\boldsymbol{x}| - \lambda, 0)_+$$

21: **Fin Si**

3.1.2 Proyección sobre el espacio $\mathfrak{B}_{\ell_1+\ell_2}$ de restricción Elastic net

La proyección de un vector sobre el espacio $\mathfrak{B}_{\ell_1+\ell_2}$ fue propuesta en el año 2010 (Mairal et al., 2010). Mairal, et al. (2010) proponen en su trabajo la resolución del problema de optimización restringido para la proyección euclídea de un vector sobre la región $\mathfrak{B}_{\ell_1+\ell_2}$.

$$\boldsymbol{y} := P_{\tau}^{\mathfrak{B}_{\ell_1+\ell_2}}(\boldsymbol{x}) = \left\{ \underset{\boldsymbol{y}\in\mathbb{R}^J}{\operatorname{argmin}} \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2 \,;\, s.a\colon \|\boldsymbol{y}\|_1 + \frac{\gamma}{2}\|\boldsymbol{y}\|_2^2 \le \tau \right\} \tag{3.12}$$

O, equivalentemente, de forma penalizada:

$$\underset{\boldsymbol{x}\in\mathbb{R}^J}{\operatorname{argmin}} \ \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2 + \lambda(\|\boldsymbol{y}\|_1 + \frac{\gamma}{2}\|\boldsymbol{y}\|_2^2) \tag{3.13}$$

La búsqueda de la solución a este problema se lleva a cabo mediante la función Lagrangiana del problema (3.13).

$$\mathrm{L}(y,\lambda) = \underset{\boldsymbol{y}\in\mathbb{R}^J}{\operatorname{argmin}} \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2 \ + \ \lambda\left(\|\boldsymbol{y}\|_1 + \frac{\gamma}{2}\|\boldsymbol{y}\|_2^2 - \tau\right) \tag{3.14}$$

Minimizando el Lagrangiano con respecto a λ, Mairal, Bach, Ponce, & Sapiro (2010) siguen la teoría propuesta por Duchi, Shalev-Shwartz, Singer y Chandra (2008) y muestran que este problema admite una solución cerrada de la forma:

$$\boldsymbol{y} := S_{\lambda}^{*}(\boldsymbol{x}) = \frac{S_{\lambda}(x)}{1+\lambda\gamma} = \frac{\operatorname{sign}(x_j)\left(|x_j|-\lambda\right)_+}{1+\lambda\gamma} \quad \forall x_j \in \boldsymbol{x} \tag{3.15}$$

En su trabajo, Mairal, Bach, Ponce, & Sapiro (2010) afirman que trabajando sobre el problema dual a (3.13), el Teorema de Holguras Complementarias (*Complementary Slackness Theorem*) permite encontrar la solución óptima cuando se conoce la solución óptima de (3.14) (y viceversa). Este teorema afirma que si una variable dual es mayor a cero, entonces la

correspondiente restricción primal debe ser una igualdad y viceversa. Esto se traduce en que, obviando el caso $\lambda = 0$ por no ser una solución, la condición de holgura complementaria permite afirmar que si ambas formulaciones son equivalentes tiene que existir $\lambda \geq 0$ solución a:

$$\|S_\lambda^*(\tilde{\boldsymbol{x}})\|_1 + \frac{\gamma}{2}\|S_\lambda^*(\tilde{\boldsymbol{x}})\|_2^2 = \tau \tag{3.16}$$

donde $\tilde{\boldsymbol{x}} = (\tilde{x}_1, \dots, \tilde{x}_J)$ está formado por los valores absolutos de los elementos de $\boldsymbol{x}$ ordenados de manera decreciente. Para un seguimiento adecuado de la siguiente metodología, se añade un componente más al vector a proyectar $\tilde{\boldsymbol{x}}$ que se supondrá nulo, es decir: $\tilde{x}_{J+1} = 0$.

1) Para $\boldsymbol{\lambda} \in [\mathbf{0}, \tilde{\boldsymbol{x}}_1]$, existe la función $\boldsymbol{\theta(\lambda)} = \|\boldsymbol{S_\lambda^*(\tilde{x})}\|_1 + \frac{\gamma}{2}\|\boldsymbol{S_\lambda^*(\tilde{x})}\|_2^2$ continua y decreciente.

Se define la función:

$$\theta(\lambda) \coloneqq \|S_\lambda^*(\tilde{\boldsymbol{x}})\|_1 + \frac{\gamma}{2}\|S_\lambda^*(\tilde{\boldsymbol{x}})\|_2^2 \tag{3.17}$$

Donde

$$\|S_\lambda^*(\tilde{\boldsymbol{x}})\|_1 = \sum_{i=1}^{j}\left(\frac{\tilde{x}_i - \lambda}{1+\gamma\lambda}\right) = \frac{1}{1+\gamma\lambda}\sum_{i=1}^{j}(\tilde{x}_i - \lambda) = \frac{1}{1+\gamma\lambda}\left[\sum_{i=1}^{j}\tilde{x}_i - \sum_{i=1}^{j}\lambda\right] \tag{3.18}$$

$$\begin{aligned}\|S_\lambda^*(\tilde{\boldsymbol{x}})\|_2^2 &= \sum_{i=1}^{j}\left(\frac{\tilde{x}_i - \lambda}{1+\gamma\lambda}\right)^2 = \frac{1}{(1+\gamma\lambda)^2}\sum_{i=1}^{j}(\tilde{x}_i - \lambda)^2 \\ &= \frac{1}{(1+\gamma\lambda)^2}\sum_{i=1}^{j}(\tilde{x}_i^2 - 2\lambda\tilde{x}_i + \lambda^2) \\ &= \frac{1}{(1+\gamma\lambda)^2}\left[\sum_{i=1}^{j}\tilde{x}_i^2 - 2\lambda\sum_{i=1}^{j}\tilde{x}_i + \sum_{i=1}^{j}\lambda^2\right]\end{aligned} \tag{3.19}$$

Con todo ello:

$$\theta(\lambda) \coloneqq \|S_\lambda^*(\tilde{\boldsymbol{x}})\|_1 + \frac{\gamma}{2}\|S_\lambda^*(\tilde{\boldsymbol{x}})\|_2^2$$

$$= \frac{1}{1+\gamma\lambda}\left[\sum_{i=1}^{j}\tilde{x}_i - \sum_{i=1}^{j}\lambda\right] + \frac{\gamma}{2}\left[\frac{1}{(1+\gamma\lambda)^2}\left[\sum_{i=1}^{j}\tilde{x}_i^2 - 2\lambda\sum_{i=1}^{j}\tilde{x}_i + \sum_{i=1}^{j}\lambda^2\right]\right]$$

$$= \frac{\gamma}{2}\cdot\frac{1}{(1+\gamma\lambda)^2}\sum_{i=1}^{j}\tilde{x}_i^2 + \frac{1}{1+\gamma\lambda}\sum_{i=1}^{j}\tilde{x}_i + \frac{\gamma}{2}\cdot\frac{1}{(1+\gamma\lambda)^2}\cdot(-2\lambda)\sum_{i=1}^{j}\tilde{x}_i$$

$$-\frac{1}{1+\gamma\lambda}\sum_{i=1}^{j}\lambda + \frac{\gamma}{2}\cdot\frac{1}{(1+\gamma\lambda)^2}\sum_{i=1}^{j}\lambda^2$$

$$= \frac{\gamma}{2}\cdot\frac{1}{(1+\gamma\lambda)^2}\sum_{i=1}^{j}\tilde{x}_i^2 + \left(\frac{1}{1+\gamma\lambda} - \frac{\gamma\lambda}{(1+\gamma\lambda)^2}\right)\sum_{i=1}^{j}\tilde{x}_i + \frac{1}{(1+\gamma\lambda)^2}$$

$$\cdot\left[\frac{\gamma}{2}\sum_{i=1}^{j}\lambda^2 - (1+\gamma\lambda)\sum_{i=1}^{j}\lambda\right]$$

$$= \frac{\gamma}{2}\cdot\frac{1}{(1+\gamma\lambda)^2}\sum_{i=1}^{j}\tilde{x}_i^2 + \left(\frac{(1+\gamma\lambda)-\gamma\lambda}{(1+\gamma\lambda)^2}\right)\sum_{i=1}^{j}\tilde{x}_i + \frac{1}{(1+\gamma\lambda)^2}$$

$$\cdot\left[\frac{\gamma}{2}\sum_{i=1}^{j}\lambda^2 - \sum_{i=1}^{j}\lambda - \gamma\sum_{i=1}^{j}\lambda^2\right]$$

$$= \frac{\gamma}{2}\cdot\frac{1}{(1+\gamma\lambda)^2}\sum_{i=1}^{j}\tilde{x}_i^2 + \frac{1}{(1+\gamma\lambda)^2}\sum_{i=1}^{j}\tilde{x}_i - \frac{1}{(1+\gamma\lambda)^2}\cdot\left[\frac{\gamma}{2}\sum_{i=1}^{j}\lambda^2 + \sum_{i=1}^{j}\lambda\right]$$

$$= \frac{1}{(1+\gamma\lambda)^2}\left[\frac{\gamma}{2}\sum_{i=1}^{j}\tilde{x}_i^2 + \sum_{i=1}^{j}\tilde{x}_i - \left(\frac{\gamma}{2}\sum_{i=1}^{j}\lambda^2 + \sum_{i=1}^{j}\lambda\right)\right]$$

$$= \frac{1}{(1+\gamma\lambda)^2}\left[\sum_{i=1}^{j}\left(\frac{\gamma}{2}\cdot\tilde{x}_i^2 + \tilde{x}_i - \left(\frac{\gamma}{2}\cdot\lambda^2 + \lambda\right)\right)\right]$$

$$= \frac{1}{(1+\gamma\lambda)^2}\left[\sum_{i=1}^{j}\left(\frac{\gamma}{2}\cdot\tilde{x}_i^2 + \tilde{x}_i - \lambda\left(1 + \frac{\gamma}{2}\cdot\lambda\right)\right)\right]$$

Por tanto,

$$\theta(\lambda) \coloneqq \|S_\lambda^*(\tilde{\boldsymbol{x}})\|_1 + \frac{\gamma}{2}\|S_\lambda^*(\tilde{\boldsymbol{x}})\|_2^2 \tag{3.20}$$
$$= \frac{1}{(1+\gamma\lambda)^2}\left[\sum_{i=1}^{j}\left(\frac{\gamma}{2}\cdot\tilde{x}_i^2 + \tilde{x}_i - \lambda\left(1+\frac{\gamma}{2}\cdot\lambda\right)\right)\right]$$

i) La función θ es una función continua por ser una combinación de funciones continuas.

ii) A partir de la forma cerrada de la solución, Mairal, Bach, Ponce, & Sapiro (2010) afirman que la función $\theta: \lambda \to \|\tilde{\boldsymbol{x}}^*(\lambda)\|_1 + \frac{\gamma}{2}\|\tilde{\boldsymbol{x}}^*(\lambda)\|_2^2$ es estrictamente decreciente en λ. Aunque los autores no demuestran esta afirmación, se desarrolla a continuación su por qué.

Demostración.

Dado que $\theta(\lambda)$ es una función continua, si $\theta'(\lambda) < 0$ quedará probado que la función θ es una función estrictamente decreciente.

$$\theta'(\lambda) = (\|S_\lambda^*(\tilde{\boldsymbol{x}})\|_1)' + \frac{\gamma}{2}(\|S_\lambda^*(\tilde{\boldsymbol{x}})\|_2^2)' \tag{3.21}$$

Derivando con respecto a λ cada una de las expresiones anteriores:

$$(\|S_\lambda^*(\tilde{\boldsymbol{x}})\|_1)' = \frac{0(1+\gamma\lambda) - \gamma\sum\tilde{x}_i}{(1+\gamma\lambda)^2} - \frac{j(1+\gamma\lambda) - j\gamma\lambda}{(1+\gamma\lambda)^2}$$
$$= \frac{-\gamma\sum\tilde{x}_i - j + j\gamma\lambda - j\gamma\lambda}{(1+\gamma\lambda)^2} = \frac{-\gamma\sum\tilde{x}_i - j}{(1+\gamma\lambda)^2}$$

$$
\begin{aligned}
&(\|S_\lambda^*(\tilde{\boldsymbol{x}})\|_2^2)' \\
&= \frac{-2\gamma(1+\gamma\lambda)\sum \tilde{x}_i{}^2}{(1+\gamma\lambda)^4} - \frac{2(1+\gamma\lambda)^2\sum \tilde{x}_i - 4\gamma\lambda(1+\gamma\lambda)\sum \tilde{x}_i}{(1+\gamma\lambda)^4} \\
&+ \frac{2j\lambda(1+\gamma\lambda)^2 - 2j\gamma\lambda^2(1+\gamma\lambda)}{(1+\gamma\lambda)^4} \\
&= \frac{-2\gamma\sum \tilde{x}_i{}^2 - 2(1+\gamma\lambda)\sum \tilde{x}_i - 4\gamma\lambda\sum \tilde{x}_i + 2j\lambda(1+\gamma\lambda) - 2j\gamma\lambda^2}{(1+\gamma\lambda)^3} \\
&= \frac{-2\gamma\sum \tilde{x}_i{}^2 - 2\sum \tilde{x}_i - 6\gamma\lambda\sum \tilde{x}_i + 2j\lambda}{(1+\gamma\lambda)^3}
\end{aligned}
$$

$$
\begin{aligned}
\theta'(\lambda) &= \left(\frac{-\gamma\sum \tilde{x}_i - j}{(1+\gamma\lambda)^2}\right) + \frac{\gamma}{2}\cdot\left(\frac{-2\gamma\sum \tilde{x}_i{}^2 - 2\sum \tilde{x}_i - 6\gamma\lambda\sum \tilde{x}_i + 2j\lambda}{(1+\gamma\lambda)^3}\right) \\
&= \left(\frac{-\gamma\sum \tilde{x}_i - j}{(1+\gamma\lambda)^2}\right) + \gamma\cdot\left(\frac{-\gamma\sum \tilde{x}_i{}^2 - \sum \tilde{x}_i - 3\gamma\lambda\sum \tilde{x}_i + j\lambda}{(1+\gamma\lambda)^3}\right) \\
&= \frac{-\gamma^2\sum \tilde{x}_i{}^2}{(1+\gamma\lambda)^3} - \frac{\gamma\sum \tilde{x}_i}{(1+\gamma\lambda)^2} - \frac{\gamma\sum \tilde{x}_i}{(1+\gamma\lambda)^3} - \frac{3\gamma^2\lambda\sum \tilde{x}_i}{(1+\gamma\lambda)^3} \\
&+ \frac{-j(1+\gamma\lambda) + j\lambda\gamma}{(1+\gamma\lambda)^3} \\
&= \frac{1}{(1+\gamma\lambda)^3}\left[-\gamma^2\sum \tilde{x}_i{}^2 - 2\gamma\sum \tilde{x}_i - 4\gamma^2\lambda\sum \tilde{x}_i - j\right]
\end{aligned}
\tag{3.22}
$$

Por tanto, $\theta'(\lambda) < 0$ por definición de $\lambda, \gamma, j, \tilde{x}_i > 0$. Como $\theta(\lambda)$ es continua y $\theta'(\lambda) < 0$, puede concluirse que $\theta(\lambda)$ es una función decreciente. ■

Además, $\theta(\lambda)$ decrece desde $\theta(0) = \|\tilde{\boldsymbol{x}}\|_1 + \frac{\gamma}{2}\|\tilde{\boldsymbol{x}}\|_2^2$, hasta $\theta(\tilde{x}_1) = 0$:

$$
\begin{aligned}
\theta(0) &= \|S_0^*(\tilde{\boldsymbol{x}})\|_1 + \frac{\gamma}{2}\|S_0^*(\tilde{\boldsymbol{x}})\|_2^2 \\
&= \left\|\frac{\operatorname{sign}(\tilde{x}_j)\left(|\tilde{x}_j| - 0\right)_+}{1+0}\right\|_1 + \frac{\gamma}{2}\left\|\frac{\operatorname{sign}(\tilde{x}_j)\left(|\tilde{x}_j| - 0\right)_+}{1+0}\right\|_2^2 = \\
&= \|\tilde{\boldsymbol{x}}\|_1 + \frac{\gamma}{2}\|\tilde{\boldsymbol{x}}\|_2^2
\end{aligned}
$$

$$\theta(\tilde{x}_1) = \left\|S^*_{\tilde{x}_1}(\tilde{\boldsymbol{x}})\right\|_1 + \frac{\gamma}{2}\left\|S^*_{\tilde{x}_1}(\tilde{\boldsymbol{x}})\right\|_2^2$$

$$= \left\|\frac{\text{sign}(\tilde{x}_j)\left(|\tilde{x}_j| - \tilde{x}_1\right)_+}{1+\gamma\tilde{x}_1}\right\|_1 + \frac{\gamma}{2}\left\|\frac{\text{sign}(\tilde{x}_j)\left(|\tilde{x}_j| - \tilde{x}_1\right)_+}{1+\gamma\tilde{x}_1}\right\|_2^2 =$$

$$= \left\|\frac{0}{1+\gamma\tilde{x}_1}\right\|_1 + \frac{\gamma}{2}\left\|\frac{0}{1+\gamma\tilde{x}_1}\right\|_2^2 = 0$$

Recuérdese que el objetivo era encontrar λ ($\tilde{x}_{k+1} \leq \lambda < \tilde{x}_k$, k elementos no nulos) tal que:

$$\|S^*_\lambda(\tilde{\boldsymbol{x}})\|_1 + \frac{\gamma}{2}\|S^*_\lambda(\tilde{\boldsymbol{x}})\|_2^2 = \tau \tag{3.23}$$

2) A partir de la definición de la función $\theta(\lambda)$, hallar λ es equivalente a resolver una ecuación de segundo grado. Nótese que:

Supóngase a continuación que k es conocido. La función $\theta(\lambda)$ es una función decreciente en λ, por lo que existe una única raíz solución a la ecuación de segundo grado, que proviene de desarrollar (3.23)

$$\frac{1}{(1+\gamma\lambda)^2}\left[\sum_{i=1}^{j}\left(\frac{\gamma}{2}\cdot\tilde{x}_i^2 + \tilde{x}_i - \lambda\left(1+\frac{\gamma}{2}\cdot\lambda\right)\right)\right] = \tau \tag{3.24}$$

$$\frac{\gamma}{2}\sum_{i=1}^{j}\tilde{x}_i^2 + \sum_{i=1}^{j}\tilde{x}_i - \lambda j - j\frac{\gamma}{2}\lambda^2 = \tau(1+\gamma\lambda)^2$$

$$\frac{\gamma}{2}\sum_{i=1}^{j}\tilde{x}_i^2 + \sum_{i=1}^{j}\tilde{x}_i - \lambda j - j\frac{\gamma}{2}\lambda^2 - \tau - 2\tau\gamma\lambda - \tau\gamma^2\lambda^2 = 0$$

Escribiéndolo como una ecuación de segundo grado de incógnita λ:

$$\left(\frac{-j\gamma}{2} - \tau\gamma^2\right)\lambda^2 + (-j - 2\tau\gamma)\lambda + \left(\frac{\gamma}{2}\sum_{i=1}^{j}\tilde{x}_i^2 + \sum_{i=1}^{j}\tilde{x}_i - \tau\right) = 0$$

$$\left(\frac{j\gamma}{2} + \tau\gamma^2\right)\lambda^2 + (j + 2\tau\gamma)\lambda + \left(\tau - \left(\frac{\gamma}{2}\sum_{i=1}^{j}\tilde{x}_i^2 + \sum_{i=1}^{j}\tilde{x}_i\right)\right) = 0$$

Llamando $s = \frac{\gamma}{2}\sum_{i=1}^{j}\tilde{x}_i^2 + \sum_{i=1}^{j}\tilde{x}_i$,

$$\lambda = \frac{-(j+2\tau\gamma) \pm \sqrt{(j+2\tau\gamma)^2 - 4\left(\frac{j\gamma}{2}+\tau\gamma^2\right)(\tau - s)}}{2\left(\frac{j\gamma}{2}+\tau\gamma^2\right)}$$

de donde la única solución válida es:

$$\lambda = \frac{-(j+2\tau\gamma) + \sqrt{(j+2\tau\gamma)^2 - 4\left(\frac{j\gamma}{2}+\tau\gamma^2\right)(\tau - s)}}{2\left(\frac{j\gamma}{2}+\tau\gamma^2\right)} \tag{3.25}$$

Si $\lambda = \frac{-b-\sqrt{b^2-4ac}}{2a}$, entonces $\lambda < 0$ y la solución no sería válida. Una vez encontrada la solución λ para el problema anterior, la construcción de la solución al problema de optimización (3.12) es inmediata:

$$\boldsymbol{y} \coloneqq S_\lambda^*(\boldsymbol{x}) = \frac{S_\lambda(\boldsymbol{x})}{1+\lambda\gamma} \tag{3.26}$$

Un resumen del método puede verse en la Figura 49.

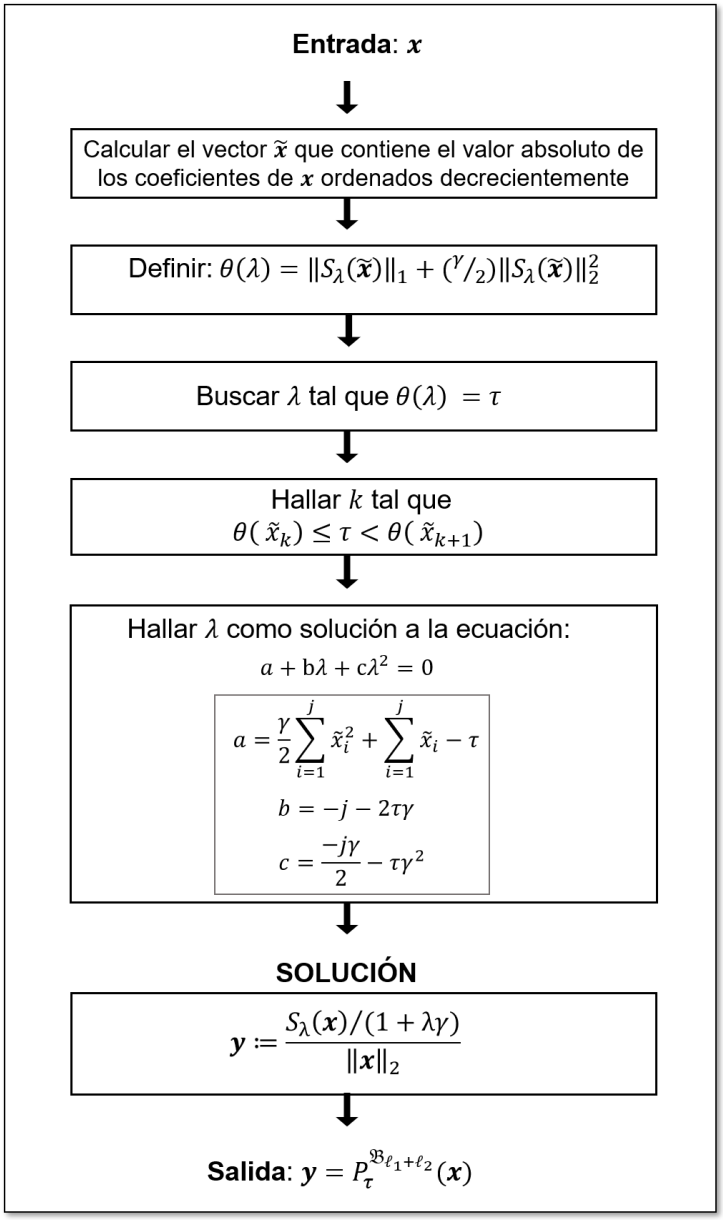


***Figura 49**. Esquema para la proyección sobre la norma $\ell_1 + \ell_2$ propuesto en (Mairal et al., 2010)*

Algoritmo de proyección sobre $\mathfrak{B}_{\ell_1+\ell_2}$ en tiempo lineal

El algoritmo propuesto por Mairal et al. (2010) para resolver la proyección sobre la restricción Elastic net sigue la teoría de Duchi et al. (2008) y Maculan & de Paula (1989) y el enfoque de los algoritmos *divide y vencerás* para encontrar el entero $k \in \{1, \dots, J-1\}$ tal que $\theta(\tilde{x}_k) \leq \tau < \theta(\tilde{x}_{k+1})$ (Tabla 18). El enfoque de resolución es similar al algoritmo de proyección de un vector sobre la bola $\mathfrak{B}_{\ell_1}$.

Duchi et al., 2008 proponen un método eficiente para la proyección de un vector con restricción sobre su norma ℓ_1, basándose en la idea de que tras

ordenar de manera decreciente el vector a proyectar, es posible calcular la proyección exacta de este en tiempo lineal. La idea (estudiada anteriormente por otros autores como Crammer & Singer (2002) y Hazan (2006)) se basa en una técnica de proyección euclídea sobre el Simplex probabilístico que posteriormente adaptan para la proyección euclídea de un vector sobre el balón $\mathfrak{B}_{\ell_1}$.

El desarrollo de la teoría está basada en un algoritmo que ordena decrecientemente las componentes del vector a proyectar (como se ha visto anteriormente en el desarrollo teórico de los métodos), con una complejidad computacional de $O(nlog(n))$. Sin embargo, este algoritmo es mejorado reemplazando el paso de la ordenación del vector por una modificación del algoritmo de búsqueda de la mediana aleatorizado (Cormen, Leiserson, Rivest, & Stein, 2009), dando lugar a un método de complejidad computacional esperada del orden de $O(n)$. Posteriormente, el algoritmo identifica la solución del problema sin ordenar el vector inicialmente siguiendo la idea de los métodos *divide y vencerás*. Mairal et al. (2010) extienden posteriormente un algoritmo similar para la proyección euclídea sobre la norma Elastic net, el que se presenta a continuación.

***Tabla 18**. Algoritmo de proyección de un vector sobre la restricción Elastic net (Mairal et al., 2010)*

Algoritmo E-NET: Proyección sobre la restricción Elastic net	
Entrada:	$\boldsymbol{x} \in \mathbb{R}^J, \gamma \in \mathbb{R}$ con $\gamma > 0, \tau \in \mathbb{R}$ con $\tau \in [1, \sqrt{J}]$
Salida	$\boldsymbol{y} = P_{\tau}^{\mathfrak{B}_{\ell_1+\ell_2}}(\boldsymbol{x})$
Inicialización:	$s = 0, \Delta s = 0, p = 0, \Delta p = 0, U = \{1, \dots, J\}$

1: **Si** $\|\boldsymbol{x}\|_1 + \frac{\gamma}{2}\|\boldsymbol{x}\|_2^2 \leq \tau$ **entonces devuelve** $\boldsymbol{y} = \boldsymbol{x}$ #Ya cumple la restricción

2: **Si no**

3: **Mientras** que $U \neq \emptyset$ hacer:

4: Seleccionar $k \in U$ aleatorio

5: #Particionar U en dos conjuntos disjuntos:

$$G = \{j \in U \,/\, |x_j| \geq |x_k|\}$$

$$L = \{j \in U \,/\, |x_j| < |x_k|\}$$

6: $\Delta p = |G|$

7: $$\Delta s = \sum_{j \in G} \left(|x_j| + \frac{\gamma}{2}|x_j|^2\right)$$

8: **Si** $s + \Delta s - (p + \Delta p)\left(1 + \frac{\gamma}{2}|x_k|\right)|x_k| < \tau(1 + \gamma|x_k|)^2$ **entonces**

9: Actualización de s y p: $s = s + \Delta s \;;\; p = p + \Delta p$

Actualización de U: $U = L$

10: **Si no**

11: Actualización de U: $U = G - \{k\}$

Fin Si

12: **Fin Mientras**

13: #Encontrar la penalización λ resolviendo la ecuación de 2º grado

$$\left(\gamma^2\tau + \frac{\gamma}{2}p\right)\lambda^2 + (2\gamma\tau + p)\lambda + (\tau - s) = 0$$

14: # Solución: $\boldsymbol{y} = P_{\tau}^{\mathfrak{B}_{\ell_1+\ell_2}}(\boldsymbol{x})$

$$\boldsymbol{y} \coloneqq \frac{S_{\lambda}(\boldsymbol{x})}{1 + \lambda\gamma}$$

15: **Fin Si**

3.1.3 Proyección de un vector sobre la intersección de regiones convexas: el espacio $\mathfrak{B}_{\ell_1} \cap \mathfrak{B}_{\ell_2}$

Se define la proyección de un vector $\boldsymbol{x}$ en la intersección de la región $\mathfrak{B}_{\ell_1}$de radio τ y el balón $\mathfrak{B}_{\ell_2}$ de radio 1 como la solución al siguiente problema de optimización con restricción (Guillemot et al., 2019):

$$\boldsymbol{y} := P_{\tau,1}^{\mathfrak{B}_{\ell_1} \cap \mathfrak{B}_{\ell_2}}(\boldsymbol{x}) = \left\{ \underset{\boldsymbol{y} \in \mathbb{R}^J}{\operatorname{argmin}} \|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \, ; s.a: \|\boldsymbol{y}\|_1 \leq \tau, \|\boldsymbol{y}\|_2 \leq 1 \right\} \tag{3.27}$$

Sea $\tilde{\boldsymbol{x}} = \left(\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_J\right)$ el vector compuesto por los valores absolutos de los coeficientes de $\boldsymbol{x}$ ordenados de manera decreciente ($\tilde{x}_1 \geq \tilde{x}_2 \geq \cdots \geq \tilde{x}_J$). Con el objetivo de resolver el problema de optimización (3.27). De manera similar a las proyecciones sobre las regiones $\mathfrak{B}_{\ell_1}$ y $\mathfrak{B}_{\ell_1+\ell_2}$ se define la función:

$$\psi(\lambda) = \frac{\|S_\lambda(\tilde{\boldsymbol{x}})\|_1}{\|S_\lambda(\tilde{\boldsymbol{x}})\|_2} \tag{3.28}$$

donde:

$$\|S_\lambda(\tilde{\boldsymbol{x}})\|_1 = \sum_{i=1}^{j} \tilde{x}_i - j \cdot \lambda \tag{3.29}$$

$$\|S_\lambda(\tilde{\boldsymbol{x}})\|_2^2 = \sum_{i=1}^{j} (\tilde{x}_i - \lambda)^2 = \sum_{i=1}^{j} \tilde{x}_i^2 - 2\lambda \sum_{i=1}^{j} \tilde{x}_i + j \cdot \lambda^2 \tag{3.30}$$

para algún j tal que $\tilde{x}_{j+1} \leq \lambda$. Al igual que en los casos anteriores, el objetivo es encontrar λ tal que:

$$\psi(\lambda) = \frac{\|S_\lambda(\tilde{\boldsymbol{x}})\|_1}{\|S_\lambda(\tilde{\boldsymbol{x}})\|_2} = \tau \tag{3.31}$$

La función ψ cumple las siguientes propiedades:

1- Es continua (por ser composición de funciones continuas)

2- Decreciente $(\psi'(\lambda) \leq 0)$ desde $\psi(0) \leq \sqrt{J}$ hasta $\psi(v) \leq \sqrt{J_{MAX}}$, con $v \in [\tilde{x}_2, \tilde{x}_1]$ y J_{MAX} la frecuencia absoluta de elementos de $\boldsymbol{x}$ iguales a $\tilde{x}_1$.

Por un lado:

$$\psi(v) = \frac{\|S_v(\tilde{\boldsymbol{x}})\|_1}{\|S_v(\tilde{\boldsymbol{x}})\|_2} = \frac{\|\text{sign}(\tilde{\boldsymbol{x}})(|\tilde{\boldsymbol{x}}| - v)_+\|_1}{\|\text{sign}(\tilde{\boldsymbol{x}})(|\tilde{\boldsymbol{x}}| - v)_+\|_2} = \frac{(\tilde{x}_1 - v) \cdot J_{MAX}}{(\tilde{x}_1 - v) \cdot \sqrt{J_{MAX}}}$$

$$= \frac{J_{MAX}}{\sqrt{J_{MAX}}} = \frac{J_{MAX}}{\sqrt{J_{MAX}}} \cdot \frac{\sqrt{J_{MAX}}}{\sqrt{J_{MAX}}} = \sqrt{J_{MAX}}$$

Por otro lado,

$$\psi(0) = \frac{\|S_0(\tilde{\boldsymbol{x}})\|_1}{\|S_0(\tilde{\boldsymbol{x}})\|_2} = \frac{\|\text{sign}(\tilde{\boldsymbol{x}})(|\tilde{\boldsymbol{x}}| - 0)_+\|_1}{\|\text{sign}(\tilde{\boldsymbol{x}})(|\tilde{\boldsymbol{x}}| - 0)_+\|_2} = \frac{\|\tilde{\boldsymbol{x}}\|_1}{\|\tilde{\boldsymbol{x}}\|_2} = \frac{\|\boldsymbol{x}\|_1}{\|\boldsymbol{x}\|_2}$$

De la relación entre normas, se verifica el siguiente el resultado.

Lema 1. Sea $\boldsymbol{x} \in \mathbb{R}^J$, entonces se cumple que:

$$\|\boldsymbol{x}\|_2 \leq \|\boldsymbol{x}\|_1 \leq \sqrt{J}\|\boldsymbol{x}\|_2$$

La demostración del Lema se puede encontrar en los anexos del trabajo (Guillemot et al., 2019).

Por la aplicación del lema 1 sobre el resultado anterior se cumple que:

$$\psi(0) = \frac{\|\boldsymbol{x}\|_1}{\|\boldsymbol{x}\|_2} \leq \frac{\sqrt{J}\|\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2} = \sqrt{J}$$

Por último, recuérdese que una función $f(x)$ continua es decreciente si $f'(x) < 0 \ \forall x \in \text{Dom}(f(x))$. Por tanto, para demostrar que la función $\psi(\lambda)$ es decreciente basta con demostrar que $\psi'(\lambda) < 0$. Por definición,

$$\psi(\lambda) = \frac{\|S_\lambda(\tilde{\boldsymbol{x}})\|_1}{\|S_\lambda(\tilde{\boldsymbol{x}})\|_2}$$

Luego:

$$\psi'(\lambda) = \frac{(\|S_\lambda(\tilde{\boldsymbol{x}})\|_1)'\|S_\lambda(\tilde{\boldsymbol{x}})\|_2 - \|S_\lambda(\tilde{\boldsymbol{x}})\|_1(\|S_\lambda(\tilde{\boldsymbol{x}})\|_2)'}{\|S_\lambda(\tilde{\boldsymbol{x}})\|_2{}^2}$$

Como $\|S_\lambda(\tilde{\boldsymbol{x}})\|_1 = \sum_{i=1}^{j} \tilde{x}_i - j \cdot \lambda$,

$$\|S_\lambda(\tilde{\boldsymbol{x}})\|_1' = -j$$

Además, como:

$$\|S_\lambda(\tilde{\boldsymbol{x}})\|_2^2 = \sum_{i=1}^{j}(\tilde{\boldsymbol{x}}_i - \lambda)^2 = \sum_{i=1}^{j}\tilde{\boldsymbol{x}}_i^2 - 2\lambda\sum_{i=1}^{j}\tilde{\boldsymbol{x}}_i + j\cdot\lambda^2,$$

Entonces:

$$\|S_\lambda(\tilde{\boldsymbol{x}})\|_2' = \frac{1}{2}(\|S_\lambda(\tilde{\boldsymbol{x}})\|_2^2)^{\frac{1}{2}-1}\cdot\left(-2\sum_{i=1}^{j}\tilde{x}_i + 2\lambda j\right) = \frac{-2\sum_{i=1}^{j}\tilde{x}_i + 2\lambda j}{2\|S_\lambda(\tilde{\boldsymbol{x}})\|_2}$$

$$= \frac{-2\left(\sum_{i=1}^{j}\tilde{x}_i - \lambda j\right)}{\|S_\lambda(\tilde{\boldsymbol{x}})\|_2} = \frac{-\|S_\lambda(\tilde{\boldsymbol{x}})\|_1}{\|S_\lambda(\tilde{\boldsymbol{x}})\|_2}$$

Integrando ambos resultados,

$$\psi'(\lambda) = \frac{-j\|S_\lambda(\tilde{\boldsymbol{x}})\|_2 - \|S_\lambda(\tilde{\boldsymbol{x}})\|_1\cdot\frac{-\|S_\lambda(\tilde{\boldsymbol{x}})\|_1}{\|S_\lambda(\tilde{\boldsymbol{x}})\|_2}}{\|S_\lambda(\tilde{\boldsymbol{x}})\|_2{}^2} \tag{3.32}$$

$$= \frac{\frac{\|S_\lambda(\tilde{\boldsymbol{x}})\|_1{}^2}{\|S_\lambda(\tilde{\boldsymbol{x}})\|_2} - j\|S_\lambda(\tilde{\boldsymbol{x}})\|_2}{\|S_\lambda(\tilde{\boldsymbol{x}})\|_2{}^2}$$

$$= \frac{\|S_\lambda(\tilde{\boldsymbol{x}})\|_1{}^2 - j\|S_\lambda(\tilde{\boldsymbol{x}})\|_2{}^2}{\|S_\lambda(\tilde{\boldsymbol{x}})\|_2{}^3}$$

$$= \frac{\|S_\lambda(\tilde{\boldsymbol{x}})\|_1{}^2}{\|S_\lambda(\tilde{\boldsymbol{x}})\|_2{}^3} - j\frac{\|S_\lambda(\tilde{\boldsymbol{x}})\|_2{}^2}{\|S_\lambda(\tilde{\boldsymbol{x}})\|_2{}^3}$$

$$= \frac{1}{\|S_\lambda(\tilde{\boldsymbol{x}})\|_2}\cdot\left[\frac{\|S_\lambda(\tilde{\boldsymbol{x}})\|_1{}^2}{\|S_\lambda(\tilde{\boldsymbol{x}})\|_2{}^2} - j\right]$$

$$= \frac{1}{\|S_\lambda(\tilde{\boldsymbol{x}})\|_2}\cdot[\psi^2(\lambda) - j]$$

Por el lema 1 y dado que el número de elementos no nulos de $S_\lambda(\tilde{\boldsymbol{x}})$ es igual a j, se verifica que:

$$\|S_\lambda(\tilde{\boldsymbol{x}})\|_1 \leq \sqrt{j}\|S_\lambda(\tilde{\boldsymbol{x}})\|_2$$

Por ello,

$$\frac{\|S_\lambda(\tilde{\boldsymbol{x}})\|_1}{\|S_\lambda(\tilde{\boldsymbol{x}})\|_2} \leq \sqrt{j} \quad \rightarrow \quad \left(\frac{\|S_\lambda(\tilde{\boldsymbol{x}})\|_1}{\|S_\lambda(\tilde{\boldsymbol{x}})\|_2}\right)^2 \leq j$$

$$\frac{\|S_\lambda(\tilde{\boldsymbol{x}})\|_1{}^2}{\|S_\lambda(\tilde{\boldsymbol{x}})\|_2{}^2} \leq j \quad \rightarrow \quad \psi(\lambda)^2 \leq j$$

De este resultado en (3.32) puede seguirse fácilmente que $\frac{1}{\|S_\lambda(\tilde{\boldsymbol{x}})\|_2} > 0$ y $(\psi^2(\lambda) - j) \leq 0$, luego:

$$\psi'(\lambda) = \frac{1}{\|S_\lambda(\tilde{\boldsymbol{x}})\|_2} \cdot [\psi^2(\lambda) - j] \leq 0$$

Así se demuestra que $\psi(\lambda)$ es una función decreciente, por ser una función continua de derivada negativa.

■

Por ser $\psi(\lambda)$ continua y cumplir que $\psi(0) \leq \sqrt{J}$ y $\psi(v) \leq \sqrt{J_{MAX}}$ para $\tilde{\boldsymbol{x}} = (\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_J)$, $\forall \tau \in [\sqrt{J_{MAX}}, \sqrt{J}]$ existe $k \in \mathbb{Z}$, $k \leq J$, tal que:

$$\psi(\tilde{x}_k) \leq \tau < \psi(\tilde{x}_{k+1}) \tag{3.33}$$

(recuérdese que $\tilde{x}_k \geq \tilde{x}_{k+1}$) o, equivalentemente existe $\delta \in [0, \tilde{x}_k - \tilde{x}_{k+1}]$ de manera que:

$$\psi(\tilde{x}_k - \delta) = \tau \tag{3.34}$$

Por la definición de ψ,

$$\psi(\tilde{x}_k - \delta) = \frac{\left\|S_{\tilde{x}_k - \delta}(\tilde{\boldsymbol{x}})\right\|_1}{\left\|S_{\tilde{x}_k - \delta}(\tilde{\boldsymbol{x}})\right\|_2} \tag{3.35}$$

Desarrollando esta expresión:

$$\begin{aligned}\left\|S_{\tilde{x}_k-\delta}(\tilde{\boldsymbol{x}})\right\|_1 &= \sum_{i=1}^{k}(\tilde{x}_i-(\tilde{x}_k-\delta)) = \sum_{i=1}^{k}\tilde{x}_i - \sum_{i=1}^{k}(\tilde{x}_k-\delta) \\ &= \sum_{i=1}^{k}\tilde{x}_i - k\cdot(\tilde{x}_k-\delta) = \sum_{i=1}^{k}\tilde{x}_i - k\tilde{x}_k + k\delta \\ &= \left(\sum_{i=1}^{k}\tilde{x}_i - k\tilde{x}_k\right) + k\delta = \left\|S_{\tilde{x}_k}(\tilde{\boldsymbol{x}})\right\|_1 + k\delta\end{aligned} \tag{3.36}$$

$$\begin{aligned}\left\|S_{\tilde{x}_k-\delta}(\tilde{\boldsymbol{x}})\right\|_2^2 &= \sum_{i=1}^{k}(\tilde{x}_i-(\tilde{x}_k-\delta))^2 = \sum_{i=1}^{k}\left((\tilde{x}_i-\tilde{x}_k)+\delta\right)^2 \\ &= \sum_{i=1}^{k}((\tilde{x}_i-\tilde{x}_k)^2 + 2\delta(\tilde{x}_i-\tilde{x}_k)+\delta^2) \\ &= \sum_{i=1}^{k}(\tilde{x}_i-\tilde{x}_k)^2 + 2\delta\sum_{i=1}^{k}(\tilde{x}_i-\tilde{x}_k) + k\delta^2 \\ &= \left\|S_{\tilde{x}_k}(\tilde{\boldsymbol{x}})\right\|_2^2 + 2\delta\left\|S_{\tilde{x}_k}(\tilde{\boldsymbol{x}})\right\|_1 + k\delta^2\end{aligned} \tag{3.37}$$

Por tanto, como:

$$\begin{cases}\psi(\tilde{x}_k-\delta) = \dfrac{\left\|S_{\tilde{x}_k-\delta}(\tilde{\boldsymbol{x}})\right\|_1}{\left\|S_{\tilde{x}_k-\delta}(\tilde{\boldsymbol{x}})\right\|_2} \\ \psi(\tilde{x}_k-\delta)=\tau\end{cases} \quad \rightarrow \quad \frac{\left\|S_{\tilde{x}_k-\delta}(\tilde{\boldsymbol{x}})\right\|_1}{\left\|S_{\tilde{x}_k-\delta}(\tilde{\boldsymbol{x}})\right\|_2} = \tau \tag{3.38}$$

Entonces,

$$\left\|S_{\tilde{x}_k-\delta}(\tilde{\boldsymbol{x}})\right\|_1 = \tau\cdot\left\|S_{\tilde{x}_k-\delta}(\tilde{\boldsymbol{x}})\right\|_2 \tag{3.39}$$

O lo que es lo mismo:

$$\left\|S_{\tilde{x}_k-\delta}(\tilde{\boldsymbol{x}})\right\|_1^{\,2} = \tau^2\cdot\left\|S_{\tilde{x}_k-\delta}(\tilde{\boldsymbol{x}})\right\|_2^{\,2}$$

Y utilizando las ecuaciones (3.36) y (3.37), se sigue que:

$$(\left\|S_{\tilde{x}_k}(\tilde{\boldsymbol{x}})\right\|_1 + k\delta)^2 = \tau^2\cdot\left(\left\|S_{\tilde{x}_k}(\tilde{\boldsymbol{x}})\right\|_2^2 + 2\delta\left\|S_{\tilde{x}_k}(\tilde{\boldsymbol{x}})\right\|_1 + k\delta^2\right)$$

Si desarrollamos esta expresión y utilizamos la notación $l_1 = \left\|S_{\tilde{x}_k}(\tilde{\boldsymbol{x}})\right\|_1$ y $l_2 = \left\|S_{\tilde{x}_k}(\tilde{\boldsymbol{x}})\right\|_2$ tal y como proponen los autores, podemos reescribir la ecuación anterior más fácilmente.

$$(l_1 + k\delta)^2 = \tau^2 \cdot (l_2^2 + 2\delta l_1 + k\delta^2)$$

Desarrollando las identidades notables y escribiéndolo como un polinomio de incógnita δ:

$$l_1^2 + k^2\delta^2 + 2l_1 k\delta = l_2^2\tau^2 + 2\delta l_1\tau^2 + k\tau^2\delta^2$$

$$k^2\delta^2 - k\tau^2\delta^2 + 2l_1 k\delta - 2\delta l_1\tau^2 + l_1^2 - l_2^2\tau^2 = 0$$

$$\delta^2(k^2 - k\tau^2) + 2l_1\delta(k - \tau^2) + (l_1^2 - l_2^2\tau^2) = 0 \qquad (3.40)$$

Por tanto, la solución a $\psi(\tilde{x}_k - \delta) = \tau$ se corresponde con la raíz positiva solución de la ecuación (3.40):

$$\delta = \frac{-2l_1(k - \tau^2) + \sqrt{4l_1^2(k^2 + \tau^4 - 2k\tau^2) - 4(k^2 - k\tau^2)(l_1^2 - l_2^2\tau^2)}}{2k(k - \tau^2)} \qquad (3.41)$$

$$= \frac{-2l_1(k - \tau^2) + 2\sqrt{l_1^2\tau^4 - l_1^2 k\tau^2 + l_2^2 k^2\tau^2 - l_2^2 k\tau^4}}{2k(k - \tau^2)}$$

$$= \frac{-2l_1(k - \tau^2) + 2\tau\sqrt{l_1^2(\tau^2 - k) + kl_2^2(k - \tau^2)}}{2k(k - \tau^2)}$$

$$= \frac{-2l_1(k - \tau^2) + 2\tau\sqrt{(l_1^2 - kl_2^2)(\tau^2 - k))}}{2k(k - \tau^2)}$$

$$= \frac{-l_1(k - \tau^2)}{k(k - \tau^2)} + \frac{\tau\sqrt{(l_1^2 - kl_2^2)(\tau^2 - k))}}{k(k - \tau^2)}$$

$$= \frac{-l_1}{k} + \frac{\tau}{k} \cdot \sqrt{\frac{(l_1^2 - kl_2^2)(\tau^2 - k))}{(k - \tau^2)^2}} = \frac{-l_1}{k} + \frac{\tau}{k} \cdot \sqrt{\frac{(l_1^2 - kl_2^2)}{(k - \tau^2)}}$$

Asumido que el parámetro k es conocido, se ha definido el valor del coeficiente δ tal que $\phi(a_k - \delta) = \tau$. Por tanto, el coeficiente λ que cumple $\psi(\lambda) = \tau$ es:

$$\lambda = \tilde{x}_k - \delta \qquad (3.42)$$

Usando (3.41), eso implica que

$$\lambda = \tilde{x}_k - \left(\frac{-l_1}{k} + \frac{\tau}{k} \cdot \sqrt{\frac{(l_1^2 - k l_2^2)}{(k - \tau^2)}} \right) \tag{3.43}$$

Una vez definido λ**,** el vector solución $\boldsymbol{x}$ viene dado por el operador *soft-thresholding* y la normalización del vector:

$$\boldsymbol{y} := P_{\tau,1}^{\mathfrak{B}_{\ell_1} \cap \mathfrak{B}_{\ell_2}}(\boldsymbol{x}) = \frac{S_\lambda(\boldsymbol{x})}{\|\boldsymbol{x}\|_2} = \frac{\operatorname{sign}(\boldsymbol{x}) \cdot (|\boldsymbol{x}| - \lambda)_+}{\|\boldsymbol{x}\|_2} \tag{3.44}$$

El esquema global del proceso puede verse en la Figura 50.

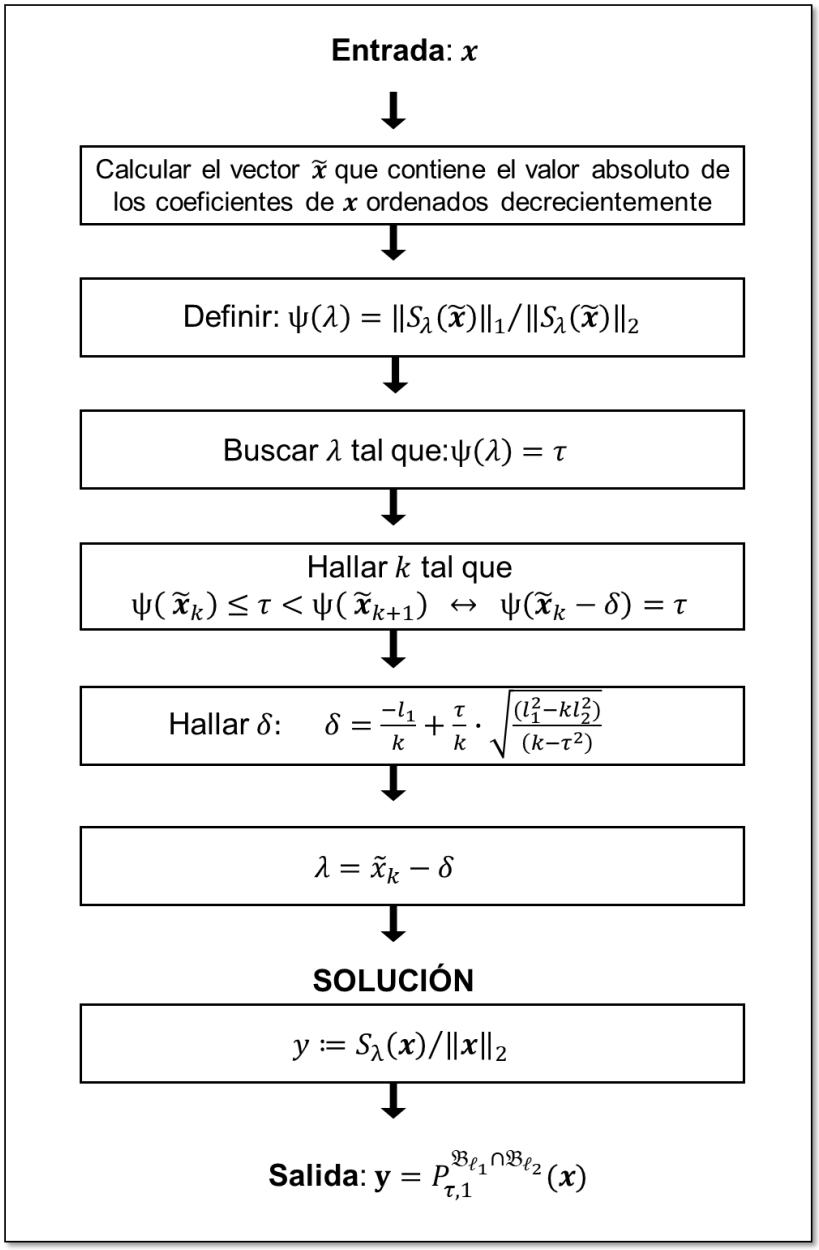


***Figura 50**. Esquema del método propuesto en (Guillemot et al., 2019) para la proyección sobre la región $\mathfrak{B}_{\ell_1} \cap \mathfrak{B}_{\ell_2}$*

Algoritmo de proyección de un vector sobre el espacio $\mathfrak{B}_{\ell_1} \cap \mathfrak{B}_{\ell_2}$ en tiempo lineal

La Tabla 19 refleja el algoritmo que Guillemot et al. (2019) proponen en su trabajo para la proyección euclídea de un vector $\boldsymbol{x}$ sobre la intersección de los balones $\mathfrak{B}_{\ell_1} \cap \mathfrak{B}_{\ell_2}$ de radios τ y 1 respectivamente. La programación de sus algoritmos en R está disponible en el repositorio público *github* (véase https://github.com/vguillemot/csvd) como parte de la librería csvd. Los autores

basan su algoritmo en el algoritmo de proyección en conjuntos convexos POCS; es decir, en la proyección alterna de un vector en ambos espacios convexos.

***Tabla 19**. Algoritmo de proyección de un vector sobre la restricción $\mathfrak{B}_{\ell_1} \cap \mathfrak{B}_{\ell_2}$ (Guillemot et al. 2019)*

Algoritmo CSVD-1: Proyección de v sobre el $\mathfrak{B}_{\ell_1}(\tau) \cap \mathfrak{B}_{\ell_2}(1)$ de radio τ	
Entrada:	$\boldsymbol{x} \in \mathbb{R}^J, \tau \in \mathbb{R}$ con $\tau \in [1, \sqrt{J}]$
Salida	$\boldsymbol{y} = P_{\tau}^{\mathfrak{B}_{\ell_1} \cap \mathfrak{B}_{\ell_2}}(\boldsymbol{x})$
Inicialización:	$s_1 = 0, s_2 = 0, nb = 0, p = \lvert\boldsymbol{x}^*\rvert$ con $\boldsymbol{x}^* = \{x_i \in \boldsymbol{v}, i = 1, \dots, J / x_i \neq 0\}$

1: **Si** $\|\boldsymbol{x}\|_2 = 0$ **entonces devuelve** $\boldsymbol{y} = \boldsymbol{x}$

2: **Si** $\|\boldsymbol{x}\|_1 / \|\boldsymbol{x}\|_2 \leq \tau$ **entonces devuelve** $\boldsymbol{y} = \boldsymbol{x}$ #Ya cumple la restricción

3: **Si no**

4: **Mientras** TRUE hacer:

5: $N = length(p)$

6: Seleccionar $r \in \{1, \dots, N\}$ aleatorio con $a_r \neq \max(p)$

7: $a_k = p[r]$

8: #Partición de p en dos conjuntos disjuntos:

$H = \{p_j \text{ tal que } p_j < a_k\}$

$L = \{p_j \text{ tal que } p_j > a_k\}$

9: #Evaluación de k

$nb_{a_k} = longitud(\{p_j \text{ tal que } p_j == a_k\})$

10: $K = nb + longitud(L) + nb_{a_k}$

11: $$s_{low_1} = \sum_{i \in L} p_i + nb_{a_k} \cdot a_k$$

12: $$s_{low_2} = \sum_{i \in L} {p_i}^2 + nb_{a_k} \cdot {a_k}^2$$

13: $$\psi(a_k) = \frac{s_1 + s_{low_1} - k a_k}{\sqrt{s_2 + s_{low_2} - 2a_k(s_1 + s_{low_1}) + k{a_k}^2}}$$

14: **Si** $\psi(a_k) > \tau$ **entonces**

15: **Si** $L = \emptyset$ **entonces salir**

16: Actualización de p: $p = \{p_i\}_{i \in L}$

17: **Si no**

18: $a_{k+1} = \max(\{p_i\}_{i \in H})$

19: $$\psi(a_{k+1}) = \frac{s_1 + s_{low_1} - k a_{k+1}}{\sqrt{s_2 + s_{low_2} - 2a_{k+1}(s_1 + s_{low_1}) + k{a_{k+1}}^2}}$$

20: **Si** $\psi(a_{k+1}) > \tau$ **entonces salir**

21: Actualización de p: $p = \{p_i\}_{i \in H}$

22: Actualización de nb: $nb = k$

23: Actualización de s_1: $s_1 = s_1 + s_{low_1}$

24: Actualización de s_2: $s_2 = s_2 + s_{low_2}$

25: **Fin Si**

26: **Fin Mientras**

27: #Encontrar la penalización λ:

$$\lambda = a_k - \left(a \cdot \sqrt{\frac{k - \psi(a_k)^2}{k - a^2}} - \psi(a_k) \right) \cdot \frac{s_1 + s_{low_1} - k a_k}{k \psi(a_k)}$$

28: # Solución: $\boldsymbol{y} = P_{\tau}^{\mathfrak{B}_{\ell_1} \cap \mathfrak{B}_{\ell_2}}(\boldsymbol{x})$

$$\boldsymbol{y} \coloneqq \frac{S_\lambda(\boldsymbol{x})}{\|\boldsymbol{x}\|_2} = \frac{\text{sign}(\boldsymbol{x})(|\boldsymbol{x}| - \lambda, 0)_+}{\|\boldsymbol{x}\|_2}$$

29: **Fin Si**

3.2 Código en R

3.2.1 Código proyección de un vector sobre $\mathfrak{B}_{\ell_1}(\tau)$ en R

La siguiente función permite proyectar un vector v cualquiera sobre $\mathfrak{B}_{\ell_1}(\tau)$. Los pasos 4 y 5 del algoritmo de Berg et al. (2008) para la selección inicial del pivote han sido modificados. La selección del pivote según el algoritmo de mediana de medianas (Cormen et al., 2009) ha sido sustituido por la selección aleatoria del índice k, tal y como proponen los autores y tal y como realizan (Guillemot et al., 2019) en su código para la proyección en el balón $\mathfrak{B}_{\ell_1}(\tau) \cap \mathfrak{B}_{\ell_2}(1)$. Guillemot et al. (2019) proponen un algoritmo diferente para la proyección en el espacio $\mathfrak{B}_{\ell_1}(\tau)$, disponible para su consulta también en: https://github.com/vguillemot/csvd

```
lasso.proj <- function(b, tau){
 if ( norm1(b) <= tau ) return(x=b) #Ya cumple la restricción
 #Inicialización
s <- 0
nb <- 0
p <- abs(b[b != 0]) #Paso 1. Cálculo del vector de coeficientes en valores absolutos
#Paso 2. Procedimiento para encontrar k que verifica ϕ(a_k) ≤ τ < ϕ(a_{k+1})
while (T)
{
        N <- length(p)
        #Primera elección de k aleatoria
        a_k <- p[sample(1:N,1)]
        while (a_k == max(p)) {a_k <- p [sample (1: N,1)]}
        #Partición de p:
        p_inf_ak<-p<a_k
        p_sup_ak<-p>a_k
        p_high<-p[p_inf_ak]
        p_low<-p[p_sup_ak]
        #Cálculo de k
        nb_a_k<-sum(p==a_k)
        k<-nb + sum(p_sup_ak) + nb_a_k
        #Cálculo de la función ϕ
        slow<-sum(p_low) + nb_a_k*a_k
        phi_a_k<-s + slow-k*a_k
```

```
        #Divide y vencerás: selección de la partición en función de la restricción τ
        if (phi_a_k>tau)
                {
                if (length(p_low) ==0) break
                p<-p_low
                }
        else
                {
                if (length(p_high) ==0) {break}
                else
                        {
                        a_k_1<-max(p_high)
                        phi_a_k_1<-s+slow-k*a_k_1
                        if (phi_a_k_1>tau) {break}
                        p<-p_high #Actualización de p
                        nb<-k #Actualización de k
                        s<-s+slow #Actualización de s
                        }
                }
}
#Paso 3.Cálculo de δ y λ
lambda<-a_k-((tau-phi_a_k)/k)
#Paso 4.Cálculo del vector solución a partir del operador soft-thresholding
b.proj<-sign(b)*pmax(0,abs(b)-lambda)
```

```
return(x=b.proj)
}
```

3.2.2 Código proyección de un vector sobre $\mathfrak{B}_{\ell_1+\ell_2}(\tau,\gamma)$ en R

Siguiendo las instrucciones del algoritmo propuesto por (Mairal et al., 2009) para la proyección de un vector sobre la restricción Elastic net $\mathfrak{B}_{\ell_1+\ell_2}(\tau,\gamma)$ se ha programado el siguiente código en R:

```
enet.proj<-function(b, tau, gamma)
{
 norm.en<-(norm1(b))+(gamma/2)*norm2(b)
 if1 (norm.en<=tau) #Comprobación de que no se cumpla ya la restricción ENET
      {return(b)}
 else1
      {
           bb<-b #Funciona bien sin el valor absoluto
           s=0
           p=0
           p_inc=0
           s_inc=0
           U=1: length(bb)
           while(length(U)>0)
           {
                b_U<-bb[U]
```

```
#selección de k aleatoria:

k=sample(U,1)

#Particion de U:

G=U[which(abs(b_U)>=abs(b_U[which(U==k)]))]

L=U[which(abs(b_U)<abs(b_U[which(U==k)]))]

#Definicion de p y s

p_inc<-length(G)

s_inc<-sum(abs(bb[G])+(gamma/2)*(abs(bb[G])^2))

if2 ((s+s_inc-((p+p_inc)*(1+(gamma/2)*abs(b_U[which(U==k)]))*abs(b_U[which(U==k)]))<
(tau*((1+gamma*abs(b_U[which(U==k)]))^2))

    {s<-s+s_inc ; p<-p+p_inc  ;  U<-L} IF2


 else2

    {U=G[-which(G==k)]} ELSE2


}WHILE


a.sol<-(gamma^2)*tau+(gamma/2)*p  #Gamma no puede ser 0. Como mucho 10^(-6)

b.sol<-2*gamma*tau+p

c.sol<-tau-s

#Si lambda=0 la ecuación de 2o grado no se puede resolver:

lambda<-(-b.sol+sqrt((b.sol^2)-(4*a.sol*c.sol)))/(2*a.sol)

#Solución vector proyectado en Elastic Net:
```

```
        b.en.proj<-pmax(abs(b)-lambda,0)*sign(b)/(1+lambda*gamma)
return (list (v.original=b,penalizacion=lambda,proj.enet= b.en.proj))
  }
}
```

3.2.3 Código proyección de un vector sobre $\mathfrak{B}_{\ell_1}(\tau) \cap \mathfrak{B}_{\ell_2}(1)$ en R

El código en R para la proyección de un vector en la intersección de los balones ℓ_1 y ℓ_2 (Guillemot et al., 2019) puede consultarse en: https://github.com/vguillemot/csvd/blob/master/R/proj12.R

CAPÍTULO 4

DESCOMPOSICIÓN EN VALORES SINGULARES *RESTRINGIDA* C_{ENET}*SVD*: SOLUCIONES ORTOGONALES Y *SPARSE*

La SVD clásica propuesta por Eckart y Young (1936) supuso un antes y un después en el desarrollo de la Estadística Multivariante (Abdi, 2007; Puntanen, 2011). Se define como una herramienta algebraica de descomposición matricial, que aproxima una base de datos X_{IxJ} en el producto de tres matrices diferentes, proporcionando la mejor aproximación de bajo rango de la matriz inicial en el sentido de los mínimos cuadrados (Björck, 2015).

Como en todos los casos, facilitar la interpretación de los resultados obtenidos mediante distintas técnicas ha sido siempre un objetivo común. En este sentido la extensión de las técnicas clásicas a los métodos sparse ha supuesto un gran avance, al poder identificar subconjuntos de variables con mayor importancia en un sentido u otro. Los métodos sparse también han sido incorporados a la SVD, como la sparse SVD propuesta por (Lee et al., 2010) o la CSVD de (Guillemot et al., 2019).

En este capítulo se propone una extensión de la descomposición en valores singulares restringida (CSVD) (Guillemot et al., 2019) que generará vectores sparse y ortogonales al mismo tiempo. La metodología matemática propuesta en este capítulo, a la que denominaremos C_{enet}SVD tratará de lograr la restricción sparsity, penalizando la norma Elastic net de los vectores singulares; esto es, restringirá estos a la bola $\mathfrak{B}_{\ell_1+\ell_2}$. Esta propiedad es deseable cuando el número de variables excede en gran medida al número de observaciones, como puede suceder en genómica, neurociencia, … y cuando el usuario necesita facilitar la interpretación de sus resultados, como en el campo del análisis psicométrico de cuestionarios. Por otro lado, la ortogonalidad es una característica deseable que da lugar a vectores singulares no correlacionados y facilita el hecho de darles un significado. Además, esta característica de los nuevos vectores restringidos es necesaria para el desarrollo de otro tipo de técnicas de análisis de datos

mediante el uso de la SVD, como es el caso de los métodos de reducción de la dimensión.

Este capítulo se dirige a obtener dos objetivos diferentes. Por un lado, introducir una nueva formulación matemática de la CSVD de Guillemot et al. (2019), introduciendo restricciones de sparsity y ortogonalidad en el problema de optimización. Con este fin, en este capítulo se conseguirá la restricción sparse de los vectores pseudo-singulares limitando la norma Elastic net (combinación de las normas Lasso y Ridge) de los vectores. Por otro lado, implementar dos alternativas de métodos de reducción de la dimensión restringidas: el análisis de componentes principales restringido (sparse & ortogonal) C_{enet}PCA y las técnicas Biplot clásicas restringidas (sparse & ortogonal) C_{enet}Biplots. El PCA trata de representar el comportamiento de una muestra en un espacio formado por un menor número de variables latentes, conocidas como componentes principales y que se calculan a partir de la combinación lineal de todas las variables del modelo de partida. En el caso del C_{enet}PCA, cada una de estas componentes principales restringidas se obtendrá como una combinación de tan solo un subconjunto de las variables originales, de manera que los patrones de las observaciones puedan ser explicados de manera más sencilla a partir de un solo grupo de variables. Los métodos Biplot son técnicas de reducción de la dimensión, basadas en el PCA, cuyo fin último es la representación gráfica de observaciones y variables en un mismo sistema de referencia.

Para lograr estos objetivos, el capítulo comenzará con el que supondrá ser el cimiento teórico de las metodologías propuestas posteriores (C_{enet}SVD, C_{enet}PCA, C_{enet}Biplots): la proyección de un vector sobre la intersección de la región definida por la norma Elastic net y su normalización. Posteriormente se desarrollará la formulación matemática de las técnicas sparse mencionadas y su implementación en el software R. Finalmente, las metodologías propuestas se aplicarán al análisis de datos reales de alta dimensión. Considerando todo esto, se concluye que los métodos que aquí se proponen son herramientas muy prometedoras para el análisis multivariante, como lo ha sido todo el círculo de técnicas en torno a los métodos sparse, y abren un abanico de posibilidades en múltiples disciplinas de investigación.

4.1 Marco teórico

Como se ha descrito anteriormente, uno de los principales propósitos al examinar bases de datos multivariantes es obtener una reproducción de los datos en un espacio reproducido. En este sentido la SVD representa el pilar principal de una gran cantidad de metodologías estadísticas del análisis de dos vías (Beaton, Chin Fatt, & Abdi, 2014) como el PCA (Jolliffe et al., 2016), el MDS (Borg & Groenen, 2003), CA (Greenacre, 2017) y numerosas técnicas de análisis multivía como el MFA (Abdi et al., 2013), STATIS (Abdi et al., 2012) o JIVE (E. F. Lock et al., 2013).

De todas las técnicas de reducción de la dimensión, el PCA es la más utilizada con diferencia en el campo estadístico (Jolliffe et al., 2016). Para ello, identifica y extra un conjunto de nuevas variables latentes, PCs, que se calculan como una combinación lineal de las variables originales de manera que los coeficientes de dichas combinaciones (cargas) denotan la contribución de cada característica original a la formación de las PCs. En la práctica, una situación ideal llevaría a la obtención de cargas exactamente cero (cargas sparse), de manera que la interpretación de las PCs dependiese solo de un subconjunto de las variables originales. Desafortunadamente, esto no suele ocurrir en el análisis de datos reales de manera que no existe garantía de proporcionar un significado a esos conceptos matemáticos escogidos por sus propiedades óptimas de síntesis. Este hecho esconde la capacidad informativa de los datos y, más aún, en el análisis de matrices de altas dimensiones donde uno de los objetivos principales de los estudios suele ser la selección automática de las características más relevantes.

A lo largo de los años la literatura ha reunido diferentes enfoques para generar cargas nulas. Cadima y Jolliffe (1995) sugieren el uso de la umbralización; esto es, asumir como nulas aquellas cargas cuyo valor absoluto se encuentre por debajo de un cierto umbral. Anaya-Izquierdo, Critchley y Vines (2011) y Vines (2000) proponen restringir el valor de las cargas a un cierto subconjunto de enteros. Los puntos de vista más modernos se centran en los métodos de regularización para producir coeficientes nulos. Estos incluyen penalización que promuevan la propiedad *sparsity* (variables con coeficientes

exactamente cero) en la formulación del problema de optimización (Hastie & Tibshirani, 2015). Los métodos de regularización han ganado mucha popularidad porque permiten controlar el sobreajuste en la estimación de parámetros de modelos así como realizar selección de variables (Huang, Liu, & Liang, 2016; X. Liu et al., 2019). Su objetivo es introducir una restricción sobre la norma Lp de un vector en la función de pérdida del problema de optimización para conseguir convertirlo en un vector sparse. El método de regularización más utilizado con diferencia es Lasso (*Least Absolute Shrinkage and Selection Operator*) (Tibshirani, 1996), que restringe la norma L1 de un vector penalizando la suma de los valores absolutos de sus coeficientes. Lasso es una técnica realmente útil en el análisis de datos de altas dimensiones, donde la identificación de las variables relevantes es un hecho fundamental (como dice el refrán, "*como buscar una aguja en un pajar…")* (Wong et al., 2019).

Como se ha comentado anteriormente, Lasso presenta algunos inconvenientes. No es un procedimiento oracle (Fan & Li, 2001) y permite que aparezcan características redundantes en el modelo estimado. Por otra parte, la esencia de los métodos multivariantes radica en aprovechar la relación entre las variables para explicar los patrones de los datos. Sin embargo, si hay un grupo de variables correlacionadas, Lasso tiende a seleccionar una variable del grupo. Esto supone una inconsistencia práctica en diversas disciplinas, como en el análisis de la expresión génica de microarrays, donde es importante tener en cuenta la actividad conjunta de los genes en múltiples mecanismos biológicos (Hore et al., 2016; Wang et al., 2015), o en el análisis psicométrico de cuestionarios en psicología, donde cada una de las construcciones latentes está compuesta por un conjunto de ítems (Barahona et al., 2018; Vega-Hernández et al., 2018). Así, para superar esto, Zou y Hastie (2005) proponen Elastic net (ENET, L1+L2), una combinación lineal de L1 (Lasso) y L2 (Ridge) que permite que las variables relacionadas aparezcan juntas en el modelo sparse.

Estrechamente relacionado con este trabajo, se han propuesto en la literatura diferentes métodos de factorización matricial penalizada (Kim & Park, 2008; M. Lee et al., 2010; Witten et al., 2009). Desafortunadamente, estas técnicas proporcionan soluciones dispersas u ortogonales, pero no ambas propiedades simultáneamente. Recientemente, Guillemot et al. (2019) proponen

la SVD restringida (CSVD, por sus siglas en inglés), método que integra simultáneamente la sparsity vía Lasso y la ortogonalidad, basado en la anterior Factorización Matricial Penalizada (PMD, por sus siglas en inglés) de Witten et al. (2009). Nuestra propuesta se centra en una de las líneas de futuro propuestas por Guillemot et al. (2019): la incorporación de diferentes restricciones a la CSVD, siempre que estas restricciones puedan expresarse como proyección sobre conjuntos convexos. Siguiendo sus recomendaciones, el tema de este trabajo es diseñar una CSVD sparse y ortogonal, pero restringiendo la norma ENET de los vectores pseudo-singulares (C_{enet}SVD).

4.2 C_{enet}SVD

Considerando todo lo expuesto anteriormente, esta sección se organiza de la siguiente forma. A continuación se definirá la CSVD restringida a Elastic net como solución a un problema de optimización convexa con restricciones. Para ello previamente se mostrará la metodología seguida para lograr proyectar un vector normalizado sobre el espacio Elastic net $\mathfrak{B}_{\ell_1+\ell_2} \cap \mathfrak{B}_{\ell_2}(\tau, 1)$**.** La proyección de un vector sobre la intersección convexa de Elastic net y la norma L2 es el pilar teórico fundamental del método. La C_{enet}SVD que proponemos es la extensión de la CSVD de Guillemot et al. (2019), quien enuncia la posibilidad de extender su propuesta a otros tipos de restricciones convexas. En este caso, se mostrará su extensión a la regularización Elastic net, que restringe los coeficientes a cero a la vez que asegura que variables correlacionadas puedan presentar coeficientes similares y aparecer juntas en el modelo. C_{enet}SVD es una herramienta efectiva tanto para bases de datos de altas dimensiones como para bases de datos tradicionales en las que el número de observaciones es superior al número de variables.

A continuación se mostrará la formulación de C_{enet}SVD y su pseudocógido, unificando el algoritmo POCS con las ideas de los algoritmos *divide y vencerás*. Posteriormente se hablará acerca de la selección del parámetro de regularización. Para finalizar, se presentará la implementación del sparse y ortogonal C_{enet}PCA con restricción sobre Elastic net, así como la versión modificada sparse y ortogonal de los métodos Biplot Clásicos o C_{enet}Biplots. Por último se mostrará la utilidad de la metodología propuesta mediante el análisis

de un conjunto de datos real en el campo de la genómica: análisis de expresión en muestras con leucemia.

4.2.1 Notación

Se recuerda brevemente parte de la notación útil para seguir adecuadamente este capítulo. La matriz traspuesta de $\boldsymbol{X}$ se denota como $\boldsymbol{X}^{\mathrm{T}}$ y su inversa como $\boldsymbol{X^{-1}}$. La norma de Frobenius de una matriz $\boldsymbol{X}$ se define como $\|\boldsymbol{X}\|_{\boldsymbol{F}}^{\boldsymbol{2}} = \mathrm{traza}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})$. La norma L2 de un vector es calculada mediante $\sqrt{\sum {x_{ij}}^2}$ y la norma L1 se estima como $\sum|x_{ij}|$. Diremos que un vector está normalizado cuando es dividido por su norma L2. Habitualmente, $\boldsymbol{U}$ denotará la matriz de vectores singulares a izquierda de la SVD clásica y $\boldsymbol{V}$ la matriz compuesta por los vectores singulares a derecha. $\boldsymbol{D}$ es una matriz diagonal que contiene los valores singulares de la SVD. Las regiones de restricción son bolas definidas como $\mathfrak{B}_{\tau}^{\ell_2}(\boldsymbol{x}) = \{\boldsymbol{x} \,/\, \|\boldsymbol{x}\|_2^2 \leq \tau\}$, $\mathfrak{B}_{\tau}^{\ell_1}(\boldsymbol{x}) = \{\boldsymbol{x} \,/\, \|\boldsymbol{x}\|_1 \leq \tau\}$ y $\mathfrak{B}_{\tau}^{\ell_1+\ell_2}(\boldsymbol{x}) = \{\boldsymbol{x} \,/\, (1-\alpha)\|\boldsymbol{x}\|_1 + \alpha\|\boldsymbol{x}\|_2^2 \leq \tau\}$, para algún $\alpha \in [0,1]$.

4.2.2 Descomposición en valores singulares clásica

Dada una matriz $\boldsymbol{X}_{IxJ}$ de rango $R \leq \min(I,J)$ que se supone centrada y estandarizada sin pérdida de generalidad. La SVD de $\boldsymbol{X}$ se define como la aproximación en bajo rango de la matriz original expresada mediante el producto:

$$\boldsymbol{X}_{IxJ} = \boldsymbol{U}_{IxR}\boldsymbol{D}_{RxR}{\boldsymbol{V}^T}_{RxJ} \tag{4.1}$$

donde $\boldsymbol{U} = [\boldsymbol{U}_1, \ldots, \boldsymbol{U}_I]$ y $\boldsymbol{V} = [\boldsymbol{V}_1, \ldots, \boldsymbol{V}_J]$ son matrices ortonormales, $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}$ y $\boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}$, cuyos vectores columna son los vectores singulares a izquierda y derecha respectivamente y $\boldsymbol{D}$ es la matriz diagonal que almacena los valores singulares de $\boldsymbol{X}$, expresados convenientemente de manera que $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_R \geq 0$. Para un $Q \leq R$ óptimo, la SVD proporciona la mejor aproximación de rango Q, $\widehat{\boldsymbol{X}}_Q$, de la matriz original $\boldsymbol{X}$ en el sentido de los mínimos cuadrados. Para ello se minimiza la norma de Frobenius entre la base de datos inicial y la matriz reconstruida (Eckart & Young, 1936; Shen & Huang, 2008).

$$\left\|\boldsymbol{X} - \widehat{\boldsymbol{X}}_Q\right\|_F^2 = \|\boldsymbol{X} - UDV^T\|_F^2 \tag{4.2}$$

y $\widehat{\boldsymbol{X}}_Q, Q \leq R$:

$$\widehat{\boldsymbol{X}}_Q = \boldsymbol{U}_{IxQ}\boldsymbol{D}_{QxQ}\boldsymbol{V}^T{}_{QxJ} = \sum_{q=1}^{Q} d_{\mathrm{q}}\boldsymbol{u}_q\boldsymbol{v}_q{}^T \tag{4.3}$$

con $\boldsymbol{u}_q{}^T\boldsymbol{u}_q = \boldsymbol{v}_q{}^T\boldsymbol{v}_q = 1$ y $\boldsymbol{u}_q{}^T\boldsymbol{u}_{q'} = \boldsymbol{v}_q{}^T\boldsymbol{v}_{q'} = 0$ $\forall q \neq q'$. La SVD se expresa como la solución al problema de optimización:

$$\begin{aligned} &\underset{d,\boldsymbol{u},\boldsymbol{v}}{\operatorname{argmin}} \frac{1}{2}\left\|\boldsymbol{X} - d_q\boldsymbol{u}_q\boldsymbol{v}_q{}^T\right\|_F^2 \\ &\text{s.a. } \left\{\boldsymbol{u}_q{}^T\boldsymbol{u}_q = \boldsymbol{v}_q{}^T\boldsymbol{v}_q = 1, \boldsymbol{u}_q{}^T\boldsymbol{u}_{q'} = \boldsymbol{v}_q{}^T\boldsymbol{v}_{q'} = 0 \quad \forall q \neq q'\right. \end{aligned} \tag{4.4}$$

Guillemot et al. (2019) proponen un algoritmo basado en la proyección de un vector en la intersección de un conjunto de espacios convexos para resolver el problema anterior, basándose en el algoritmo POCS (Bauschke & Combettes, 2017). Proponen reemplazar el proceso de ortogonalización de los vectores singulares por la proyección de los mismos en la intersección del espacio convexo definido por la norma L2 y el espacio ortogonal $M^{\perp}$ a los vectores singulares ya estimados. El código que proponen es el que se muestra en la Tabla 20.

Tabla 20. *Algoritmo para la implementación de la SVD clásica basado en el algoritmo POCS* (Guillemot et al., 2019)

Algoritmo: POCS – SVD. Proyección de un vector en el espacio $\mathfrak{B}_{\ell_2}(\mathbf{1}) \cap M^{\perp}$	
Entrada:	$\boldsymbol{X} \in \mathbb{R}^{IxJ}$, rango Q, $\varepsilon \approx 0$
Salida:	$\boldsymbol{U} \in \mathbb{R}^{IxQ}, \boldsymbol{D} \in \mathbb{R}^{QxQ}, \boldsymbol{V} \in \mathbb{R}^{JxQ}$
Inicialización:	$\boldsymbol{U} = \mathbf{0}, \boldsymbol{V} = \mathbf{0},$ $\lambda_0 = 0$
1: **Para** q en $1{:}\,Q$ hacer:	
2: Inicializar $\boldsymbol{u}_0, \boldsymbol{v}_0$ aleatoriamente	
3: $\lambda_1 = \boldsymbol{u}_0^T\boldsymbol{X}\boldsymbol{v}_0$	
$t = 0$	
Mientras $\lvert\lambda_{t+1} - \lambda_t\rvert \geq \varepsilon$ hacer:	
$\boldsymbol{u}_{t+1} = \mathrm{proj}^{\mathfrak{B}_{\ell_2}(1)\cap\boldsymbol{U}^{\perp}}(\boldsymbol{X}\boldsymbol{v}_t)$	
$\boldsymbol{v}_{t+1} = \mathrm{proj}^{\mathfrak{B}_{\ell_2}(1)\cap\boldsymbol{V}^{\perp}}(\boldsymbol{X}^T\boldsymbol{u}_{t+1})$	
$\lambda_{t+1} = \boldsymbol{u}_{t+1}^T\boldsymbol{X}\boldsymbol{v}_{t+1}$	
$t = t + 1$	
4: $\boldsymbol{d} = [\boldsymbol{d}, \lambda_{t+1}]$	
5: $\boldsymbol{U} = [\boldsymbol{U}, \boldsymbol{u}_{t+1}]$	
6: $\boldsymbol{V} = [\boldsymbol{V}, \boldsymbol{v}_{t+1}]$	
7: **Fin**	
8: $\boldsymbol{D} = \mathrm{diag}(d)$	
9: **Fin**	

Antes de extender el algoritmo de proyección en espacios convexos para la obtención de la SVD ortogonal y penalizada sobre Elastic net, se desarrolla el método de proyección de un vector sobre la intersección convexa de los espacios Elastic net y la norma L2 ($\mathfrak{B}_{(\ell_1+\ell_2)\cap\ell_2}(\tau,1)$). Este es el punto esencial para la propuesta posterior de CenetSVD.

4.2.3 Solución general al problema de proyección de un vector sobre el espacio $\mathfrak{B}_{\ell_1+\ell_2}(\tau)$

Se presenta a continuación una formulación diferente y equivalente a la presentada por Mairal et al. en la sección 2 para la proyección de un vector sobre la bola $\mathfrak{B}_{\ell_1+\ell_2}$. Dado $\boldsymbol{x} \in \mathbb{R}^J$, Mairal et al. (2010) definen la proyección de un vector $\boldsymbol{y} = P_\tau^{\mathfrak{B}_{\ell_1+\ell_2}}(\boldsymbol{x}) \in \mathbb{R}^J$ con $\gamma > 0$ sobre el espacio restringido de la norma Elastic net como la solución al problema de optimización con restricción:

$$\boldsymbol{y} := P_\tau^{\mathfrak{B}_{\ell_1+\ell_2}}(\boldsymbol{x}) = \left\{ \underset{\boldsymbol{y}\in\mathbb{R}^J}{\operatorname{argmin}} \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{x}\|_2^2 \,;\, s.a\colon \|\boldsymbol{y}\|_1 + \frac{\gamma}{2}\|\boldsymbol{y}\|_2^2 \le \tau \right\} \tag{4.5}$$

Encontrar la solución al problema de optimización (4.5) es equivalente a encontrar la solución del problema de optimización restringido:

$$\boldsymbol{y} := P_\tau^{\mathfrak{B}_{\ell_1+\ell_2}}(\boldsymbol{x}) = \left\{ \underset{\boldsymbol{y}\in\mathbb{R}^J}{\operatorname{argmin}} \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{x}\|_2^2 \,;\, s.a\colon (1-\alpha)\|\boldsymbol{y}\|_1 + \alpha\|\boldsymbol{y}\|_2^2 \le \tau \right\}$$

con $\alpha \in [0,1]$. La función $(1-\alpha)\|\boldsymbol{y}\|_1 + \alpha\|\boldsymbol{y}\|_2^2$, definida como la penalización Elastic net, es una combinación convexa de las penalizaciones Lasso y Ridge, poseyendo las propiedades óptimas de ambos operadores. Para $\alpha \in (0,1)$, esta norma se convierte en un método automático de selección de variables (debido a que incluye en su definición una restricción sobre la norma Lasso), siendo así apta para análisis de bases de datos de altas dimensiones, y de selección de grupos de variables altamente correlacionadas (por la inclusión de la restricción Ridge). Es inmediato comprobar que si $\alpha = 0$, el problema de optimización se convierte en un problema de proyección del vector sobre el balón de norma ℓ_1. Por el contrario, si $\alpha = 1$, el problema se restringe a la proyección del vector sobre la norma Ridge.

La solución a este problema de optimización se determina a continuación. Supuesto $\boldsymbol{x} \geq 0$, se construye el problema de optimización dual y la función Lagrangiana asociada al problema de proyección, introduciendo el multiplicador de Lagrange $\lambda \in \mathbb{R}$, como sigue:

$$L(\boldsymbol{y}, \lambda) = \underset{\boldsymbol{y} \in \mathbb{R}^J}{\operatorname{argmin}} \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{x}\|_2^2 \; + \; \lambda((1-\alpha)\|\boldsymbol{y}\|_1 + \alpha\|\boldsymbol{y}\|_2^2 - \tau)$$

$$L(\boldsymbol{y}, \lambda) = \frac{1}{2}(\boldsymbol{y} - \boldsymbol{x})^2 + \lambda(1-\alpha)\|\boldsymbol{y}\|_1 + \lambda\alpha\|\boldsymbol{y}\|_2^2 - \lambda\tau$$

$$L(\boldsymbol{y}, \lambda) = \frac{1}{2}(\boldsymbol{y}^t\boldsymbol{y} + \boldsymbol{x}^t\boldsymbol{x} - 2\boldsymbol{y}^{\mathrm{t}}\boldsymbol{x}) + \lambda(1-\alpha)\boldsymbol{y} + \lambda\alpha\boldsymbol{y}^{\mathrm{t}}\boldsymbol{y} - \lambda\tau$$

El problema de Lagrange dual a (46) es aquel que trata de maximizar $L(\boldsymbol{y}, \lambda)$ para $\lambda \geq 0$:

$$\operatorname{argmax} L(\boldsymbol{y}, \lambda) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{x}\|_2^2 + \lambda(1-\alpha)\|\boldsymbol{y}\|_1 + \alpha\lambda\|\boldsymbol{y}\|_2^2 - \lambda\tau \qquad (4.6)$$

$$s.a: \lambda \geq 0$$

con $\lambda \in \mathbb{R}$. Los valores óptimos de (4.5) y (4.6) son los mismos, pero el cálculo de la solución es más eficiente en (4.6), pues la variable a optimizar es un escalar. Por las condiciones de Karush–Kuhn–Tucker (KKT) de optimización matemática, generalización del método de los multiplicadores de Lagrange, la solución λ óptima viene dada por punto donde el gradiente de $L(\boldsymbol{y}, \lambda)$ se anula.

$$\nabla_{\boldsymbol{y}} L(\boldsymbol{y}, \lambda) = 0$$

$$\nabla_{\boldsymbol{y}} L(\boldsymbol{y}, \lambda) = \frac{1}{2}(2\boldsymbol{y} - 2\boldsymbol{x}) + \lambda(1-\alpha) + 2\lambda\alpha\boldsymbol{y} = 0$$

$$\nabla_{\boldsymbol{y}} L(\boldsymbol{y}, \lambda) = \boldsymbol{y} - \boldsymbol{x} + \lambda(1-\alpha) + 2\lambda\alpha\boldsymbol{y} = 0$$

$$\boldsymbol{y}(1 + 2\lambda\alpha) \; - \boldsymbol{x} + \lambda(1-\alpha) = 0$$

$$\boldsymbol{y} = \frac{\boldsymbol{x} - \lambda(1-\alpha)}{1 + 2\lambda\alpha}$$

Una vez calculado el punto estacionario, si este es positivo entonces se trata de la solución óptima. En caso contrario, la solución óptima será 0. Generalizando:

$$\lambda = \max(0, \boldsymbol{y})$$

donde $\boldsymbol{y}$ es la solución de $\nabla_{\boldsymbol{y}} L(\boldsymbol{y}, \lambda) = 0$. Con todo ello, la solución al problema de optimización (46) viene dada por:

$$\boldsymbol{y} = \frac{S_{\lambda(1-\alpha)}(\boldsymbol{x})}{1 + 2\lambda\alpha} = \begin{cases} (x_j + \lambda(1-\alpha))/1 + 2\lambda\alpha & si\ x_j < \lambda(1-\alpha) \\ 0 & si\ x_j \in [-\lambda(1-\alpha), \lambda(1-\alpha)] \\ (x_j - \lambda(1-\alpha))/1 + 2\lambda\alpha & si\ x_j > \lambda(1-\alpha) \end{cases}$$

Una vez generalizada la solución, es inmediato comprobar que se verifica para $\alpha = 1$ (Ridge) que la solución coincide con la solución del problema de proyección en la norma ℓ_2^2:

$$\boldsymbol{y} = \frac{S_0(\boldsymbol{x})}{1 + 2\lambda} = \frac{\boldsymbol{x}}{1 + 2\lambda}$$

Si $\alpha = 0$, la solución que se obtiene es la correspondiente al operador *soft-thresholding*, solución a la proyección sobre la restricción de la norma Lasso:

$$\boldsymbol{y} = S_\lambda(\boldsymbol{x})$$

El vector solución sparse $\boldsymbol{y} := P_{\tau,1}^{\mathfrak{B}_{\ell_1+\ell_2} \cap \mathfrak{B}_{\ell_2}}(\boldsymbol{x})$ puede obtenerse, así, como una composición de los operadores solución de Lasso y Ridge, al ser Elastic net una función de regularización que combina la norma ℓ_1 y la norma clásica ℓ_2^2:

$$\boldsymbol{y} = Proj_{\lambda(\alpha\|\cdot\|_1 + (1-\alpha)\|\cdot\|_2^2)}(\boldsymbol{x}) = \left(Proj_{\lambda\alpha\|\cdot\|_2^2} \circ Proj_{\lambda(1-\alpha)\|\cdot\|_1}\right)(\boldsymbol{x})$$

con $Proj_{\lambda(1-\alpha)\|\cdot\|_1}$ el operador *soft-thresholding:*

$$Proj_{\lambda(1-\alpha)\|\cdot\|_1}(\boldsymbol{x}) := \text{sign}(x_j)(|x_j| - \lambda(1-\alpha))_+$$

y $Proj_{\lambda\alpha\|\cdot\|_2^2}$ el operador escalar solución de la regresión Ridge:

$$Proj_{\mu\|\cdot\|_2^2}(\boldsymbol{x}) := \frac{1}{1+\mu}(\boldsymbol{x})$$

Con todo ello,

$$\boldsymbol{y} = \frac{1}{1+\lambda\alpha} Proj_{\lambda(1-\alpha)\|\cdot\|_1}(\boldsymbol{x}) = \frac{\text{sign}(x_j)(|x_j| - \lambda(1-\alpha))_+}{1+2\lambda\alpha} = \frac{S_{\lambda(1-\alpha)}(\boldsymbol{x})}{1+2\lambda\alpha}$$

que coincide con la solución obtenida anteriormente.

4.2.4 Proyección de un vector sobre el espacio $\mathfrak{B}_{\ell_1+\ell_2}(\tau) \cap \mathfrak{B}_{\ell_2}(1)$

Se muestra a continuación el problema de optimización que permite obtener la proyección de un vector $\boldsymbol{x}$ sobre el espacio $\mathfrak{B}_{\ell_1+\ell_2} \cap \mathfrak{B}_{\ell_2}$.Dado $\boldsymbol{x} \in \mathbb{R}^J$, se define la proyección de un vector $\boldsymbol{y} = P_{\tau,1}^{\mathfrak{B}_{\ell_1+\ell_2} \cap \mathfrak{B}_{\ell_2}}(\boldsymbol{x}) \in \mathbb{R}^J$ con $\gamma > 0$ sobre el espacio restringido de la norma Elastic net como la solución al problema de optimización con restricción:

$$\begin{aligned} \boldsymbol{y} &:= P_{\tau,1}^{\mathfrak{B}_{\ell_1+\ell_2} \cap \mathfrak{B}_{\ell_2}}(\boldsymbol{x}) \\ &= \left\{ \underset{\boldsymbol{y}\in\mathbb{R}^J}{\text{argmin}} \ \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{x}\|_2^2 \, ; s.a: (1-\alpha)\|\boldsymbol{y}\|_1 + \alpha\|\boldsymbol{y}\|_2^2 \right. \\ &\left. \leq \tau, \|\boldsymbol{y}\|_2 \leq 1 \right\} \end{aligned} \tag{4.7}$$

Sea $\tilde{\boldsymbol{x}} = (\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_J)$ el vector compuesto por los valores absolutos de los coeficientes de $\boldsymbol{x}$ ordenados de manera decreciente ($\tilde{x}_1 \geq \tilde{x}_2 \geq \cdots \geq \tilde{x}_J$). Se define la función:

$$\Omega(\lambda) = \frac{(1-\alpha)\left\|S^*_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})\right\|_1 + \alpha\left\|S^*_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})\right\|_2^2}{\left\|S^*_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})\right\|_2} \tag{4.8}$$

con:

$$S^*_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}}) = \frac{S_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})}{1+2\lambda\alpha} \tag{4.9}$$

Por la definición de $S^*_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})$, la función $\Omega(\lambda)$ puede reescribirse como:

$$\Omega(\lambda) = \frac{(1-\alpha)\left\|S^*_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})\right\|_1 + \alpha\left\|S^*_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})\right\|_2^2}{\left\|S^*_{\lambda}(\tilde{\boldsymbol{x}})\right\|_2}$$

$$= \frac{(1-\alpha)\left\|\frac{S_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})}{1+2\lambda\alpha}\right\|_1 + \alpha\left\|\frac{S_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})}{(1+2\lambda\alpha)^2}\right\|_2^2}{\left\|\frac{S_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})}{1+2\lambda\alpha}\right\|_2}$$

$$= \frac{\frac{(1-\alpha)}{1+2\lambda\alpha}\left\|S_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})\right\|_1 + \frac{\alpha}{(1+2\lambda\alpha)^2}\left\|S_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})\right\|_2^2}{\frac{1}{1+2\lambda\alpha}\left\|S_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})\right\|_2}$$

$$= \frac{(1-\alpha)\left\|S_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})\right\|_1 + \frac{\alpha}{(1+2\lambda\alpha)}\left\|S_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})\right\|_2^2}{\left\|S_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})\right\|_2}$$

$$= \frac{(1-\alpha)\left\|S_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})\right\|_1}{\left\|S_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})\right\|_2} + \frac{\alpha}{(1+2\lambda\alpha)}\left\|S_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})\right\|_2$$

$$= (1-\alpha)\psi(\lambda(1-\alpha)) + \frac{\alpha}{(1+2\lambda\alpha)}\left\|S_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})\right\|_2$$

de donde se sigue que para $\alpha = 0$, $\Omega(\lambda) = \psi(\lambda)$; esto es, nos encontraríamos en el caso de la proyección $P_{\tau,1}^{\mathfrak{B}_{\ell_1}\cap\mathfrak{B}_{\ell_2}}$ del vector $\boldsymbol{x}$ sobre el espacio de restricciones $\|\boldsymbol{x}\|_1 \leq \tau$ y $\|\boldsymbol{x}\|_2 \leq 1$.

Al igual que en los casos anteriores, el objetivo es encontrar λ tal que:

$$\Omega(\lambda) = \frac{(1-\alpha)\left\|S^*_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})\right\|_1 + \alpha\left\|S^*_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})\right\|_2^2}{\left\|S^*_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})\right\|_2} = \tau \tag{4.10}$$

con $\lambda \in [\tilde{x}_1, \tilde{x}_J]$ y $\tau \in [\Omega(\tilde{x}_1), \Omega(0)]$. La función Ω cumple las siguientes propiedades:

1- Es continua (por ser composición de funciones continuas)
2- Es una función decreciente. Por definición,

$$\Omega(\lambda) = \frac{(1-\alpha)\left\|S_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})\right\|_1}{\left\|S_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})\right\|_2} + \frac{\alpha}{(1+2\lambda\alpha)}\left\|S_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})\right\|_2$$

Anteriormente se ha visto que:

$$\psi'(\lambda) = \left(\frac{\|S_\lambda(\tilde{\boldsymbol{x}})\|_1}{\|S_\lambda(\tilde{\boldsymbol{x}})\|_2}\right)' = \frac{1}{\|S_\lambda(\tilde{\boldsymbol{x}})\|_2} \cdot \left[\frac{\|S_\lambda(\tilde{\boldsymbol{x}})\|_1^2}{\|S_\lambda(\tilde{\boldsymbol{x}})\|_2^2} - j\right] \leq 0$$

Extendiendo este resultado a la penalización $\lambda(1-\alpha)$:

$$\psi'(\lambda(1-\alpha)) = \left(\frac{\left\|S_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})\right\|_1}{\left\|S_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})\right\|_2}\right)' = \frac{1}{\left\|S_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})\right\|_2} \cdot \left[\frac{\left\|S_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})\right\|_1^2}{\left\|S_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})\right\|_2^2} - j\right] \leq 0$$

Por el Lema 1:

$$\left\|S_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})\right\|_1 \leq \sqrt{j}\left\|S_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})\right\|_2$$

$$\frac{\left\|S_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})\right\|_1^{\,2}}{\left\|S_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})\right\|_2^{\,2}} \leq j$$

Y añadiendo que:

$$\left\|S_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})\right\|_2' = \frac{-\left\|S_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})\right\|_1}{\left\|S_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})\right\|_2}$$

Se tiene:

$$\Omega'(\lambda) = \frac{(1-\alpha)}{\left\|S_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})\right\|_2} \cdot \left[\frac{\left\|S_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})\right\|_1^2}{\left\|S_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})\right\|_2^2} - j\right]$$

$$+ \left(\frac{-2\alpha^2}{(1+2\lambda\alpha)^2}\left\|S_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})\right\|_2 - \frac{\alpha}{(1+2\lambda\alpha)}\frac{\left\|S_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})\right\|_1}{\left\|S_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})\right\|_2}\right)$$

$$= \frac{(1-\alpha)}{\left\|S_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})\right\|_2} \cdot \left[\frac{\left\|S_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})\right\|_1^2}{\left\|S_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})\right\|_2^2} - j\right]$$

$$- \frac{\alpha}{(1+2\lambda\alpha)}\left[\frac{\frac{2\alpha}{(1+2\lambda\alpha)}\left\|S_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})\right\|_2^2 + \left\|S_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})\right\|_1}{\left\|S_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})\right\|_2}\right]$$

$$= \frac{1}{\left\|S_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})\right\|_2}$$

$$\cdot \left[(1-\alpha)\left(\frac{\left\|S_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})\right\|_1^2}{\left\|S_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})\right\|_2^2} - j\right)\right.$$

$$\left. - \frac{\alpha}{(1+2\lambda\alpha)}\left[\frac{2\alpha}{(1+2\lambda\alpha)}\left\|S_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})\right\|_2^2 + \left\|S_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})\right\|_1\right]\right] \leq 0$$

Por ser una función continúa de primera derivada negativa, se concluye que $\Omega(\lambda)$ es una función decreciente, que decrece desde $\Omega(0)$ hasta $\Omega(v)$, con $v \in [\tilde{x}_2, \tilde{x}_1]$[2]:

$$\Omega(0) = \Omega\left(\tilde{x}_{J+1}\right) = \frac{(1-\alpha)\|S_0^*(\tilde{\boldsymbol{x}})\|_1 + \alpha\|S_0^*(\tilde{\boldsymbol{x}})\|_2^2}{\|S_0^*(\tilde{\boldsymbol{x}})\|_2}$$

$$= \frac{(1-\alpha)\left\|\frac{S_0(\tilde{\boldsymbol{x}})}{1+0}\right\|_1 + \alpha\left\|\frac{S_0(\tilde{\boldsymbol{x}})}{(1+0)^2}\right\|_2^2}{\left\|\frac{S_0(\tilde{\boldsymbol{x}})}{1+0}\right\|_2} = \frac{(1-\alpha)\|\tilde{\boldsymbol{x}}\|_1 + \alpha\|\tilde{\boldsymbol{x}}\|_2^2}{\|\tilde{\boldsymbol{x}}\|_2}$$

Y por el Lema 1:

[2] No es posible considerar $v = \tilde{x}_1$ como en casos anteriores puesto que la situación llevaría a una indeterminación, por ser $\|\tilde{\boldsymbol{x}}^*(v)\|_2 = 0$.

$$\Omega(0) = \Omega\left(\tilde{x}_{J+1}\right) \leq \frac{(1-\alpha)\sqrt{J}\|\tilde{\boldsymbol{x}}\|_2 + \alpha\|\tilde{\boldsymbol{x}}\|_2^2}{\|\tilde{\boldsymbol{x}}\|_2} = (1-\alpha)\sqrt{J} + \alpha\|\tilde{\boldsymbol{x}}\|_2$$

Por otra parte,

$$\Omega(\tilde{x}_1) = \frac{(1-\alpha)\left\|S^*_{\tilde{x}_1}(\tilde{\boldsymbol{x}})\right\|_1 + \alpha\left\|S^*_{\tilde{x}_1}(\tilde{\boldsymbol{x}})\right\|_2^2}{\left\|S^*_{\tilde{x}_1}(\tilde{\boldsymbol{x}})\right\|_2} = \frac{(1-\alpha)\|\mathbf{0}\|_1 + \alpha\|\mathbf{0}\|_2^2}{\|\mathbf{0}\|_2} = 0$$

Dado que, por definición, $\tilde{x}_1 > \tilde{x}_2 > \ldots > \tilde{x}_J > \tilde{x}_{J+1} = 0$ y $\Omega\left(\tilde{x}_{J+1}\right) = \Omega(0) > \Omega\left(\tilde{x}_J\right) > \ldots > \Omega(\tilde{x}_2) > \Omega(\tilde{x}_1)$, a partir de las propiedades de $\Omega(\lambda)$ puede deducirse que $\forall\tau \in [\Omega(\tilde{x}_1), \Omega(0)] = [0, (1-\alpha)\sqrt{J} + \alpha\|\tilde{\boldsymbol{x}}\|_2]$ existe $k \in \mathbb{Z}$, $k \leq J$, tal que:

$$\Omega(\tilde{\boldsymbol{x}}_k) \leq \tau < \Omega(\tilde{\boldsymbol{x}}_{k+1}) \tag{4.11}$$

Se busca λ^*, con $\tilde{\boldsymbol{x}}_{k+1} \leq \lambda^* < \tilde{\boldsymbol{x}}_k$, tal que $\Omega(\lambda^*) = \tau$, con $\lambda^* = \lambda(1-\alpha)$ luego:

$$\frac{(1-\alpha)\|S^*_{\lambda*}(\tilde{\boldsymbol{x}})\|_1 + \alpha\|S^*_{\lambda*}(\tilde{\boldsymbol{x}})\|_2^2}{\left\|S^*_{\lambda*}(\tilde{\boldsymbol{x}})\right\|_2} = \tau \tag{4.12}$$

Supóngase k conocido. De manera similar a la definición de $\theta(\lambda)$ vista en el punto 2, se desarrolla esta expresión algebraicamente, donde el objetivo principal es definir la solución λ^*.

$$\left\|S^*_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})\right\|_1 = \sum_{i=1}^{j}\left(\frac{\tilde{x}_i - \lambda(1-\alpha)}{1+2\lambda\alpha}\right) = \frac{1}{1+2\lambda\alpha}\sum_{i=1}^{j}(\tilde{x}_i - \lambda(1-\alpha))$$

$$= \frac{1}{1+2\lambda\alpha}\left[\sum_{i=1}^{j}\tilde{x}_i - \sum_{i=1}^{j}\lambda(1-\alpha)\right]$$

$$\left\|S^*_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})\right\|_2^2 = \sum_{i=1}^{j}\left(\frac{\tilde{x}_i - \lambda(1-\alpha)}{1+2\lambda\alpha}\right)^2 = \frac{1}{(1+2\lambda\alpha)^2}\sum_{i=1}^{j}(\tilde{x}_i - \lambda(1-\alpha))^2$$

$$= \frac{1}{(1+2\lambda\alpha)^2}\sum_{i=1}^{j}(\tilde{x}_i^2 - 2\lambda(1-\alpha)\tilde{x}_i + \lambda^2(1-\alpha)^2)$$

$$= \frac{1}{(1+2\lambda\alpha)^2}\left[\sum_{i=1}^{j}\tilde{x}_i^2 - 2\lambda(1-\alpha)\sum_{i=1}^{j}\tilde{x}_i + \sum_{i=1}^{j}\lambda^2(1-\alpha)^2\right]$$

Por tanto:

$$(1-\alpha)\left\|S^*_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})\right\|_1 + \alpha\left\|S^*_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})\right\|_2^2$$

$$= \frac{1-\alpha}{1+2\lambda\alpha}\left(\sum_{i=1}^{j}\tilde{x}_i - \lambda(1-\alpha)\right)$$

$$+ \frac{\alpha}{(1+2\lambda\alpha)^2}\left(\sum_{i=1}^{j}\tilde{x}_i^2 - 2\lambda(1-\alpha)\sum_{i=1}^{j}\tilde{x}_i + \sum_{i=1}^{j}\lambda^2(1-\alpha)^2\right)$$

$$= \frac{1-\alpha}{1+2\lambda\alpha}\sum_{i=1}^{j}\tilde{x}_i - \frac{1-\alpha}{1+2\lambda\alpha}\sum_{i=1}^{j}\lambda(1-\alpha) + \frac{\alpha}{(1+2\lambda\alpha)^2}\sum_{i=1}^{j}\tilde{x}_i^2$$

$$- \frac{\alpha}{(1+2\lambda\alpha)^2}2\lambda(1-\alpha)\sum_{i=1}^{j}\tilde{x}_i + \frac{\alpha}{(1+2\lambda\alpha)^2}\sum_{i=1}^{j}\lambda^2(1-\alpha)^2$$

$$= \frac{\alpha}{(1+2\lambda\alpha)^2}\sum_{i=1}^{j}\tilde{x}_i^2 + \left[\frac{1-\alpha}{1+2\lambda\alpha} - \frac{\alpha}{(1+2\lambda\alpha)^2}2\lambda(1-\alpha)\right]\sum_{i=1}^{j}\tilde{x}_i$$

$$+ \frac{1}{(1+2\lambda\alpha)^2}\left[\alpha\sum_{i=1}^{j}\lambda^2(1-\alpha)^2 - (1+2\lambda\alpha)(1-\alpha)\sum_{i=1}^{j}\lambda(1-\alpha)\right]$$

$$= \frac{\alpha}{(1+2\lambda\alpha)^2}\sum_{i=1}^{j}\tilde{x}_i^2 + \frac{(1-\alpha)(1+2\lambda\alpha) - 2\lambda\alpha(1-\alpha)}{(1+2\lambda\alpha)^2}\sum_{i=1}^{j}\tilde{x}_i$$

$$+ \frac{1}{(1+2\lambda\alpha)^2}\left[\alpha(1-\alpha)^2\sum_{i=1}^{j}\lambda^2 - (1-\alpha)^2\left(\sum_{i=1}^{j}\lambda + 2\alpha\sum_{i=1}^{j}\lambda^2\right)\right]$$

$$= \frac{\alpha}{(1+2\lambda\alpha)^2}\sum_{i=1}^{j}\tilde{x}_i^2 + \frac{(1-\alpha)}{(1+2\lambda\alpha)^2}\sum_{i=1}^{j}\tilde{x}_i$$

$$- \frac{1}{(1+2\lambda\alpha)^2}\left[\alpha(1-\alpha)^2\sum_{i=1}^{j}\lambda^2 + (1-\alpha)^2\sum_{i=1}^{j}\lambda\right]$$

$$= \frac{1}{(1+2\lambda\alpha)^2}\left[\sum_{i=1}^{j}\alpha\tilde{x}_i^2 + (1-\alpha)\sum_{i=1}^{j}\tilde{x}_i\right.$$

$$\left. - \left(\alpha(1-\alpha)^2\sum_{i=1}^{j}\lambda^2 + (1-\alpha)^2\sum_{i=1}^{j}\lambda\right)\right]$$

$$= \frac{1}{(1+2\lambda\alpha)^2}\sum_{i=1}^{j}(\alpha\tilde{x}_i^2 + (1-\alpha)\tilde{x}_i - (\alpha(1-\alpha)^2)\lambda^2 + (1-\alpha)^2\lambda)$$

$$= \frac{1}{(1+2\lambda\alpha)^2}\sum_{i=1}^{j}\left(\alpha\tilde{x}_i^2 + (1-\alpha)\tilde{x}_i - \lambda(1-\alpha)^2(1+\lambda\alpha)\right)$$

De manera que si:

$$\frac{(1-\alpha)\|S_{\lambda*}^*(\tilde{\boldsymbol{x}})\|_1 + \alpha\|S_{\lambda*}^*(\tilde{\boldsymbol{x}})\|_2^2}{\left\|S_{\lambda*}^*(\tilde{\boldsymbol{x}})\right\|_2} = \tau$$

es lo mismo que:

$$\frac{\frac{1}{(1+2\lambda\alpha)^2}\sum_{i=1}^{k}\left(\alpha\tilde{x}_i^2 + (1-\alpha)\tilde{x}_i - \lambda(1-\alpha)^2(1+\lambda\alpha)\right)}{\frac{1}{1+2\lambda\alpha}\left\|S_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})\right\|_2} = \tau \tag{4.13}$$

$$\frac{\sum_{i=1}^{k}\left(\alpha\tilde{x}_i^2 + (1-\alpha)\tilde{x}_i - \lambda(1-\alpha)^2(1+\lambda\alpha)\right)}{(1+2\lambda\alpha)\left\|S_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})\right\|_2} = \tau$$

$$\frac{\sum_{i=1}^{k}\left(\alpha\tilde{x}_i^2 + (1-\alpha)\tilde{x}_i - \lambda(1-\alpha)^2(1+\lambda\alpha)\right)}{\sqrt{\sum_{i=1}^{k}(\tilde{x}_i^2 - 2\lambda(1-\alpha)\tilde{x}_i + \lambda^2(1-\alpha)^2)}} = \tau(1+2\lambda\alpha)$$

$$\sum_{i=1}^{k}\left(\alpha\tilde{x}_i^2 + (1-\alpha)\tilde{x}_i - \lambda(1-\alpha)^2(1+\lambda\alpha)\right)$$

$$= \tau(1+2\lambda\alpha)\sqrt{\sum_{i=1}^{k}(\tilde{x}_i^2 - 2\lambda(1-\alpha)\tilde{x}_i + \lambda^2(1-\alpha)^2)}$$

Elevando al cuadrado ambas expresiones:

$$\left(\alpha\sum_{i=1}^{k}\tilde{x}_i^2 + (1-\alpha)\sum_{i=1}^{k}\tilde{x}_i - k\left(\lambda(1-\alpha)^2(1+\lambda\alpha)\right)\right)^2$$

$$= \tau^2(1+2\lambda\alpha)^2\left[\sum_{i=1}^{k}\tilde{x}_i^2 - 2\lambda(1-\alpha)\sum_{i=1}^{k}\tilde{x}_i + k\lambda^2(1-\alpha)^2\right]$$

Usando la notación $l_1 = \sum_{i=1}^{k}\tilde{x}_i$ y $l_2 = \sum_{i=1}^{k}\tilde{x}_i{}^2$ tenemos por un lado que:

$$\left(\alpha\sum_{i=1}^{k}\tilde{x}_i^2+(1-\alpha)\sum_{i=1}^{k}\tilde{x}_i-k\big(\lambda(1-\alpha)^2(1+\lambda\alpha)\big)\right)^2$$

$$=\Big(\alpha l_2+(1-\alpha)l_1-k\big(\lambda(1-\alpha)^2(1+\lambda\alpha)\big)\Big)^2$$
$$=\alpha^2l_2^2+2\alpha l_2(1-\alpha)l_1-2k\alpha l_2\big(\lambda(1-\alpha)^2(1+\lambda\alpha)\big)+(1-\alpha)^2l_1^2$$
$$-2k\lambda l_1(1-\alpha)^3(1+\lambda\alpha)+k^2\lambda^2(1-\alpha)^4(1+\lambda\alpha)^2$$
$$=(\alpha^2l_2^2+2l_2l_1\alpha(1-\alpha)+(1-\alpha)^2l_1^2)-2k\alpha l_2(1-\alpha)^2(\lambda+\lambda^2\alpha)$$
$$-2l_1k(1-\alpha)^3(\lambda+\lambda^2\alpha)+k^2\lambda^2(1-\alpha)^4(1+2\lambda\alpha+\lambda^2\alpha^2)$$
$$=(\alpha^2l_2^2+2l_2l_1\alpha(1-\alpha)+(1-\alpha)^2l_1^2)+\lambda\cdot(-2k(1-\alpha)^2\alpha l_2$$
$$-2kl_1(1-\alpha)^3)+\lambda^2\cdot(-2k(1-\alpha)^2\alpha^2l_2-2kl_1(1-\alpha)^3\alpha$$
$$+k^2(1-\alpha)^4)+\lambda^3\cdot(2k^2(1-\alpha)^4\alpha)+\lambda^4\cdot(k^2(1-\alpha)^4\alpha^2)$$
$$=(\alpha^2l_2^2+2l_2l_1\alpha(1-\alpha)+(1-\alpha)^2l_1^2)-\lambda\cdot(2k(1-\alpha)^2\alpha l_2$$
$$+2kl_1(1-\alpha)^3)+\lambda^2\cdot(k^2(1-\alpha)^4-2k(1-\alpha)^2\alpha^2l_2$$
$$-2kl_1(1-\alpha)^3\alpha)+\lambda^3\cdot(2k^2(1-\alpha)^4\alpha)+\lambda^4\cdot(k^2(1-\alpha)^4\alpha^2)$$

Y por otro lado que:

$$\tau^2(1+2\lambda\alpha)^2\left[\sum_{i=1}^{k}\tilde{x}_i^2-2\lambda(1-\alpha)\sum_{i=1}^{k}\tilde{x}_i+k\lambda^2(1-\alpha)^2\right]$$
$$=(\tau^2+4\lambda^2\alpha^2\tau^2+4\lambda\alpha\tau^2)(l_2-2\lambda(1-\alpha)l_1+k\lambda^2(1-\alpha)^2)$$
$$=\tau^2(l_2-2\lambda(1-\alpha)l_1+k\lambda^2(1-\alpha)^2+4\lambda^2\alpha^2l_2-8\lambda^3\alpha^2(1-\alpha)l_1$$
$$+4k\lambda^4\alpha^2(1-\alpha)^2+4\lambda\alpha l_2-8\lambda^2\alpha(1-\alpha)l_1+4k\lambda^3\alpha(1-\alpha)^2)$$
$$=\tau^2(l_2+\lambda(4\alpha l_2-2(1-\alpha)l_1)+\lambda^2(k(1-\alpha)^2+4\alpha^2l_2$$
$$-8\alpha(1-\alpha)l_1)+\lambda^3(4k\alpha(1-\alpha)^2-8\alpha^2(1-\alpha)l_1)$$
$$+\lambda^4(4k\alpha^2(1-\alpha)^2))$$

Unificando ambos resultados, encontrar la solución λ es equivalente a resolver la ecuación de grado 4 siguiente, con una única solución válida por ser $\Omega(\lambda)$ decreciente y continua en el intervalo $[\tilde{x}_1,\tilde{x}_J]$:

$$
\begin{aligned}
&(\tau^2 l_2 - \alpha^2 l_2^2 - 2\alpha l_1 l_2(1-\alpha) - l_1^2(1-\alpha)^2) + \lambda \\
&\quad \cdot (4\alpha l_2 \tau^2 - 2\tau^2 l_1(1-\alpha) + 2k l_2 \alpha(1-\alpha)^2 + 2k l_1(1-\alpha)^3) \\
&\quad + \lambda^2 \\
&\quad \cdot (k\tau^2(1-\alpha)^2 + 4\alpha^2 l_2 \tau^2 - 8\alpha\tau^2(1-\alpha) l_1 \\
&\quad + 2 l_1 k\alpha(1-\alpha)^3 + 2 l_2 k\alpha^2(1-\alpha)^2 - k^2(1-\alpha)^4) + \lambda^3 \\
&\quad \cdot (4k\alpha\tau^2(1-\alpha)^2 - 2k^2\alpha(1-\alpha)^4 - 8\alpha^2\tau^2(1-\alpha) l_1) + \lambda^4 \\
&\quad \cdot (4k\alpha^2\tau^2(1-\alpha)^2 - k^2\alpha^2(1-\alpha)^4) = 0
\end{aligned}
\tag{4.14}
$$

Se deduce de todo este proceso una metodología de cuatro etapas para la proyección de un vector en la intersección del balón $\mathfrak{B}_{\ell_1+\ell_2}$ de radio τ y el balón $\mathfrak{B}_{\ell_2}$ de radio 1:

1- Calcular el vector de los coeficientes de x en valor absoluto.
2- Encontrar k que cumple la ecuación (4).
3- Calcular λ como solución a la ecuación (53).
4- Calcular el vector solución como:

$$\boldsymbol{y} = \frac{S_{\lambda(1-\alpha)}(\tilde{\boldsymbol{x}})}{1 + 2\lambda\alpha}$$

El lector puede encontrar un esquema de la técnica presentada en la Figura 51 y la Tabla 21.

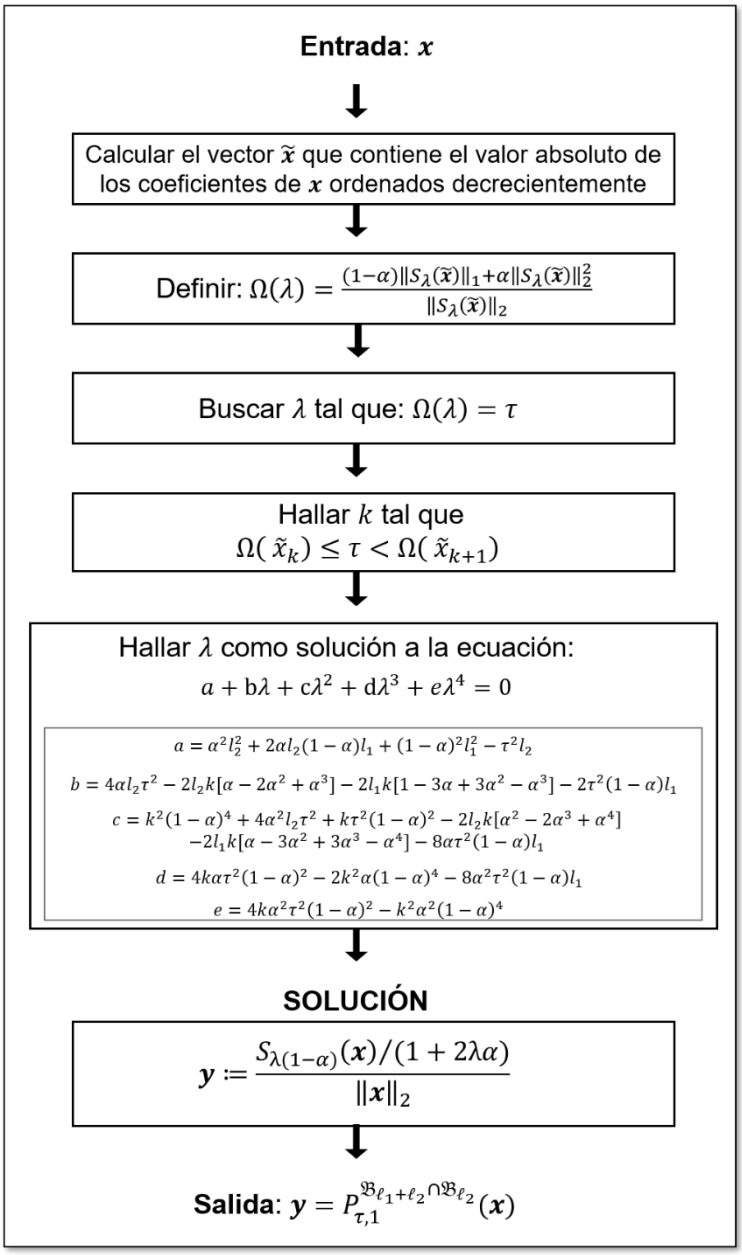


***Figura 51**. Esquema del método aquí propuesto para la proyección sobre la región $\mathfrak{B}_{\ell_1+\ell_2} \cap \mathfrak{B}_{\ell_2}$*

Algoritmo de proyección de un vector sobre el espacio $\mathfrak{B}_{\ell_1+\ell_2}(\tau) \cap \mathfrak{B}_{\ell_2}(1)$ en tiempo lineal.

Siguiendo la metodología utilizada en los métodos anteriores, se propone a continuación el algoritmo de proyección de un vector $\boldsymbol{x}$ sobre el espacio $\mathfrak{B}_{\ell_1+\ell_2}(\tau) \cap \mathfrak{B}_{\ell_2}(1)$ y que se ha implementado en R. El código se muestra al final del capítulo.

***Tabla 21**. Algoritmo de proyección de un vector $\boldsymbol{x}$ sobre la restricción $\mathfrak{B}_{\ell_1+\ell_2} \cap \mathfrak{B}_{\ell_2}$*

Algoritmo CSVD-1: Proyección de $\boldsymbol{v}$ sobre el $\mathfrak{B}_{\ell_1+\ell_2}(\tau) \cap \mathfrak{B}_{\ell_2}(1)$ de radio τ	
Entrada:	$\boldsymbol{x} \in \mathbb{R}^J, \tau \in \mathbb{R}$ con $\tau \in [1, \sqrt{J}]$
Salida	$\boldsymbol{y} = P_\tau^{\mathfrak{B}_{\ell_1+\ell_2} \cap \mathfrak{B}_{\ell_2}}(\boldsymbol{x})$
Inicialización:	$s_1 = 0, s_2 = 0, nb = 0, p = \lvert\boldsymbol{x}^*\rvert$ con $\boldsymbol{x}^* = \{x_i \in \boldsymbol{x}, i = 1, \dots, J / x_i \neq 0\}$

Si $\|\boldsymbol{x}\|_2 = 0$ **entonces devuelve** $\boldsymbol{x} = \boldsymbol{v}$

Si $\|\boldsymbol{x}\|_1 / \|\boldsymbol{x}\|_2 \leq \tau$ **entonces devuelve** $\boldsymbol{y} = \boldsymbol{x}$ #Ya cumple la restricción

Si no

Mientras TRUE hacer:

$N = \text{length}(p)$

Seleccionar $r \in \{1, \dots, N\}$ aleatorio con $\tilde{x}_r \neq \max(p)$

$\tilde{x}_k = p[r]$

#Partición de p en dos conjuntos disjuntos:

$H = \{p_j \text{ tal que } p_j < \tilde{x}_k\}$

$L = \{p_j \text{ tal que } p_j > \tilde{x}_k\}$

#Evaluación de k

$nb_{\tilde{x}_k} = \text{length}(\{p_j \text{ tal que } p_j == \tilde{x}_k\})$

$K = nb + \text{length}(L) + nb_{\tilde{x}_k}$

$$s_{low_1} = \sum_{i \in L} p_i + nb_{\tilde{x}_k} \cdot \tilde{x}_k$$

$$s_{low_2} = \sum_{i \in L} {p_i}^2 + nb_{\tilde{x}_k} \cdot {\tilde{x}_k}^2$$

$$\Omega(\tilde{x}_k) = \frac{\alpha(s_2 + s_{low_2}) + (1-\alpha)(s_1 + s_{low_1}) - k\tilde{x}_k(1-\alpha)^2 - k{\tilde{x}_k}^2\alpha(1-\alpha)^2}{(1+2\alpha\tilde{x}_k)\sqrt{(s_2 + s_{low_2}) - 2\tilde{x}_k(1-\alpha)(s_1 + s_{low_1}) + k{\tilde{x}_k}^2(1-\alpha)^2}}$$

Si $\Omega(\tilde{x}_k) > \tau$ **entonces**

Si $L = \emptyset$ **entonces salir**

Actualización de p: $p = \{p_i\}_{i \in L}$

Si no

$\tilde{x}_{k+1} = \max(\{p_i\}_{i \in H})$

$$\Omega(\tilde{x}_{k+1}) = \frac{\alpha(s_2 + s_{low_2}) + (1-\alpha)(s_1 + s_{low_1}) - k\tilde{x}_{k+1}(1-\alpha)^2 - k{\tilde{x}_{k+1}}^2\alpha(1-\alpha)^2}{(1+2\alpha\tilde{x}_k)\sqrt{(s_2 + s_{low_2}) - 2\tilde{x}_{k+1}(1-\alpha)(s_1 + s_{low_1}) + k{\tilde{x}_{k+1}}^2(1-\alpha)^2}}$$

Si $\Omega(\tilde{x}_{k+1}) > \tau$ **entonces salir**

Actualización de p: $p = \{p_i\}_{i \in H}$

Actualización de nb: $nb = k$

23: Actualización de s_1: $s_1 = s_1 + s_{low_1}$

24: Actualización de s_2: $s_2 = s_2 + s_{low_2}$

25: **Fin Si**

26: **Fin Mientras**

27: $l_1 = s_1 + s_{low_1}$

$l_2 = s_2 + s_{low_2}$

28: #Encontrar la penalización λ resolviendo la ecuación:

$$(\alpha^2 l_2^2 + 2\alpha l_2(1-\alpha)l_1 + (1-\alpha)^2 l_1^2 - \tau^2 l_2) + \lambda \cdot (4\alpha l_2 \tau^2 - 2l_2 k[\alpha - 2\alpha^2 + \alpha^3] - 2l_1 k[1 - 3\alpha + 3\alpha^2 - \alpha^3] - 2\tau^2(1-\alpha)l_1) + \lambda^2 \cdot (k^2(1-\alpha)^4 + 4\alpha^2 l_2 \tau^2 + k\tau^2(1-\alpha)^2 - 2l_2 k[\alpha^2 - 2\alpha^3 + \alpha^4] - 2l_1 k[\alpha - 3\alpha^2 + 3\alpha^3 - \alpha^4] - 8\alpha\tau^2(1-\alpha)l_1) + \lambda^3 \cdot (4k\alpha\tau^2(1-\alpha)^2 - 2k^2\alpha(1-\alpha)^4 - 8\alpha^2\tau^2(1-\alpha)l_1) + \lambda^4 \cdot (4k\alpha^2\tau^2(1-\alpha)^2 - k^2\alpha^2(1-\alpha)^4) = 0$$

29: # Solución: $y = P_{\tau,1}^{\mathfrak{B}_{\ell_1+\ell_2} \cap \mathfrak{B}_{\ell_2}}(\boldsymbol{x})$

$$\boldsymbol{y} \coloneqq \frac{S_\lambda^*(\boldsymbol{x})}{\|\boldsymbol{x}\|_2} = \frac{\frac{S_\lambda(\boldsymbol{x})}{1+\lambda\gamma}}{\|\boldsymbol{x}\|_2} = \frac{\frac{\text{sign}(\boldsymbol{x})(|\boldsymbol{x}| - \lambda(1-\alpha), 0)_+}{1+2\lambda\alpha}}{\|\boldsymbol{x}\|_2}$$

30: **Fin Si**

4.2.5 Formulación C_{enet}SVD

En consonancia con la SVD clásica, C_{enet}SVD descompone $\boldsymbol{X} \in \mathbb{R}^{IxJ}$ en función del producto de sus vectores pseudo-singulares y sus valores singulares. El punto clave de la C_{enet}SVD reside en la extracción de vectores sparse y ortogonales al mismo tiempo, incorporando restricciones convexas en los vectores singulares. La formulación general matemática de la C_{enet}SVD está basada en el problema de optimización restringida de la CSVD propuesto por Guillemot et al. (2019), con C_1 y C_2 funciones de penalización convexas definidas en $\mathbb{R}^I$ (si la penalización se produce sobre los vectores singulares a izquierda) o $\mathbb{R}^J$ (sobre los vectores singulares a derecha), respectivamente.

$$\underset{d,u,v}{\text{argmin}} \frac{1}{2}\left\|\boldsymbol{X} - d_{\text{q}} u_{\text{q}} {v_{\text{q}}}^T\right\|_F^2$$
$$\text{s.a} \begin{cases} {u_{\text{q}}}^T u_{\text{q}} = {v_{\text{q}}}^T v_{\text{q}} = 1, {u_{\text{q}}}^T u_{\text{q}'} = {v_{\text{q}}}^T v_{\text{q}'} = 0 \quad \forall q \neq q' \\ C_1(u_{\text{q}}) \leq \tau_{1,q} \; ; \; C_2(v_{\text{q}}) \leq \tau_{2,q} \end{cases} \tag{4.15}$$

En este trabajo se propone proyectar los vectores pseudo-singulares sobre la restricción convexa Elastic net (enet) (Figura 52). La norma L1+L2 o balón Elastic net se obtiene como una combinación de las normas L1 y L2:

$$(\mathbf{1}-\boldsymbol{\alpha})\|\boldsymbol{x}\|_{\mathbf{1}}+\boldsymbol{\alpha}\|\boldsymbol{x}\|_{\mathbf{2}}^{\mathbf{2}}\leq\boldsymbol{\tau}$$

Para $\alpha = 0$ la restricción enet penaliza la norma L1 de los vectores, convirtiéndose en la norma Lasso, mientras que para $\alpha = 1$, Elastic net penaliza únicamente la norma L2 del vector x. Como se observa en la Figura 52, a mayor valor de α la región de restricción será más semejante al círculo de penalización L2. En caso contrario, para valores pequeños de α la región de restricción de la norma L1+L2 será prácticamente la misma que la región de regularización Lasso.

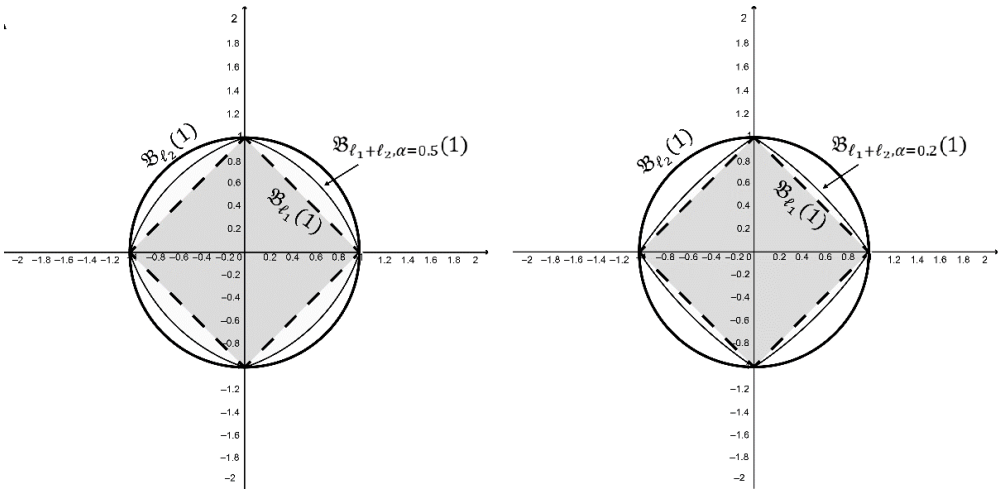


***Figura 52**. Representación gráfica en $\mathbb{R}^2$ de las restricciones de las normas ℓ_1, ℓ_2 y $\ell_1+\ell_2$ en un vector x para su proyección en la bola $\mathfrak{B}_{\ell_1+\ell_2}(\tau)\cap\mathfrak{B}_{\ell_2}(1)$ del procedimiento C_{enet}SVD. La restricción ℓ_2 restringe $\boldsymbol{x}$ tal que $\|\boldsymbol{x}\|_{\mathbf{2}}^{\mathbf{2}}\leq 1$ y la región ℓ_1 a $\|\boldsymbol{x}\|_{\mathbf{1}}\leq\tau$. En el caso del balón de restricción Elastic net, $\boldsymbol{x}$ debe verificar que $(1-\alpha)\|\boldsymbol{x}\|_{\mathbf{1}}+\alpha\|\boldsymbol{x}\|_{\mathbf{2}}^{\mathbf{2}}\leq\tau$. El panel A muestra los balones de restricción $\mathfrak{B}_{\ell_2}(1),\mathfrak{B}_{\ell_1}(1)$ y $\mathfrak{B}_{\ell_1+\ell_2}(1)$ para diferentes valores de $\boldsymbol{\alpha}$ en el balón enet (izquierda, $\alpha = 0.5$; derecha, $\alpha = 0.2$). Se observa como a mayores valores de α el balón de restricción de Elastic net es más parecido a la restricción $\mathfrak{B}_{\ell_2}$.*

Incluyendo la restricción de Elastic net en el problema de optimización (4.15) la función a minimizar que resuelve la C_{enet}SVD (Figura 53) es:

$$\underset{d,u,v}{\operatorname{argmin}} \frac{1}{2}\left\|\boldsymbol{X} - d_{\mathrm{q}}\boldsymbol{u}_{\mathrm{q}}\boldsymbol{v}_{\mathrm{q}}{}^{T}\right\|_{F}^{2}$$

$$\text{s.a} \tag{4.16}$$

$$\begin{cases} \boldsymbol{u}_{\mathrm{q}}{}^{T}\boldsymbol{u}_{\mathrm{q}} = \boldsymbol{v}_{\mathrm{q}}{}^{T}\boldsymbol{v}_{\mathrm{q}} = 1, \boldsymbol{u}_{\mathrm{q}}{}^{T}\boldsymbol{u}_{\mathrm{q}'} = \boldsymbol{v}_{\mathrm{q}}{}^{T}\boldsymbol{v}_{\mathrm{q}'} = 0 \quad \forall q \neq q' \\ (1-\alpha)\left\|\boldsymbol{u}_{\mathrm{q}}\right\|_{1} + \alpha\left\|\boldsymbol{u}_{\mathrm{q}}\right\|_{2}^{2} \leq \tau_{1,q}\ ;\ (1-\alpha)\left\|\boldsymbol{v}_{\mathrm{q}}\right\|_{1} + \alpha\left\|\boldsymbol{v}_{\mathrm{q}}\right\|_{2}^{2} \leq \tau_{2,q} \end{cases}$$

donde $\tau_{1,q}, \tau_{2,q} > 0$ son los parámetros de regularización que controlan el grado de sparsity incluida en el modelo de optimización restringido. Cuando mayor sera τ, menor será el número de coeficientes sparse. Es importante remarcar que tan solo algunos valores de $\tau_{1,q}$ y $\tau_{2,q}$ llevarán a la obtención de soluciones factibles, aunque este punto se tratará más adelante.

Para encontrar la solución de (4.16) es necesario enunciar el problema de minimización de forma equivalente. La ecuación (4.16) se puede reescribir como un problema de maximización en 1 dimensión, para cada uno de los pares de vectores $\boldsymbol{u}$ y $\boldsymbol{v}$ a calcular. Para $q \geq 1$ y dados los vectores previos $\boldsymbol{v}_{\mathrm{q}'}$ y $\boldsymbol{u}_{\mathrm{q}'}$, con $0 \leq q' < q$:

$$\underset{\boldsymbol{u},\boldsymbol{v}}{\operatorname{argmax}}\ \boldsymbol{u}^{T}\boldsymbol{X}\boldsymbol{v}$$

$$\text{s.a.} \begin{cases} \boldsymbol{u}^{T}\boldsymbol{u} \leq 1, \boldsymbol{v}^{T}\boldsymbol{v} \leq 1, \boldsymbol{u}^{T}\boldsymbol{u}_{\mathrm{q}'} = \boldsymbol{v}^{T}\boldsymbol{v}_{\mathrm{q}'} = 0 \quad \forall q' < q \\ (1-\alpha_1)\|\boldsymbol{u}\|_1 + \alpha_1\|\boldsymbol{u}\|_2^2 \leq \tau_{1,q}\ ;\ (1-\alpha_2)\|\boldsymbol{v}\|_1 + \alpha_2\|\boldsymbol{v}\|_2^2 \leq \tau_{2,q} \end{cases} \tag{4.17}$$

Esta ecuación es resuelta mediante un proceso iterativo de dos pasos diferentes, basado en un algoritmo ALS:

1) Para $\boldsymbol{v}$ fijo, el primer punto es encontrar el vector solución $\boldsymbol{u}$ que optimiza la siguiente función:

$$\underset{u}{\operatorname{argmin}} \frac{1}{2}\|\boldsymbol{u} - \boldsymbol{X}\boldsymbol{v}\|_F^2$$

$$\text{s.a. } \{\boldsymbol{u} \in \mathfrak{B}_{\ell_1+\ell_2}(\tau), \boldsymbol{u} \in \mathfrak{B}_{\ell_2}(1), \boldsymbol{u} \in \boldsymbol{U}^{\perp} \leftrightarrow \boldsymbol{u} \in \mathfrak{B}_{\ell_1+\ell_2}(\tau) \cap \mathfrak{B}_{\ell_2}(1) \tag{4.18}$$

con $\boldsymbol{U}^{\perp}$ el complemento ortogonal al espacio definido por las columnas de la matriz $\boldsymbol{U}$**.** Las restricciones añadidas al problema de optimización

pueden verse justamente como la proyección de un vector en un espacio convexo. Además**,** en este caso se trabaja sobre la proyección de un mismo vector sobre dos conjuntos convexos (la norma L2 y la norma L1+L2); es decir, el vector debe pertenecer a la intersección de ambos espacios. Ahora bien, la intersección de dos espacios convexos es también convexa así que pueden utilizarse algoritmos de proyección en espacios convexos para encontrar el vector solución. Para proyectar sobre el espacio $\mathfrak{B}_{\ell_1+\ell_2}(\tau) \cap \mathfrak{B}_{\ell_2}(1)$ se extiende el algoritmo PL1L2 que sigue la ideología POCS para la proyección de un vector en $\mathfrak{B}_{\ell_1}(\tau) \cap \mathfrak{B}_{\ell_2}(1)$ (Gloaguen et al., 2017; Guillemot et al., 2019). Siguiendo dicha metodología, la solución de la proyección buscada en $\mathfrak{B}_{\ell_1+\ell_2}(\tau) \cap \mathfrak{B}_{\ell_2}(1)$ es obtenida mediante la composición de proyecciones en cada una de las restricciones convexas establecidas. En nuestro caso, se proyectará por un lado en $\mathfrak{B}_{\ell_1+\ell_2}(\tau)$ y por otro en $\mathfrak{B}_{\ell_2}(1)$. La proyección de un vector en el espacio $\mathfrak{B}_{\ell_1+\ell_2}$ que proponemos aquí sigue la línea de los algoritmos de proyección en la bola Lasso $\mathfrak{B}_{\ell_1}$ en tiempo lineal como el PL1L2 de (Guillemot et al., 2019) basados en (Berg, Schmidt, Friedlander, & Murphy, 2008; Duchi, Shalev-Shwartz, Singer, & Chandra, 2008) y en la línea también de la formulación propuesta por Mairal et al. (2010) para la proyección en la bola Elastic net $\mathfrak{B}_{\ell_1+\ell_2}$. Una vez encontrado el vector $\boldsymbol{u}$ con $\boldsymbol{v}$ fijo, se busca ahora encontrar la solución $\boldsymbol{v}$ con $\boldsymbol{u}$ fijo:

$$\underset{u}{\operatorname{argmin}} \frac{1}{2}\|\boldsymbol{v} - \boldsymbol{X^T u}\|_F^2$$
$$\text{s.a. } \{\boldsymbol{v} \in \mathfrak{B}_{\ell_1+\ell_2}(\tau),\ \boldsymbol{v} \in \mathfrak{B}_{\ell_2}(1), \boldsymbol{v} \in \boldsymbol{V}^{\perp} \leftrightarrow\ \boldsymbol{v} \in \mathfrak{B}_{\ell_1+\ell_2}(\tau) \cap \mathfrak{B}_{\ell_2}(1) \tag{4.19}$$

con $\boldsymbol{V}^{\perp}$ el complemento ortogonal al espacio definido por las columnas de la matriz $\boldsymbol{V}$**.** La proyección $\boldsymbol{v}_{t+1} = proj^{\mathfrak{B}_{\ell_1+\ell_2}(\boldsymbol{\tau}) \cap \mathfrak{B}_{\ell_2}(1) \cap \boldsymbol{V}^{\perp}}(\boldsymbol{X}^T u_{t+1})$ se obtiene de la misma forma que se hacía en el punto anterior. El vector solución $\boldsymbol{v}$ vendrá dado por la proyección del vector $\boldsymbol{X}^T u_{t+1}$ en el espacio intersección convexo definido por las restricciones anteriores.

El pseudo-código que resuelve el problema de optimización global de la CenetSVD está descrito en la Tabla 22.

Tabla 22. *Algoritmo para la implementación de C_{enet}SVD basado en el algoritmo POCS*

Algoritmo: POCS - C_{enet}SVD. Proyección de un vector sobre el espacio $\mathfrak{B}_{\ell_1+\ell_2}(\tau) \cap \mathfrak{B}_{\ell_2}(1) \cap \boldsymbol{M}^{\perp}$	
Entrada:	$\boldsymbol{X} \in \mathbb{R}^{IxJ}$, rango Q, $\varepsilon \approx 0$, $\tau \in [1, (1-\alpha)\sqrt{J} + \alpha]$
Salida:	$\boldsymbol{U} \in \mathbb{R}^{IxQ}, \boldsymbol{D} \in \mathbb{R}^{QxQ}, \boldsymbol{V} \in \mathbb{R}^{JxQ}$
Inicialización:	$\boldsymbol{U} = \boldsymbol{0}, \boldsymbol{V} = \boldsymbol{0}, \lambda_0 = 0$
1: **Para** q en $1{:}\,q$ hacer:	
2: Inicializar aleatoriamente: $\boldsymbol{u}_0, \boldsymbol{v}_0$	
3: $\lambda_1 = \boldsymbol{u}_0^T \boldsymbol{X} \boldsymbol{v}_0$	
$t = 0$	
Mientras $\lvert\lambda_{t+1} - \lambda_t\rvert \geq \varepsilon$ hacer:	
$\boldsymbol{u}_{t+1} = proj^{\mathfrak{B}_{\ell_1+\ell_2}(\boldsymbol{\tau}) \cap \mathfrak{B}_{\ell_2}(1) \cap \boldsymbol{U}^{\perp}}(\boldsymbol{X} v_t)$	
$\boldsymbol{v}_{t+1} = proj^{\mathfrak{B}_{\ell_1+\ell_2}(\boldsymbol{\tau}) \cap \mathfrak{B}_{\ell_2}(1) \cap \boldsymbol{V}^{\perp}}(\boldsymbol{X}^T u_{t+1})$	
$\lambda_{t+1} = \boldsymbol{u}_{t+1}^T \boldsymbol{X} \boldsymbol{v}_{t+1}$	
$t = t + 1$	
4: $\boldsymbol{d} = [\boldsymbol{d}, \lambda_{t+1}]$	
5: $\boldsymbol{U} = [\boldsymbol{U}, \boldsymbol{u}_{t+1}]$	
6: $\boldsymbol{V} = [\boldsymbol{V}, \boldsymbol{v}_{t+1}]$	
7: **Fin**	
8: $\boldsymbol{D} = \mathrm{diag}(d)$	
9: **Fin**	

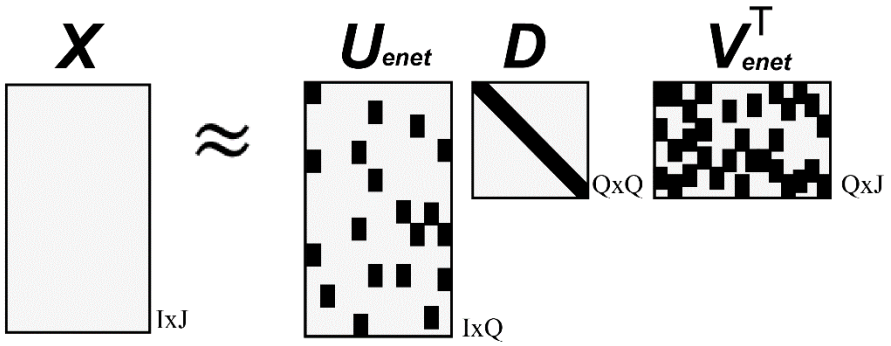


Figura 53. *Esquema de la C_{enet}SVD*

4.2.6 Posibles valores del parámetro de regularización τ. Una interpretación geométrica basada en Lasso

Como declaran Guillemot et al. (2019) y Witten, Tibshirani y Hastie (2009) para la penalización Lasso, tan solo algunos valores de restricción son posibles para alcanzar una solución válida del problema. A continuación expresamos las cotas límites para el parámetro τ en CenetSVD, siguiendo la teoría enunciada por los grupos de autores recién mencionados.

El parámetro de regularización τ debe ser un valor perteneciente al intervalo $[1, (1-\alpha)\sqrt{J}+\alpha]$ para que el problema de optimización convexa restringido a la intersección de la bola Elastic net y la norma L2 de radio 1 sea factible. La explicación de este hecho mediante su interpretación geométrica se explica a continuación. Para ello, haremos uso del lema enunciado en (Guillemot et al., 2019):

Como consecuencia del Lema 1 es fácil observar que:

$$(1-\alpha)\|\boldsymbol{x}\|_2+\alpha\|\boldsymbol{x}\|_2^2 \leq (1-\alpha)\|\boldsymbol{x}\|_1+\alpha\|\boldsymbol{x}\|_2^2 \leq (1-\alpha)\sqrt{J}\|\boldsymbol{x}\|_2+\alpha\|\boldsymbol{x}\|_2^2$$

$$\frac{(1-\alpha)\|\boldsymbol{x}\|_2+\alpha\|\boldsymbol{x}\|_2^2}{\|\boldsymbol{x}\|_2} \leq \frac{(1-\alpha)\|\boldsymbol{x}\|_1+\alpha\|\boldsymbol{x}\|_2^2}{\|\boldsymbol{x}\|_2} \leq \frac{(1-\alpha)\sqrt{J}\|\boldsymbol{x}\|_2+\alpha\|\boldsymbol{x}\|_2^2}{\|\boldsymbol{x}\|_2}$$

$$(1-\alpha)+\alpha\|\boldsymbol{x}\|_2 \leq \tau \leq (1-\alpha)\sqrt{J}+\alpha\|\boldsymbol{x}\|_2$$

Por el enfoque planteado en el problema de optimización restringido, la solución $\boldsymbol{x} \in \mathfrak{B}_{\ell_2}(1)$ y por tanto $\|\boldsymbol{x}\|_2 = 1$. Así:

$$(1-\alpha)+\alpha \leq \tau \leq (1-\alpha)\sqrt{J}+\alpha$$

Como se observa en la Figura 54, se restringe el parámetro τ a tomar valores tales que $1 \leq \tau \leq (1-\alpha)\sqrt{J}+\alpha$ para cualquier $\alpha \in [0,1]$. Obviamente, si $\alpha = 1$ entonces $\tau = 1$ y la intersección de los espacios lleva al conjunto $\mathfrak{B}_{\ell_2}(1)$. Si $\alpha = 0$ el parámetro de regularización τ se restringe a $1 \leq \tau \leq \sqrt{J}$, como se evidencia en (Guillemot et al., 2019; Witten et al., 2009).

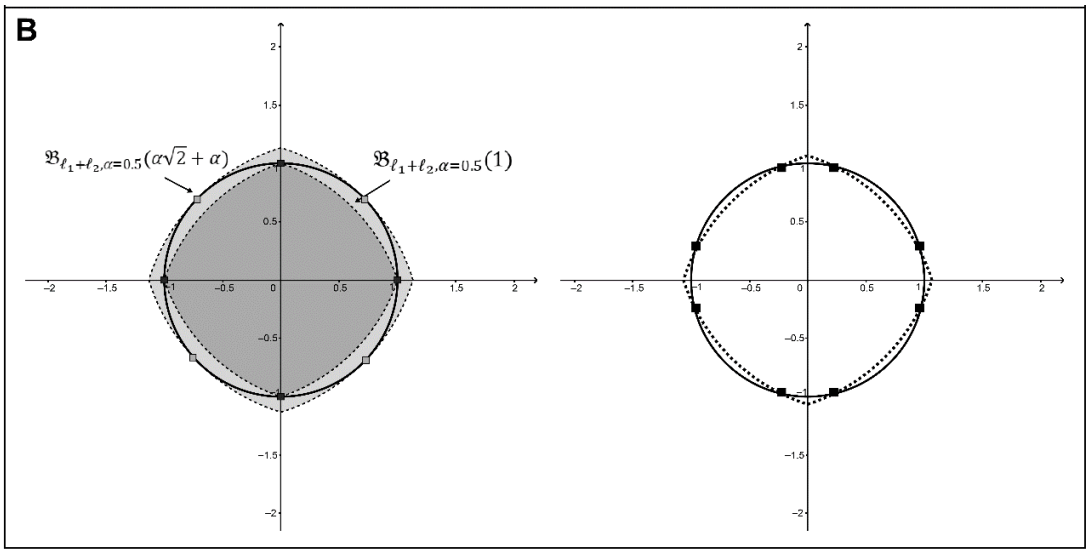


Figura 54. *Representación de restricciones* $(1-\alpha)\|\boldsymbol{x}\|_1+\alpha\|\boldsymbol{x}\|_2^2 \leq 1$ *y* $(1-\alpha)\|\boldsymbol{x}\|_1+\alpha\|\boldsymbol{x}\|_2^2 \leq (0{,}5\sqrt{2}+0{,}5)$ *usando líneas discontinuas y la restricción* $\|\boldsymbol{x}\|_2^2 \leq 1$ *mediante un círculo sólido en* $\mathbb{R}^2$*. Los rectángulos muestran los posibles puntos solución en la intersección de ambas restricciones. En el panel de la derecha se muestran las soluciones factibles para* $1 \leq \tau \leq 0{,}5\sqrt{J}+0{,}5$ *(*$J=2$*), estando simultáneamente activos los balones de restricción Ridge y Elastic net.*

4.2.7 Selección de los parámetros α y τ

La selección de los parámetros α y τ puede realizarse de manera manual. Es habitual considerar $\alpha = 0{,}5$, para que la penalización Elastic net esté compuesta por las restricciones Lasso y Ridge a partes iguales. Sin embargo, en caso de que el usuario no desee estimar estos parámetros manualmente, dicha estimación puede realizarse de manera automática en *pos* del modelo óptimo. Por otro lado, para una matriz $X \in \mathbb{R}^{IxJ}$, $\tau \in [1, (1-\alpha)\sqrt{J}+\alpha]$ para que el problema de optimización de proyección de un vector en el espacio $\mathfrak{B}_{\ell_1+\ell_2}(\tau) \cap \mathfrak{B}_{\ell_2}(1)$ presente una solución factible. En la literatura pueden encontrarse aplicaciones en las que la medida de regularización τ es seleccionada acorde al nivel de restricción de las cargas que se desea en el modelo (Guillemot et al., 2019). En otras palabras, se establecen niveles bajos, medios y altos de *sparsity* o restricción. Además, es importante recordar que cuanto menor sea el coeficiente τ, más coeficientes serán contraídos hacia 0.

A continuación se presenta el proceso de selección de los parámetros τ, de regularización de Elastic net, y α, de control de la cantidad de restricción Lasso y Ridge introducidos en el modelo general. Los valores de α y τ se encuentran

relacionados mutuamente; esto es, el valor de uno influye en el valor del otro y viceversa. Este es el motivo por el que el procedimiento de selección de los parámetros está constituido por dos etapas diferentes. En primer lugar, la selección de α se realizará mediante validación cruzada y, en segundo lugar, una vez fijado α, la selección de τ, y en consecuencia de un modelo óptimo, se llevará a cabo mediante la minimización de un criterio de información. La literatura recoge un amplio conjunto de medidas de información (Zhang, Li, & Tsai, 2010), como pueden ser el criterio de información de Akaike AIC (Akaike, 1974), el criterio de información bayesiano BIC (Kass & Raftery, 2012),... En este trabajo, se hará uso del criterio de información BIC para escoger el parámetro de regularización, decisión motivada por resultados favorables acerca de su capacidad de identificar los modelos verdaderos de manera consistente (Wang, Li, & Tsai, 2007).

Selección de α y $(1-\alpha)$

El procedimiento de elección del parámetro α está basado en un proceso iterativo de búsqueda del parámetro que minimice el error cuadrático medio introducido en la reconstrucción de la matriz original a partir de los valores y vectores singulares obtenidos mediante C_{enet}SVD mediante validación cruzada (James, Witten, Hastie, & Tibshirani, 2014).

La validación cruzada es un método de remuestreo sin reemplazamiento utilizado para evitar el sobreajuste de los parámetros de un determinado modelo, cuya popularidad se debe a la sencillez de su implementación y a la capacidad de estimar parámetros con un menor sesgo que los que se obtendrían a partir de la muestra de partida. Parte de una división inicial aleatoria de la matriz de datos original en lo que se conoce como n-pliegues (*n-folds*), de manera que $n-1$ pliegues son seleccionados como la matriz de entrenamiento (*training*) y el pliegue restante es seleccionado como un conjunto de datos de prueba (test). Así, el método es ajustado sobre los datos de entrenamiento para obtener la estimación de los parámetros necesaria y, posteriormente, dichos parámetros son validados sobre el conjunto de datos de prueba. Para ello, es habitual utilizar como medida de validación el Error Cuadrático Medio (*MSE*, por sus siglas en inglés). Para cada uno de los pliegues se define la matriz de entrenamiento como

aquella formada por las observaciones de las submatrices formadas en los $(n-1)$ pliegues, mientras que las observaciones restantes conforman la matriz de prueba. Una vez calculados los parámetros del modelo en la matriz de entrenamiento, el error de aproximación se valora midiendo el error cometido en la estimación de la matriz de prueba. Este proceso iterativo se repite n veces, de manera que cada uno de los pliegues sea seleccionado una vez como conjunto de datos de entrenamiento. Finalmente, dado $\hat{y}_i$ el vector aproximado de un vector cualquiera y_i y n el número de pliegues definido en la validación cruzada, el MSE se estima como el error promedio de los errores obtenidos en la aproximación de y_i en cada uno de los pliegues:

$$MSE = \frac{1}{n}\sum_{i=1}^{n} MSE_i = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Por otro lado, el número n de pliegues debe ser seleccionado de forma cautelosa, pues una inadecuada selección puede generar interpretaciones erróneas con respecto a la capacidad del modelo debido a un posible aumento del sesgo y la varianza de los parámetros estimados. Esto es, existe una relación balanceada entre la elección del número n de pliegues y la compensación del sesgo-varianza. Además, debe tenerse en cuenta que un número muy bajo de pliegues puede generar muestras de entrenamiento y de prueba que no sean estadísticamente representativas de la muestra de datos original. Aunque no existe una regla formal, habitualmente el número de pliegues se estipula como $n = 5$ o $n = 10$, pues a medida que n aumenta, la diferencia entre la matriz de entrenamiento y prueba se hace más pequeña, generando así sesgos cada vez más bajos y no existiendo variaciones muy altas (James et al., 2014; Kuhn, & Johnson, 2013).

El esquema del modelo definido para la selección de α puede encontrarse en la Figura 55. Como se ha mencionado anteriormente, el parámetro α permite balancear la cantidad de restricción Lasso/Ridge introducida en el modelo Elastic net. De esta manera, cuando $\alpha = 0$ el problema de optimización de Elastic net se convierte en un problema de optimización de Lasso, cuya solución viene dada por el operador *soft-thresholding*, mientras que si $\alpha = 1$ el problema de optimización queda reducido a un problema de restricción Ridge, cuya solución

viene dada por el reescalado de la solución tradicional al problema de mínimos cuadrados ordinarios para la estimación de un cierto vector. Por este motivo, el objetivo en este punto es encontrar el valor de $\alpha \in [0,1)$. De manera predefinida, se genera una secuencia de valores para el parámetro $\alpha = (\alpha_1, \ldots, \alpha_{10}) = (0{,}0,\ 0{,}1,\ 0{,}2,\ 0{,}3,\ 0{,}4,\ 0{,}5,\ 0{,}6,\ 0{,}7,\ 0{,}8,\ 0{,}9)$. El usuario podría escoger modificar esta secuencia, pero siempre proporcionando valores lógicos para el parámetro, atendiendo a su definición. Para cada posible valor del parámetro α, el algoritmo implementará un proceso de validación cruzada, segmentando inicialmente el conjunto de datos en $n = 10$ pliegues por defecto que serán utilizados después en el cálculo del error cuadrático medio obtenido en el modelo para cada valor de α.

Para $\alpha_1 = 0$, se valorará sobre cada uno de los 10 pliegues que conforman las diferentes matrices de entrenamiento y prueba, el error cuadrático medio obtenido en el siguiente sentido. Se calcula la CenetSVD de la matriz de entrenamiento (training) , almacenando la matriz V de cargas sparse obtenida en el proceso. Dicha matriz V es utilizada posteriormente para estimar el error de reconstrucción $MSE_{\alpha_1,1}$ de la matriz de prueba (test) a partir del conjunto de vectores singulares sparse obtenidos previamente:

$$MSE_{\alpha_1,1} = \left\| X_{TEST} - \hat{X}_{TEST} \right\|^2 = \| X_{TEST} - X_{TEST} V V^T \|^2$$

Los errores de reconstrucción $MSE_{\alpha_1,1}, \ldots, MSE_{\alpha_1,n}$ son obtenidos para cada uno de los n plieges, de manera que el MSE final para el primer parámetro α_1 mediante el proceso de validación cruzada para el primer parámetro es estimado como el promedio de los errores $MSE_{\alpha_1,1:n}$:

$$MSE_1 = \frac{1}{n} \sum_{i=1}^{10} MSE_{\alpha_i,1}$$

Siguiendo el mismo procedimiento se calculan los MSE para el resto de los valores de la secuencia del parámetro α predefinida de entrada. El parámetro α óptimo seleccionado para el modelo será aquel que proporcione el mínimo MSE (véase Figura 55).

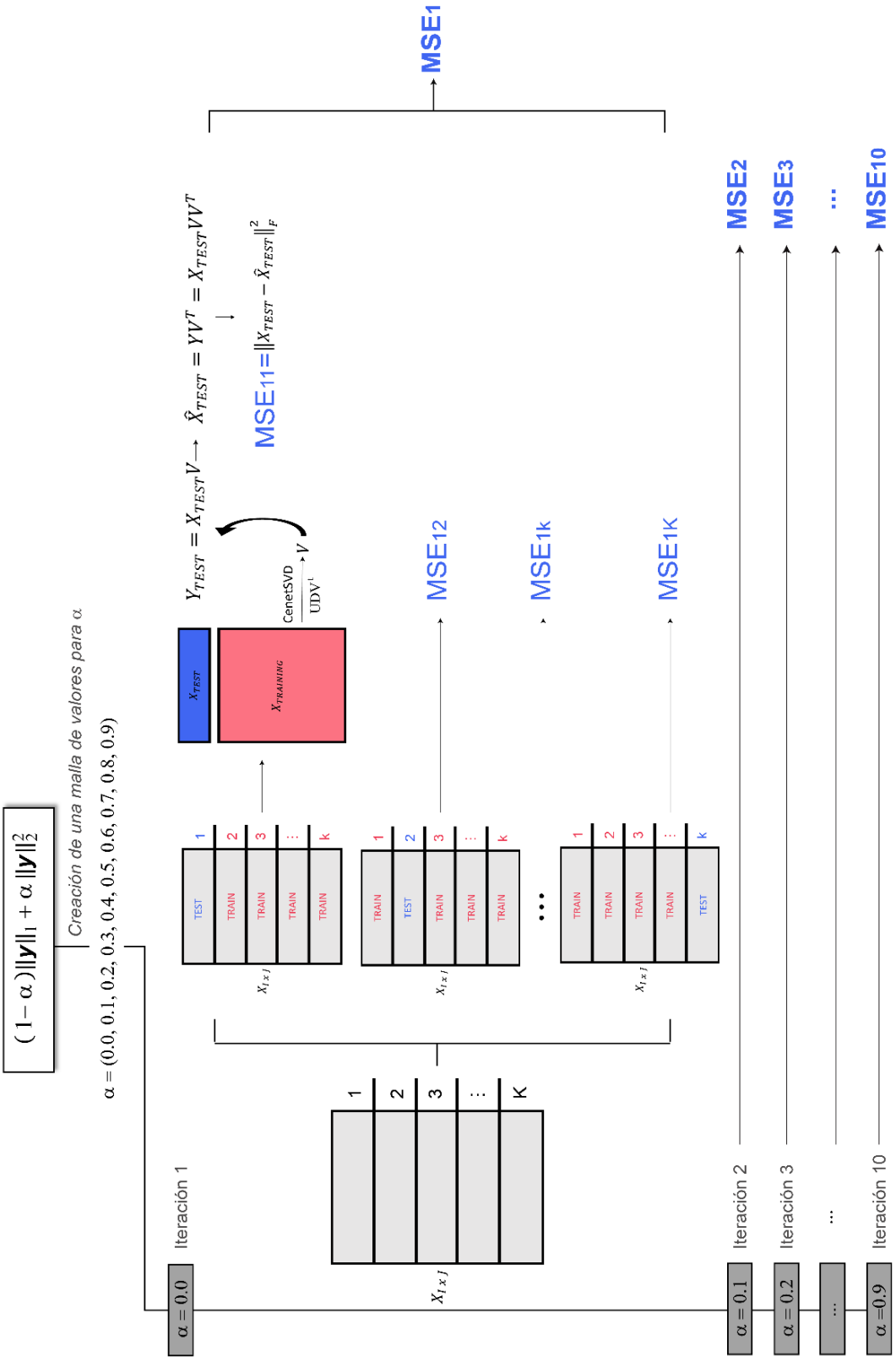


Figura 55. *Proceso general de validación cruzada implementado para la selección de α.*

Ahora bien, como se ha mencionado anteriormente, la selección del parámetro α está íntimamente relacionada con la selección de τ. Esto es así debido a que el cálculo de la CenetSVD de cada una de las matrices de entrenamiento depende de dicho parámetro. Por este motivo, en cada una de las iteraciones definidas en la Figura 55, el $MSE_{\alpha_i,1}$ se calculará como la media de los errores MSE_{α_i,n,τ_r}, a partir de una secuencia de valores τ_r definidos de manera aleatoria para cada parámetro α_i. La Figura 56 recoge una explicación esquemática de este proceso.

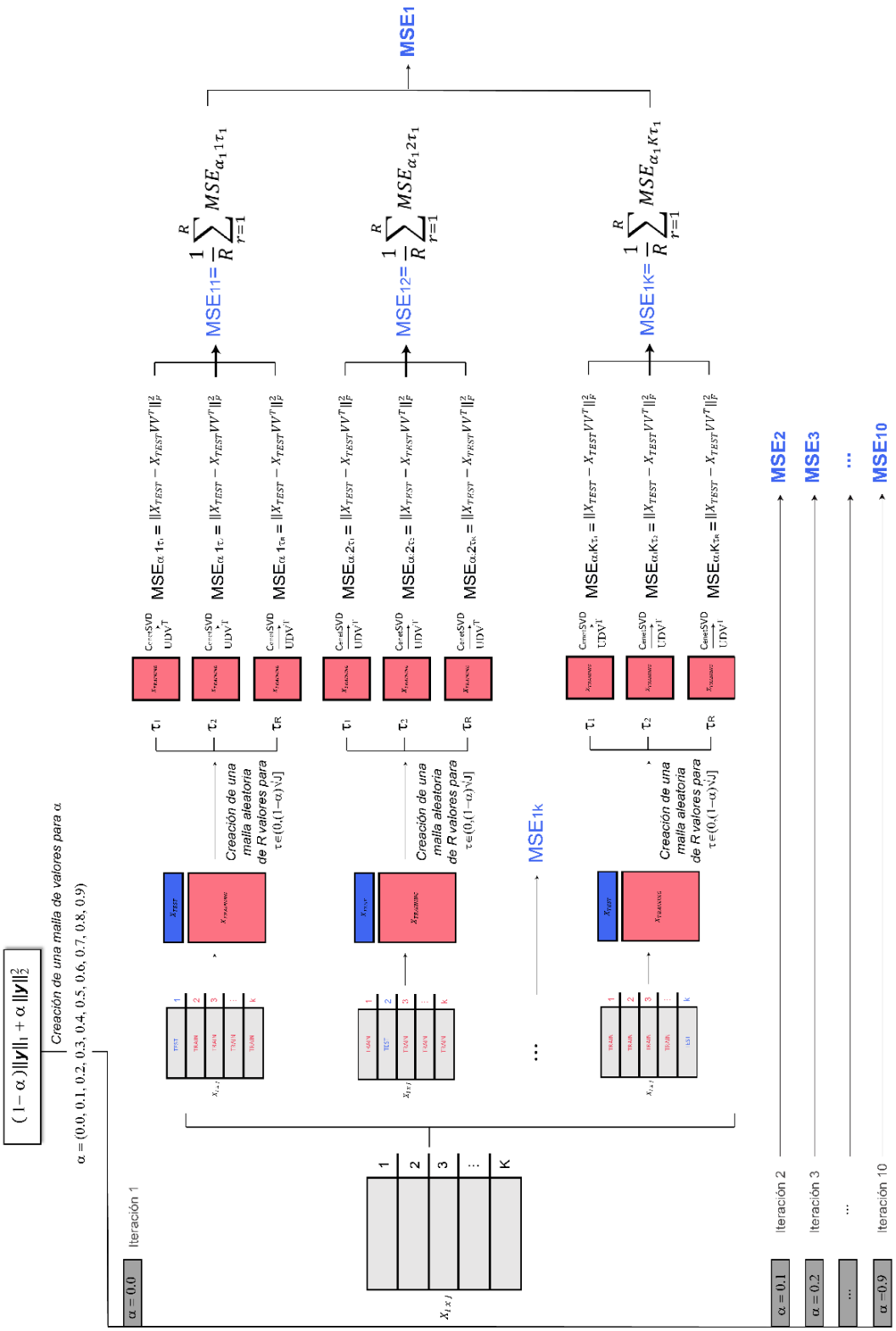


***Figura 56**. Proceso general de validación cruzada implementado para la selección de **α** incorporando **τ** en el proceso*

Selección de τ

La construcción del modelo óptimo está basada en la comparación de diversas medidas que permitan comparar la efectividad de los diferentes modelos diseñados. En este caso, aunque existe una amplia gama de métodos para la decisión, los criterios de información constituyen una de las metodologías más ampliamente utilizadas.

El parámetro de regularización τ controla el grado de sparsity incluida en el modelo mediante Elastic net. En el problema de penalización dual, cuanto mayor es el valor del parámetro de regularización λ más coeficientes son anulados. Sin embargo en la formulación restringida aquí propuesta, τ controla el grado de sparsity a la inversa: cuanto mayor es su valor, menor es la cantidad de cargas cercanas a cero o exactamente nulas. Geométricamente, τ se corresponde con el radio de la bola Elastic net y es por ello que a mayores valores del parámetro, mayor será el área de la región de restricción y se encontrará más lejos del origen de coordenadas.

En nuestro caso se propone la selección del parámetro de regularización τ usando el criterio de información BIC, de la misma forma en que lo implementaron Croux et al. (2011) y Guo, James, Levina, Michailidis y Zhu (2010):

$$BIC(\tau) = \frac{\|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{V}_{enet}\boldsymbol{V}_{enet}^{T}\|^2}{\|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{V}\boldsymbol{V}^{T}\|^2} + df(\tau)\frac{log(i)}{i} \qquad (4.20)$$

donde $\boldsymbol{X}_{IxJ}$ es la matriz de datos original, $\boldsymbol{V}_{enet}$ se refiere a la matriz de vectores pseudo-singulares a la derecha sparse obtenida de la C_{enet}SVD y que contiene las primeras q PC en columnas. La matriz $\boldsymbol{V}$ es la matriz de vectores singulares a derecha obtenida de la SVD clásica no restringida y $df(\tau)$ es el número de elementos no nulos en $\boldsymbol{V}_{enet}$. El parámetro τ que minimiza el BIC(τ) será seleccionado como adecuado de entre una secuencia de posibles valores del parámetro con $\tau \in [1, (1-\alpha)\sqrt{J} + \alpha]$.

4.3 Extensión de la C_{enet}SVD al análisis de datos de dos vías

Finalmente, extendemos la C_{enet}SVD propuesta a diferentes técnicas de reducción de dimensión para el reconocimiento de patrones entre las observaciones consideradas. De esta forma, se expone a continuación la formulación del Constrained PCA o C_{enet}PCA (sparse y ortogonal) restringido al espacio Elastic net y los Constrained Biplots clásicos o C_{enet}Biplots.

El PCA es útil para obtener una representación que permita establecer patrones de comportamiento similar en las observaciones. Sin embargo, permite representar sólo las observaciones o las variables del estudio, pero no ambas a la vez. Para evitar estas desventajas, la bibliografía propone el uso de métodos de representación Biplot. Los Biplots son técnicas de representación simultánea de observaciones y variables en un mismo sistema de referencia, de modo que las relaciones entre ellas se hacen visualmente interpretables (Gabriel, 1971; Galindo, 1986). En el campo del sparse PCA, existen muchos métodos cuyo objetivo es generar cargas sparse y facilitar así la interpretación de los resultados (Jolliffe et al., 2003; Journée & Nesterov, 2010; B. Li et al., 2016; Shen & Huang, 2008; Zou et al., 2006). Sin embargo, la ortogonalidad de la matriz de carga se pierde a expensas de la sparsity. En el caso de los Biplots sparse, la literatura sólo recoge dos técnicas relacionadas con la producción de cargas nulas: CDBiplot (Clustering Disjoint HJ-Biplot) (Nieto-Librero et al., 2017) y Elastic net HJ-Biplot (Cubilla-Montilla et al., 2019). Sin embargo, por un lado el CDBiplot genera ejes factoriales disjuntos, en los que cada variable original sólo contribuye a la generación de un eje factorial. Además, ambas técnicas generan ejes factoriales no ortogonales; es decir, correlacionados, pérdida que se produce al generar ejes sparse. Por esta razón, ampliamos el uso de C_{enet}SVD en la propuesta de un método de PCA sparse y ortogonal simultáneamente (C_{enet}PCA) y métodos Biplots con las mismas características (en lo que se denominará C_{enet}Biplot) dada la estrecha relación entre PCA/Biplot y SVD debido a que los primeros pueden ser implementados utilizando esta última.

4.3.1 Análisis de Componentes Principales restringido C_{enet}PCA (Constrained PCA): soluciones ortogonales y sparse sobre el espacio $\mathfrak{B}_{\ell_1+\ell_2}$

El objetivo principal del C_{enet}PCA es identificar y extraer Q variables latentes (PCs restringidas, *constrained* PCs) y transformar las J variables originales correlacionadas en un nuevo conjunto de variables no correlacionadas y sparse, con $Q < J$, proyectando el espacio original en un subespacio de menor dimensión. Dada $\boldsymbol{X}_{IxJ}$, la q-PC restringida (sparse y orthogonal) se define como una combinación lineal de las variables observadas x_j, con $j = 1, \dots, J$, como:

$$\boldsymbol{Y}_{IxQ} = \boldsymbol{X}_{IxJ}\, \boldsymbol{V}_{enet_{JxQ}} \tag{4.21}$$

donde $\boldsymbol{Y}_{IxQ}$ es la matriz de puntuaciones factorials que contiene en sus filas los valores de las observaciones de la tabla de datos inicial en el nuevo espacio restringido de PCs. La matriz de cargas sparse $\boldsymbol{V}_{\text{enet}}$ es la matriz de proyección en el nuevo subespacio generado por ejes ortogonales. Los vectores columna de la matriz de cargas incluyen la contribución de las variables iniciales a la formación de cada PCs restringida sparse. Se muestra en la Figura 57 un esquema de la construcción de C_{enet}PCA.

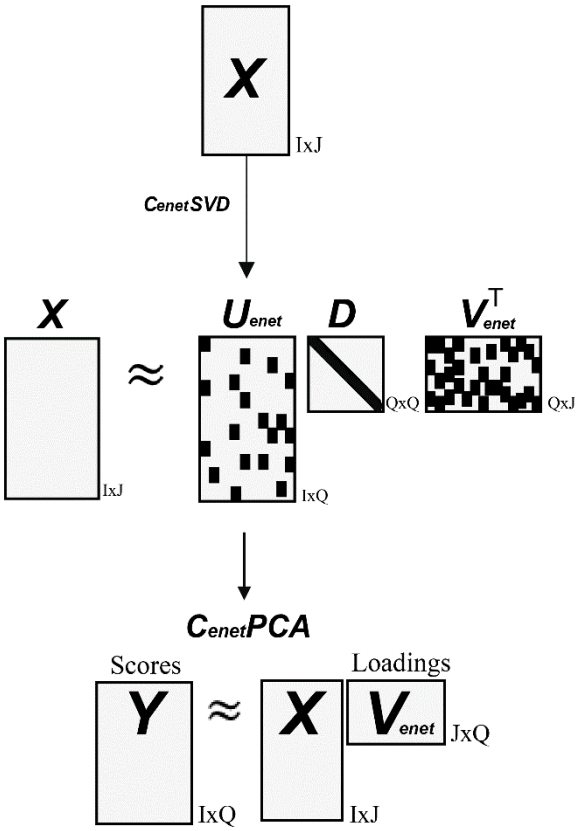


Figura 57. *PCA restringido sparse y ortogonal (CenetPCA). Construcción del modelo mediante CenetSVD*

El pseudocódigo del algoritmo para implementar el CenetPCA se recoge en la Tabla 23. Esta técnica ha sido implementada en el software libre R en la función *pca.enet* (Figura 58):

```
pca.enet<-function(X, Q=2, tau.u = 1.4, tau.v = 1.4,alpha.u=1e-16, alpha.v=1e-16,
                   itermax.pi=1000, itermax.pocs=1000,eps.pi=1e-16, eps.pocs=1e-16,
                   init.svd="svd", init.transf=1,obs.names=FALSE,plot.axis=c(1,2))
```

Figura 58*. Argumentos de la función pca.enet desarrollada en R*

Además de las matrices de puntuaciones factoriales $\boldsymbol{Z}_{enet} \in \mathbb{R}^{IxQ}$ y de cargas esta función devuelve un gráfico de dispersión bidimensional con las puntuaciones factoriales de los individuos en el plano factorial 1-2 por defecto. El plano factorial puede ser modificado en el argumento plot.axis de la función. Además, devuelve un mapa de calor con los coeficientes de las cargas factoriales en cada una de las componentes retenidas.

***Tabla 23**. Análisis de Componentes Principales Restringido sobre Elastic net con soluciones ortogonales*

Algoritmo C_{enet}PCA	
Entrada:	$\boldsymbol{X} \in \mathbb{R}^{IxJ}$, rango Q, $\varepsilon \approx 0$, $\tau_v \in [1, (1-\alpha_v)\sqrt{J} + \alpha_v]$, $\tau_u \in [1, (1-\alpha_u)\sqrt{J} + \alpha_u]$, $\alpha_u \in [0,1)$, $\alpha_u \in [0,1)$
Salida	$\boldsymbol{Z}_{enet} \in \mathbb{R}^{IxQ}$, $\boldsymbol{V}_{enet} \in \mathbb{R}^{QxJ}$, % de varianza explicada
1: **Calcular** C_{enet}SVD de $\boldsymbol{X}$:	
2: $\boldsymbol{U}_{enet} = C_{enet}SVD(\boldsymbol{X}, \text{rango } = Q, \tau_u, \alpha_v)\$\boldsymbol{U}$	
3: $\boldsymbol{V}_{enet} = C_{enet}SVD(\boldsymbol{X}, \text{rango } = Q, \tau_v, \alpha_u)\$\boldsymbol{V}$	
4: $\boldsymbol{D}_{enet} = C_{enet}SVD(\boldsymbol{X}, \text{rango } = Q, \tau_v, \alpha_v)\$\boldsymbol{D}$	
5: **Devolver:**	
$\boldsymbol{Z}_{enet} = \boldsymbol{X}\boldsymbol{V}_{enet} \in \mathbb{R}^{IxQ}$	#Matriz de puntuaciones factoriales
$\boldsymbol{V}_{enet} \in \mathbb{R}^{QxJ}$	#Matriz de cargas factoriales
$\left(\frac{\boldsymbol{D}_{enet}{}^2}{\|\boldsymbol{X}\|_F^2}\right) \cdot 100$	#Varianza explicada por cada componente C_{enet}
6: **Fin**	

4.3.2 Métodos Biplot clásicos restringidos C_{enet}Biplot (constrained Biplot): soluciones ortogonales y sparse sobre el espacio $\mathfrak{B}_{\ell_1+\ell_2}$

Los métodos Biplot son herramientas óptimas de visualización de una matriz multivariante en un espacio de baja dimensión, en el que las covariaciones entre observaciones y variables se hacen visualmente interpretables. Los métodos Biplot clásicos más utilizados son el JK-Biplot y GH-Biplot propuestos por Gabriel (1971), que asignan una óptima calidad de representación en el gráfico a filas (GH-Biplot) o columnas (JK-Biplot) en un mismo espacio euclídeo. Junto a estos se desarrolla el HJ-Biplot de Galindo (1986), que se diferencia de los anteriores en que proporciona la máxima calidad de representación tanto a filas como a columnas de la matriz de datos inicial en el mismo sistema de referencia.

La factorización Biplot trabaja descomponiendo el conjunto de datos original en el producto de dos matrices $\boldsymbol{X} = \boldsymbol{A}\boldsymbol{B}^T$, con $\boldsymbol{A} \in \mathbb{R}^{IxQ}$ y $\boldsymbol{B} \in \mathbb{R}^{QxJ}$ matrices de marcadores fila y columna respectivamente, de manera que el producto interno $a_i^{\mathrm{T}} b_j$ aproxima el elemento x_{ij} lo mejor posible. De la C_{enet}SVD

de la matriz de datos original, el JK-Biplot establece los marcadores fila y columna como $\boldsymbol{A} = \boldsymbol{U}_{enet}\boldsymbol{D}$ y $\boldsymbol{B} = \boldsymbol{V}_{enet}$, el JK-Biplot como $\boldsymbol{A} = \boldsymbol{U}_{enet}$ y $\boldsymbol{B} = \boldsymbol{V}_{enet}\boldsymbol{D}$ y por último el HJ-Biplot los establece como $\boldsymbol{A} = \boldsymbol{U}_{enet}\boldsymbol{D}$ y $\boldsymbol{B} = \boldsymbol{D}\boldsymbol{V}_{enet}$ (Figura 59).

Para realizar una correcta interpretación de los resultados de Biplot se recuerdan algunas nociones básicas: i) las observaciones están representados por puntos y las variables por vectores; ii) la distancia entre puntos muestra similitudes entre observaciones; iii) la longitud del vector se refiere a la variabilidad de la variable; iv) la relación entre variables se interpreta desde los ángulos entre los vectores respectivos (obtuso: relación inversa; agudo: relación directa; ángulo recto: independencia lineal); y v) la proyección de un punto en la dirección de un vector aproxima el valor de la variable para la observación correspondiente.

En el caso de los métodos C_{enet}Biplots debe tenerse en cuenta que cuanto mayor sea la restricción introducida en el modelo (menor τ) en cualquiera de las dos dimensiones, los marcadores fila y/o marcadores columna se encontrarán más cercanos a los ejes factoriales. En un caso extremo de restricción de las variables, el C_{enet}Biplot se parecerá al Disjoint Biplot (DBiplot) de Nieto-Librero (2015) en el que cada variable contribuye únicamente a la formación de un eje factorial y no a otro y los marcadores columna quedan dispuestos sobre el eje del plano factorial.

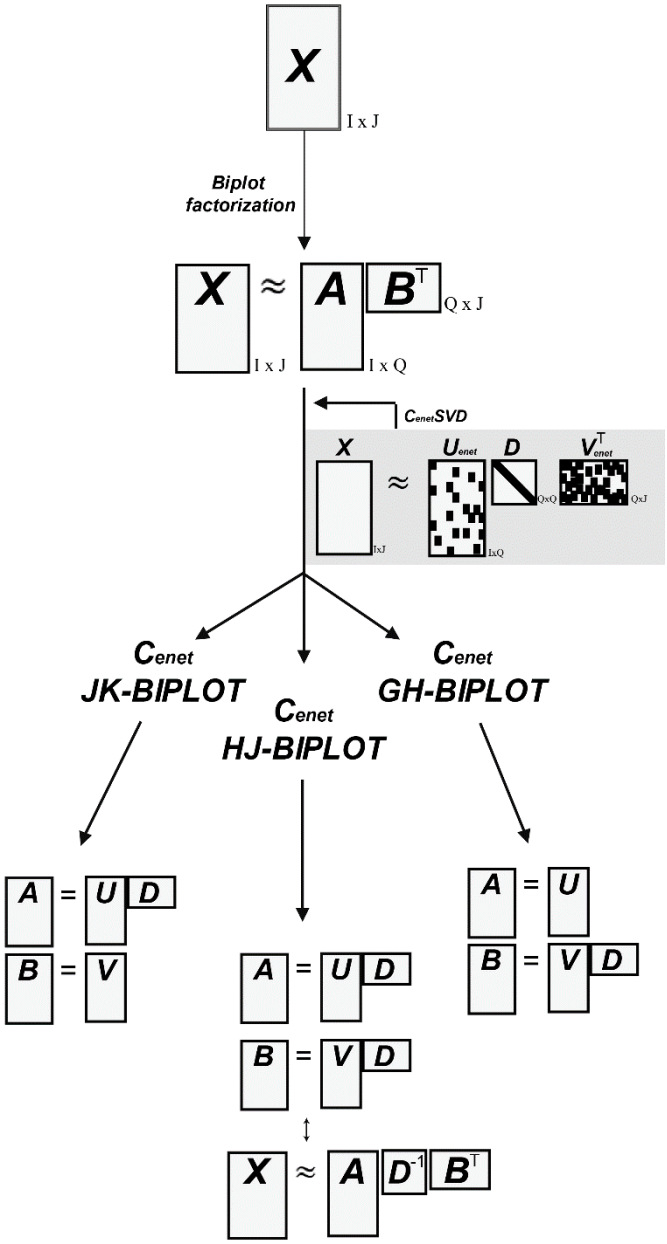


Figura 59. *Biplots clásicos (GH, JK, HJ) restringidos sparse y ortogonal (C_{enet}Biplots). Construcción del modelo mediante C_{enet}SVD*

A partir de la metodología descrita en referencia a la C_{enet}SVD, es inmediato implementar lo que se definirá de ahora en adelante como los métodos Biplots clásicos restringidos (*Constrained Biplots*) (Tabla 24, Figura 60). Además, la función incorpora la posibilidad de realizar una selección previa de variables a partir de los leverage de la descomposición CUR (véase capítulo 1), para escoger aquellas características que presenten la máxima variabilidad en las componentes a retener posteriormente en el C_{enet}Biplot. La función *Biplot.enet*

desarrollada devuelve al usuario una lista que contiene: i) la matriz de correlaciones dos a dos para las variables consideradas numérica y gráficamente; ii) las matrices de marcadores fila y columna para las componentes retenidas; iii) el porcentaje de varianza explicada por cada una de las componentes C_{enet}; iv) los resultados de la selección de la descomposición CUR si esta se ha realizado y v) la representación Biplot en los ejes escogidos a graficar mediante el argumento plot.axis.

```
Biplot.enet<-function(X, Q=2, tau.u = 1.4, tau.v = 1.4,alpha.u=1e-16, alpha.v=1e-16,biplot.type=2,
                      itermax.pi=1000, itermax.pocs=1000,eps.pi=1e-16, eps.pocs=1e-16,
                      init.svd="svd", init.transf=1,  plot.axis=c(1,2),names.obs=FALSE,
                      select.cur=FALSE, variables.cur=1, weighted.cur=FALSE, method.cur="top.scores")
```

Figura 60. Argumentos de la función Biplot.enet desarrollada en R

***Tabla 24**. Biplot restringido sobre Elastic net con soluciones ortogonales y sparse*

Algoritmo C_{enet}BIPLOT	
Entrada:	$\boldsymbol{X} \in \mathbb{R}^{IxJ}$, rango Q, $\varepsilon \approx 0$, $\tau_v \in [1,(1-\alpha_v)\sqrt{J}+\alpha_v]$, $\tau_u \in [1,(1-\alpha_u)\sqrt{J}+\alpha_u]$, $\alpha_u \in [0,1), \alpha_u \in [0,1)$, Biplot.type=2
Salida	$\boldsymbol{A}_{enet} \in \mathbb{R}^{IxQ}, \boldsymbol{B}_{enet} \in \mathbb{R}^{QxJ}$, % de varianza explicada

1: **Calcular** C_{enet}SVD de $\boldsymbol{X}$:

2: $\boldsymbol{U}_{enet} = C_{enet}SVD(\boldsymbol{X}, \text{rango } = Q, \tau_u, \alpha_v)\$\boldsymbol{U}$

3: $\boldsymbol{V}_{enet} = C_{enet}SVD(\boldsymbol{X}, \text{rango } = Q, \tau_v, \alpha_u)\$\boldsymbol{V}$

4: $\boldsymbol{D}_{enet} = C_{enet}SVD(\boldsymbol{X}, \text{rango } = Q, \tau_v, \alpha_v)\$\boldsymbol{D}$

5: **Definir** las matrices de marcadores $\boldsymbol{A}_{enet}$ y $\boldsymbol{B}_{enet}$:

6: $\boldsymbol{A}_{enet} = \boldsymbol{U}_{enet}\boldsymbol{D}_{enet}$ #Matriz de marcadores fila sparse

7: $\boldsymbol{B}_{enet} = \boldsymbol{V}_{enet}\boldsymbol{D}_{enet}$ #Matriz de marcadores columna sparse

8: **Si** $biplot.type \leq 1$

$\boldsymbol{A}_{enet} = \boldsymbol{U}_{enet}\boldsymbol{D}_{enet}^{1-biplot.type}$ #Matriz de marcadores fila sparse

$\boldsymbol{B}_{enet} = \boldsymbol{V}_{enet}\boldsymbol{D}_{enet}^{biplot.type}$ #Matriz de marcadores columna sparse

9: **Fin Si**

10: $\left(\frac{\boldsymbol{D}_{enet}{}^2}{\|\boldsymbol{X}\|_F^2}\right) \cdot 100$ #Varianza explicada por cada componente C_{enet}

11: **Reescalado de los datos:**

$sA = sum(\boldsymbol{A}_{enet}^2)/I$

$sB = sum(\boldsymbol{B}_{enet}^2)/J$

12: $\boldsymbol{A}_{enet} = \boldsymbol{A}_{enet}\sqrt{\sqrt{sB/sA}}$

13: $\boldsymbol{B}_{enet} = \boldsymbol{B}_{enet}\sqrt{\sqrt{sB/sA}}$

14: **Fin**

4.4 Aplicación a datos reales

Se analiza a continuación la base de datos de muestras de pacientes con leucemia y pacientes sanos presentada en la sección “Objetivos y Metodología”. Se trata de una base formada por un total de 216 muestras (n=71 muestras de pacientes con leucemia linfoblástica aguda (ALL), n=74 muestras de leucemia linfoblástica crónica (CLL) y n=71 muestras de pacientes control). Las 216 muestras fueron seleccionadas aleatoriamente de las 2.096 muestras disponibles en la serie GSE13204 del repositorio GEO que se coleccionaron bajo el marco del proyecto MILE (*Microarray Innovations in LEukemia*). Sobre estos pacientes se obtuvo la información de la expresión génica de un total de 54.613 probes mediante microarrays de la plataforma HGU133Plus2, de las cuales un total de 44.692 fueron analizables (referentes a un total de 21.336 genes). El resto eran probes control de Affymetrix o probes respectivas a genes de los cuales no se tiene información. Los datos fueron preprocesados según la normalización RMA. Todos los análisis se han realizado en R.

A modo ilustrativo, se seleccionaron los 2.000 genes con mayor variabilidad en las dos componentes que posteriormente serán retenidas. Habitualmente en genética los genes son filtrados o seleccionados según su variabilidad individual. Sin embargo, este enfoque puede correr muchos riesgos por no tener en cuenta lo que le ocurre al gen de manera global; es decir su variabilidad conjunta con otros genes. Así, se podrían descartar genes que actúen conjuntamente con otros y marginalmente no tengan una actividad apreciable. Por este motivo, la selección de los genes con mayor variabilidad se realiza a partir de los leverage de la descomposición CUR.

Posteriormente se realizó un PCA clásico sobre la matriz centrada y estandarizada de 216 muestras y 2.000 probes elegidas. La Figura 61 (panel A) muestra el gráfico de puntuaciones factoriales en los ejes factoriales 1-2 del PCA clásico, explicando el 45,5% de la variabilidad de los datos. El análisis de las puntuaciones revela tres subgrupos de muestras con patrones claramente diferenciados, correspondientes a cada tipo de muestra (verde: muestras normales; azul: TODAS las muestras; rojo: muestras CLL). El primer eje factorial muestra la separación entre las muestras normales y las CLL, con las CLL

ubicadas en el lado positivo del eje 1. Las ALLs fueron discriminadas de las CLL y de los tejidos normales por sus valores del eje 2. Sin embargo, a pesar de que el PCA distribuye grupos de observaciones similares, no permite encontrar la causa genética de dicha separación. La Figura 61 (panel A) muestra la contribución de cada sonda génica a la estimación de los PCs (la matriz de cargas para las dos componentes retenidas). Cada PC se computa como una combinación lineal de las 2000 sondas génicas medidas. Esto provoca que dichas variables sean poco prácticas al ser combinación de un gran número de genes, a pesar de su buena capacidad discriminatoria.

Los datos se analizaron mediante el C_{enet}PCA reteniendo nuevamente dos componentes principales penalizadas. La penalización Elastic net se definió utilizando $\alpha = 0{,}5$ mediante validación cruzada, compuesta por la misma cantidad de penalización Lasso (para lograr coeficientes sparse) y Ridge (para que genes correlacionados pudieran contribuir a la formación de una misma componente). El parámetro τ de regularización se definió como el coeficiente que generase un valor mínimo del criterio de información BIC ($\tau = 5{,}86$). La Figura 61 (panel B) recoge los resultados del C_{enet}PCA. Se contempla en el gráfico de puntuaciones factoriales de las muestras en los dos primeros ejes, la misma clasificación de muestras tumorales que en el caso del PCA clásico. Los gráficos para las cargas de la Figura 61 (panel B, medio e inferior) evidencian una clara construcción de cada una de las C_{enet}PCs, a la vez que mantienen la contribución de los genes más relevantes encontrados en los resultados del PCA clásico. De las 2.000 probes consideradas, 1.752 presentaron coeficientes exactamente nulos en las dos componentes retenidas y de las 248 restantes, 43 variables presentaban contribuciones por debajo del 0,01 en ambos ejes sparse restringidos.

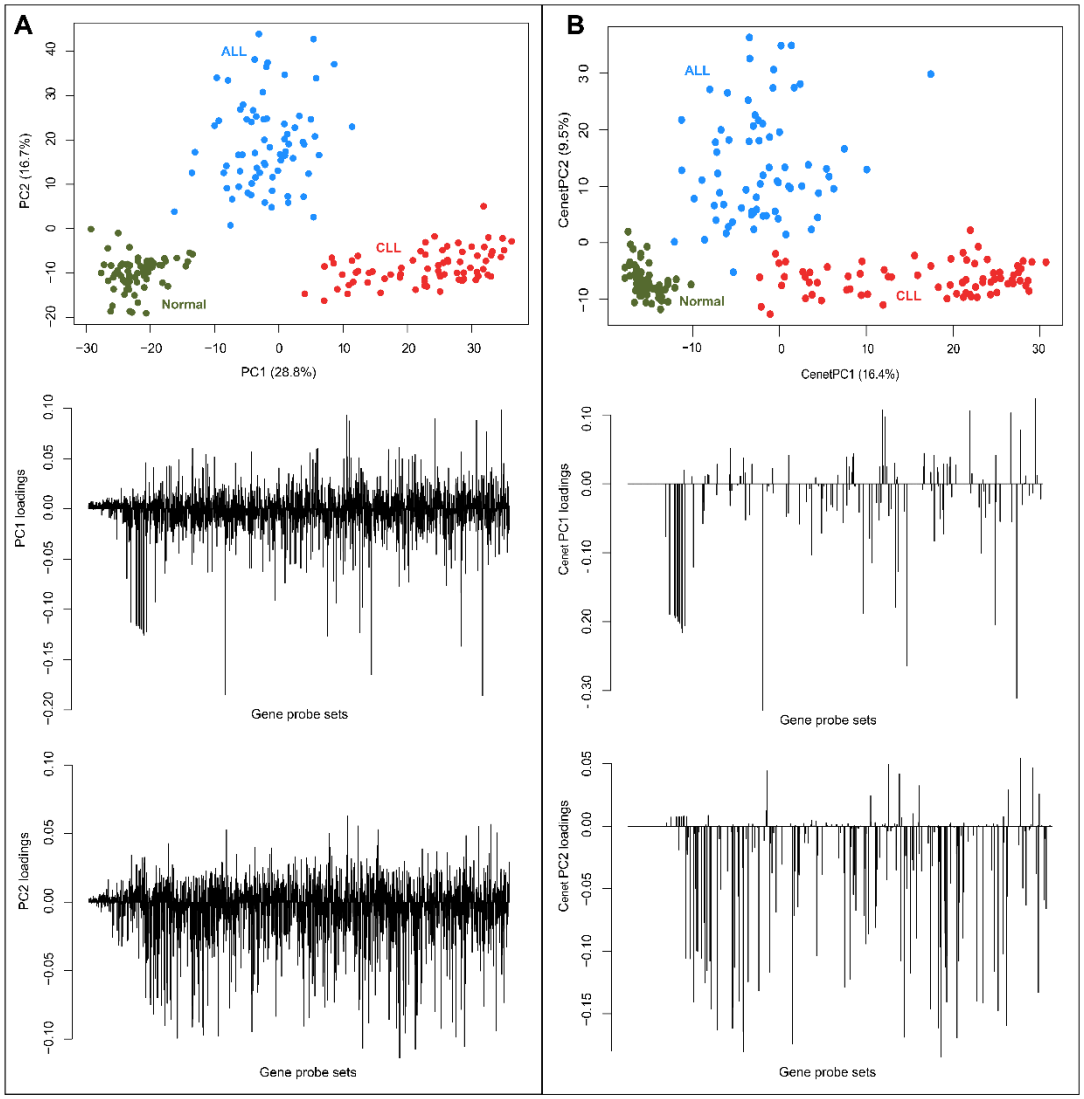


Figura 61. *Ejes factoriales 1-2 del PCA clásico (panel A) y ejes factoriales 1-2 del PCA restringido sobre la bola Elastic net (CenetPCA) (panel B). Cada color se refiere a los tres grupos de muestras de estudio (verde: muestras normales; rojo: Muestras de leucemia CLL; azul: muestras de leucemia ALL).*

Debido a que la discriminación de los grupos tumorales en un espacio de menor dimensión fue apropiada, se analizó la matriz de expresión de las 216 muestras y 205 sondas de genes no nulas identificadas en el CenetPCA. Para caracterizar los niveles de expresión de los genes característicos de cada uno de los tres grupos de muestras identificados se aplicó sobre dicha matriz de datos el CenetHJ-Biplot. El modelo HJ-Biplot fue escogido para asignar alta calidad de representación tanto a muestras como a genes en la representación gráfica del espacio de dimensión reducida. En este último análisis, como se deseaba incluir en el modelo un nivel medio de sparsity, se establece $\tau = ((1 -$

$\alpha)\sqrt{J} + \alpha) * (1/3)$ utilizando un factor de escala $(1/3)$ siguiendo la línea de lo sugerido en (V Guillemot et al., 2019). Los resultados del plano factorial 1-2 del HJ-Biplot sparse se muestran en la Figura 62. Al igual que ocurría en el plano factorial del PCA (Figura 61) los tres subgrupos de muestras son discriminados por los ejes factoriales restringidos. En este caso, el eje 1 es un gradiente de información que discrimina las muestras de CLL de las leucemias normales y de las leucemias ALL. El eje 2 es la dirección principal que diferencia las ALLs del resto. En consecuencia, la interpretación de C_{enet}HJ-Biplot hace posible el reconocimiento de una caracterización genética de cada uno de los clusters a partir de un subconjunto de los genes originales, ya que el resto han obtenido cargas nulas. En este sentido, las muestras normales se caracterizan por una alta expresión de genes que contribuyen a la formación del eje 1 negativo; es decir, *DEFA1, HBB, HBA1* y *LTF*; genes que explican la formación del eje negativo 2 (*ANXA1, TYROBP, FCN1, LYZ*) y *S100A12*. Además, las muestras normales presentan intensidades más bajas de *TCF4, POU2AF1, IGHM, FCMR* y *HLA_DPA1*. Estos últimos 5 genes son responsables de la discriminación entre muestras normales y tumorales, debido a que las leucemias ALL y CLL presentan altas intensidades de estos genes. Las ALLs se diferencian de las CLLs por expresión más alta de *LEF1, SEPTIN6, MARCKSL1, MAGED1, PMAIP1, CD99* y *PMAIP1* (las CLLs presentan expresiones más bajas) y señales más bajas de los genes *FCMR, IGHM, HLA_DPA1* (las CLLs presentan intensidades más altas).

Estas técnicas confieren al investigador la posibilidad de reconocer grupos de observación con patrones similares y las variables causales de dichos grupos. Además, se convierten en técnicas de selección de características que mejoran la interpretación de los resultados gracias a los ejes factoriales sparse, con parte de sus coeficientes exactamente cero. En el campo de la genómica, las técnicas C_{enet} se presentan como herramientas prometedoras en la clasificación del cáncer, entendiendo las causas de los grupos de muestras basadas en pequeños subconjuntos de genes.

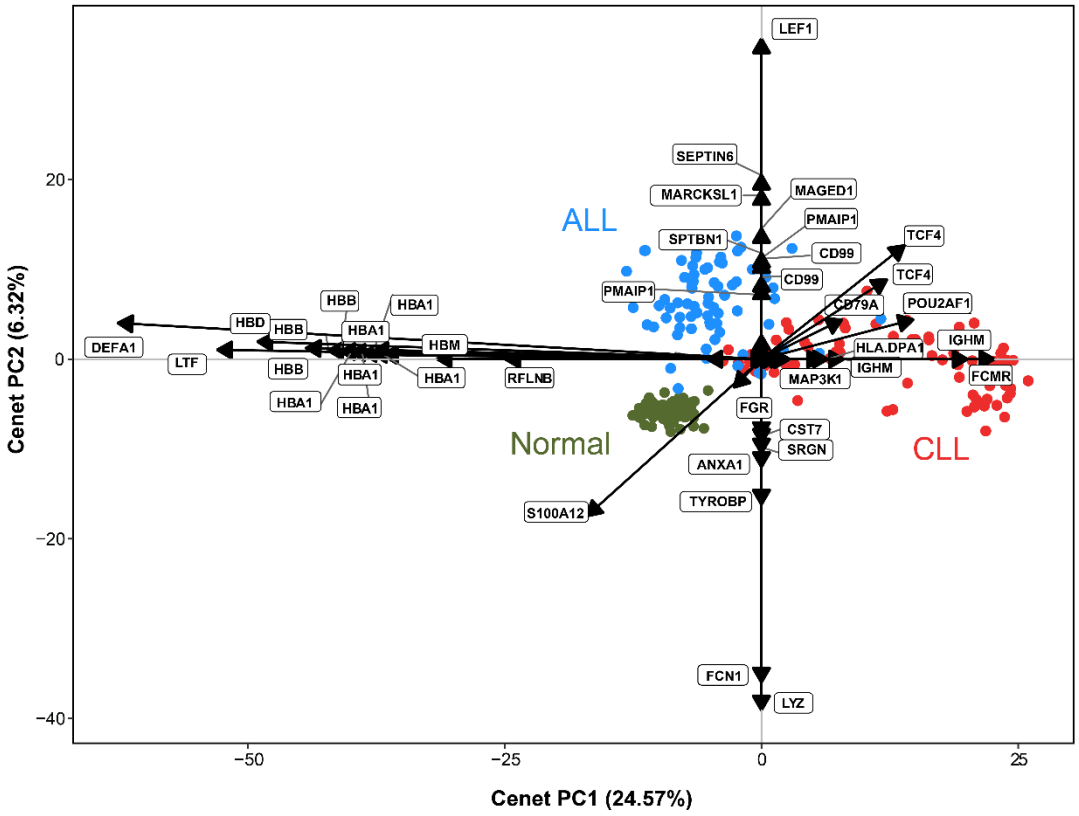
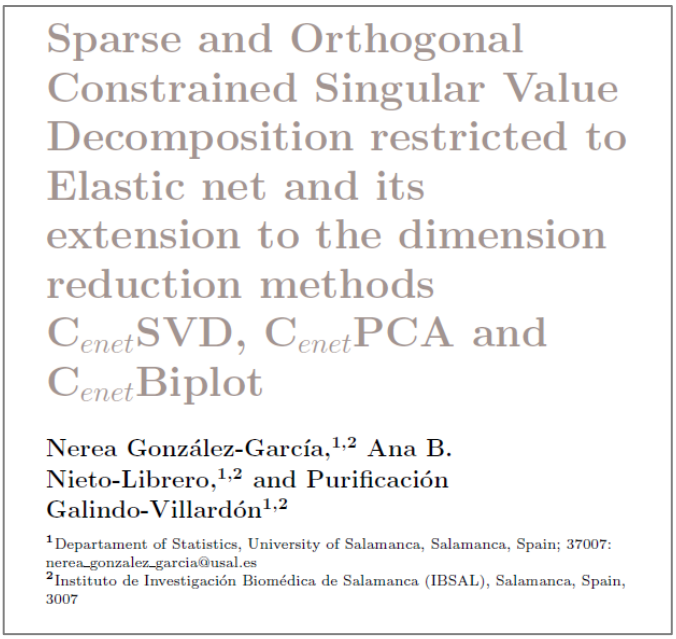


Figura 62. *Ejes factoriales 1-2 del HJ-Biplot restringido a la bola Elastic net (CenetHJ-Biplot) . Cada color se refiere a uno de los tres grupos de muestras en estudio (verde: muestras normales; rojo: muestras CLL; azul: muestras ALL)*

Toda la contribución de este capítulo ha sido sometida como artículo de investigación a la revista "*Annual Review of STATIStics and Its Application*" (JCR 3.857 Q1), bajo el título "Sparse and orthogonal constrained singular value decomposition restricted to Elastic net and its extensión to the dimensión reduction techniques: CenetSVD, CenetPCA and CenetBiplot" (Figura 63).



Sparse and Orthogonal Constrained Singular Value Decomposition restricted to Elastic net and its extension to the dimension reduction methods C_{enet}SVD, C_{enet}PCA and C_{enet}Biplot

Nerea González-García,[1,2] Ana B. Nieto-Librero,[1,2] and Purificación Galindo-Villardón[1,2]

[1]Departament of Statistics, University of Salamanca, Salamanca, Spain; 37007: nerea_gonzalez_garcia@usal.es
[2]Instituto de Investigación Biomédica de Salamanca (IBSAL), Salamanca, Spain, 3007

Figura 63. *Artículo sometido a la revista "Annual Review of STATIStics and Its Application" (JCR 3.857 Q1)*

4.5 Código de proyección de un vector sobre $\mathfrak{B}_{\ell_1+\ell_2}(\tau) \cap \mathfrak{B}_{\ell_2}(1)$ en R

A continuación se muestra el código implementado en R para la proyección de un vector sobre la intersección de los espacios convexos enet (L1+L2) y L2.

```
projL1_2L2 <- function(x, tau, alpha=1e-16){
  norm2_x <- norm2(x)
  if (norm2_x < 1e-32 ) {return(list(x=x))} #Comprobación de que el denominador no es 0
  if ((((1-alpha)*norm1(x)+(alpha)*norm2_x^2)/norm2_x) <= tau )
   { return(list(x=x/norm2_x)) } #Comprobación de que no se cumpla ya la restricción
  #Paso 1. Cálculo del vector de coeficientes en valores absolutos
  uneq<-x != 0
  p <- abs(x[x != 0])

  # Chequear τ:
  MAX=max(p)
  bMAX<-p==MAX
  nMAX<-sum(bMAX)
  #Posible modificación de τ:
  if(tau>((1-alpha)*sqrt(length(x))+alpha))
  stop("Impossible to project, maximum ratio is : ", ((1-alpha)*sqrt(length(x))+alpha*norm2(x)))
  if(tau<=0) stop("Impossible to project, minimum ratio is : ",0)
```

```
#Inicializacion
 s1 <-0
 s2 <-0
 nb <- 0
 #Paso 2. Procedimiento para encontrar k que verifica phi(a_k)<=tau<phi(a_k+1)
 while (T) {
        N <- length(p)
        #Primera elección de k aleatoria
        a_k <- p[sample(1:N,1)]
            while(a_k == MAX){ a_k <- p[sample(1:N,1)]  }
        #Partición de p:
        p_inf_ak <- p < a_k
        p_sup_ak <- p > a_k
        p_high <- p[p_inf_ak]
        p_low <- p[p_sup_ak]
        #Cálculo de k
        nb_a_k <- sum(p == a_k)
        k <- nb + sum(p_sup_ak) + nb_a_k

        #Cálculo de la función omega
        aksq <- a_k^2
        slow_1 <- sum(p_low) + nb_a_k*a_k
        slow_2 <- ssq(p_low) + nb_a_k*aksq
```

```
    omega_a_k<-(alpha*(s2 + slow_2) +(1-alpha)*(s1 + slow_1) -
k*a_k*((1-alpha)^2) - k*(a_k^2)*alpha*(1-
alpha)^2)/((1+2*alpha*a_k)*sqrt(s2 + slow_2 - 2*a_k*(1-alpha)*(s1 +
slow_1) + k*aksq*(1-alpha)^2))


    #Divide y vencerás: selección de la partición en función de la
restricción
    if (omega_a_k > tau){
      if ( length(p_low) <= 1 ) break #length(p_low) == 0
      p <- p_low
    }else{
    if (length(p_high) <=1){ break }
    else{
    a_k_1 <- max(p_high)
    omega_a_k_1<-(alpha*(s2 + slow_2) +(1-alpha)*(s1 + slow_1) -
k*a_k_1*((1-alpha)^2) - k*(a_k_1^2)*alpha*(1-
alpha)^2)/((1+2*alpha*a_k_1)*sqrt(s2 + slow_2 - 2*a_k_1*(1-alpha)*(s1 +
slow_1) + k*a_k_1^2*(1-alpha)^2))
    if (omega_a_k_1 >tau){ break }
    p <- p_high #Actualización de p
    nb <- k #Actualización de k
    s1 <- s1 + slow_1 #Actualización de s1
    s2 <- s2 + slow_2 #Actualization de s2
        }
  }
}
```

```
#Paso 3. Cálculo de lambda

l1<-s1 + slow_1

l2<-s2 + slow_2

#Término independiente:

a<-(tau^2)*l2-(alpha^2)*(l2^2)-2*alpha*l2*(1-alpha)*l1-(l1^2)*((1-alpha)^2)

#Coeficiente grado 1:

b<- 4*alpha*l2*(tau^2)-2*(tau^2)*(1-alpha)*l1+2*l2*k*((1-
alpha)^2)*alpha+2*l1*k*((1-alpha)^3)

#Coeficiente grado 2:

c<- k*(tau^2)*((1-alpha)^2)+4*(alpha^2)*l2*(tau^2)-8*alpha*(tau^2)*(1-
alpha)*l1+2*k*l1*alpha*((1-alpha)^3)+2*k*l2*(alpha^2)*((1-alpha)^2)-(k^2)*((1-
alpha)^4)

#Coeficiente grado 3:

d<-4*k*alpha*(tau^2)*((1-alpha)^2)-8*alpha^2*(tau^2)*(1-alpha)*l1-
2*(k^2)*alpha*((1-alpha)^4)

#Coeficiente grado 4:

e<-4*k*(alpha^2)*(tau^2)*((1-alpha)^2)-(k^2)*(alpha^2)*((1-alpha)^4)

#Resolución de la ecuación de grado 4:

sol<-Re(polyroot(c(a,b,c,d,e)))

sol<-sol[which(sol>=0)]

#Selección de la solución valida. Pertenece al intervalo [abs(x1),abs(xJ)]:

lambda<-sol[which(sol>=min(abs(x[uneq]))&sol<=max(abs(x[uneq])))]

#Si existe más de una válida, selección de la mínima:

if(length(lambda)>1){lambda<-min(lambda)}

if(length(lambda)<1){lambda=0}
```

```
 #Paso 4. Cálculo del vector solución a partir de la composición de los operadores soft-thresholding y L2

x_en <- (sign(x)*pmax(0, abs(x) - lambda*(1-alpha)))/(1+2*lambda*alpha)

  return( list(x=x_en / norm2(x_en) , penalizacion=lambda, sol=sol) )

 }
```

CAPÍTULO 5

ANÁLISIS SPARSE DE DATOS DE TRES VÍAS

Los datos de dos vías hacen referencia a tablas de datos en las que se almacena información de un conjunto de I observaciones, frecuentemente individuos, sobre una serie de J variables. Esta información es almacenada en matrices de tamaño IxJ como se ha mencionado en capítulos anteriores. Sin embargo, cada vez es más habitual en algunas disciplinas como las englobadas dentro de las ciencias del comportamiento (campos de la sociología, psicología, antropología humana, neurociencias, economía, geografía, ciencia política…) disponer de datos de una serie de J variables que miden características de I sujetos en K condiciones diferentes, como pueden ser distintos momentos del tiempo. En este tipo de situaciones, los datos son almacenados en arreglos tridimensionales (*arrays*) o matrices de datos de tres vías (modos, dimensiones) que en algunos ámbitos son conocidos como tensores multidimensionales; en este caso, de 3 dimensiones. Algunos ejemplos de tensores de distintas dimensiones pueden verse en la Figura 64.

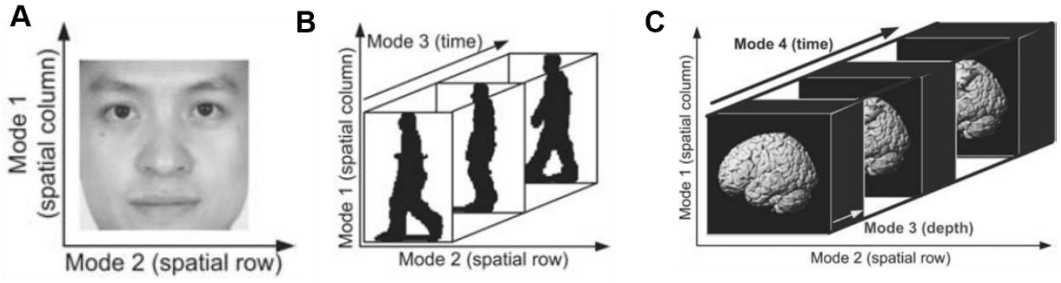


***Figura 64**. Ejemplo de tensores de orden 2 (panel A), de orden 3 (panel B) y de orden 4 (panel C). Fuente: (Lu, Plataniotis, Venetsanopoulos, & More, 2013)*

Aunque algunas investigaciones trabajan con datos estructurados en más de tres dimensiones, se tomará como referencia en este proyecto el análisis de datos de tres vías/modos: conjunto de datos que se representa en un bloque tridimensional (modo 1=individuos, modo 2=variables y modo 3=condiciones).

Al igual que ocurría en su equivalente bidimensional, las técnicas de factorización matricial, en este caso de matrices multidimensionales, son una

herramienta básica para descubrir patrones latentes en los datos y poder representar la información en espacios de menor dimensión donde sean interpretables. En este sentido, la literatura destaca dos grandes bloques de técnicas propuestas: las llamadas técnicas asimétricas, que parten de una estructura desdoblada del arreglo de tres vías original, y las llamadas técnicas simétricas, que mantienen la estructura tridimensional del cubo de datos original (Figura 65). Las técnicas asimétricas son relacionadas con los métodos franceses y algunas de las más conocidas son el MFA (Escofier & Pagès, 1983) o la familia de los métodos STATIS (Abdi et al., 2012; L'Hermier des Plantes, 1976). Este tipo de técnicas parten de una estructura desdoblada de la matriz de datos, perdiendo la estructura tridimensional por uno de los modos (normalmente el modo K, condiciones). Así, se dispone de información de un conjunto de individuos sobre los que se han medido diferentes conjuntos de variables o un mismo conjunto de variables, medidas sobre distintos conjuntos de individuos. El desarrollo de estos métodos se centra en tres etapas principales: estudio de la interestructura, compromiso e intraestructura. La interestructura hace referencia al análisis de la relación entre las distintas tablas que conforman la matriz original mediante sus operadores (matrices de correlaciones o covariación) y de medidas de relación para comparar dichas configuraciones. Estos son los llamados coeficientes de correlación vectorial entre matrices calculados a partir del producto interno de Hilbert-Schmidt. El estudio de la estructura factorial de los operadores lleva a la etapa del compromiso, que tiene como fin encontrar una estructura común *media* que resuma la información de todas las configuraciones. Finalmente, sobre esta matriz compromiso se realiza el estudio del comportamiento de individuos y variables, así como de las relaciones entre ellos. Esta es la etapa conocida como intraestructura. Debido a que este tipo de métodos se centra en la variabilidad común explicada entre tablas, a lo largo de los años han surgido otro tipo de métodos de análisis de datos de tres vías que tratan de poner su foco de atención tanto en la variabilidad conjunta como en la variabilidad individual o específica de cada una de las matrices de partida. Con este objetivo surgen en los últimos años métodos como JIVE (Lock, Hoadley, Marron & Nobel, 2013), haciendo uso de la SVD, o ioNMF (Strazar et al., 2016) que parte de la NMF.

En el caso de los métodos anglosajones, siguen la estructura de los métodos de descomposición matricial, como la SVD, o los métodos de reducción de la dimensión, como el PCA, y es por eso por lo que se conocen como métodos de descomposición tensorial. Al igual que sus equivalentes bidimensionales, se caracterizan por tratar de ajustar modelos que reproduzcan con el mínimo error los datos originales. En este caso, ninguno de los modos es descuidado y se parte de la matriz de datos en su estructura multidimensional. Algunos de los más conocidos son el CANDECOMP/PARAFAC (Carroll & Chang, 1970; Harshman, 1970) y los modelos Tucker (Tucker, 1966), sobre los que se trabajará en el marco de esta tesis doctoral.

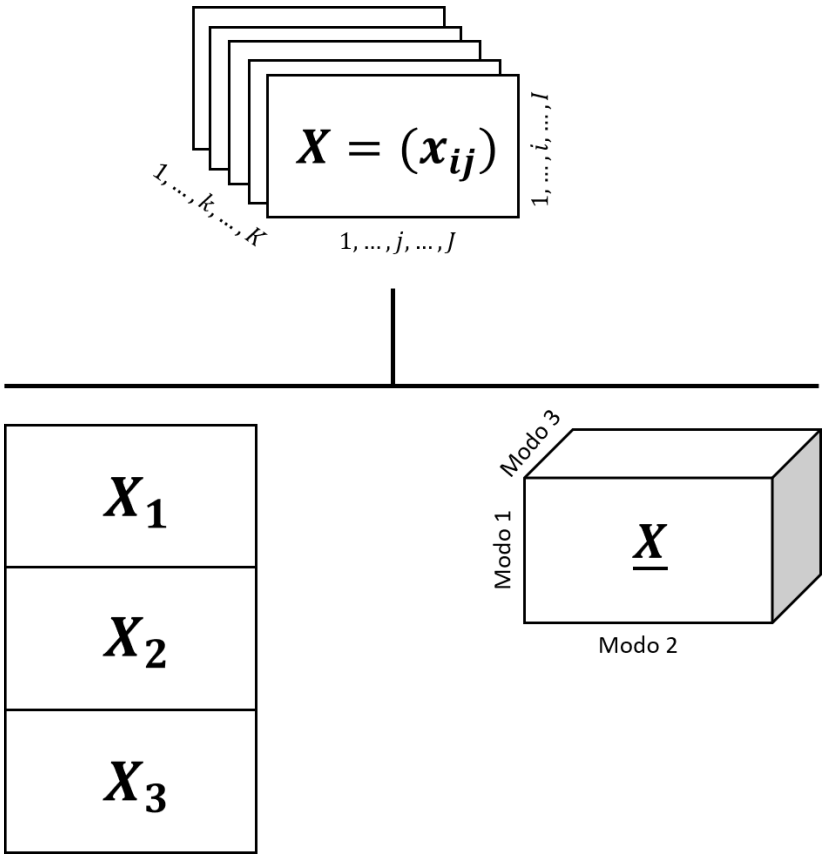


***Figura 65**. Estructura inicial de los datos para métodos de factorización de matrices de tres vías franceses (asimétricos, izquierda) y anglosajones (simétricos, derecha)*

5.1 Motivación

En el año 2016-2017 recibimos la visita del profesor y doctor en matemáticas Nickolay Trendavilov, procedente del departamento de Matemáticas y Estadística de la universidad de Londres “The Open University”, quien compartió junto a la directora de esta tesis doctoral Purificación Galindo el proyecto: “*Sparse principal component analysis in multi-way data context*” financiado por el banco Santander para la realización de una estancia del investigador en la sede del Departamento de Estadística de la Universidad de Salamanca. Su colaboración ha sido clave pues se trata del autor pionero al proponer el primer método de Sparse PCA conocido en la literatura (SCoTLASS) y extensión de los métodos Sparse a otros campos teóricos de desarrollo metodológico.

El fin principal del proyecto de investigación consistía en iniciar la investigación del Sparse PCA y sentar las bases de este en el análisis de un tipo especial de datos conocidos en estadística como datos multivía y, más recientemente en minería de datos, como datos tensoriales. Nuestro desafío consistía en desarrollar un método de Sparse PCA generalizado. La investigación sobre este tema es relativamente reciente y extremadamente desafiante debido a la falta general de una definición única de descomposición en valores singulares para los tensores.

Dicho encuentro permitió sentar las bases de la extensión del análisis de componentes principales Sparse en el contexto de análisis de datos multivía, que posteriormente daría lugar a la contribución teórica que aquí se presenta: análisis Sparse Tensores Multidimensionales, en dos vertientes bien diferenciadas: concatenación de matrices (C_{enet}JIVE) y del análisis de tensores multidimensionales y modelos C_{enet}Tucker. El Dr. Trendafilov abrió en el departamento una nueva rama de posibilidades al introducirnos en el análisis de datos de tres vías asimétrico, desde el punto de vista de los métodos que tratan de explicar la variabilidad compartida entre matrices y la variabilidad específica de cada una de ellas, como ya ocurría en el FA clásico. Así, se comenzó la revisión de un nuevo tipo de técnicas, entre las que estarían los métodos JAMMIT y JIVE.

Dentro del objetivo de este proyecto pueden tomarse dos directrices. Este algoritmo puede estar englobado dentro de las técnicas de análisis de 3 vías francesas, como el STATIS, o dentro de las inglesas, como el Tucker. El objetivo final es lograr una descomposición de tensores que tenga en cuenta todos los modos de la matriz multidimensional original; por lo tanto, el objetivo final de este proyecto consiste en la proposición de un método que se atenga, por similitud, a las definiciones de los modelos Tucker. Ahora bien, la extensión a este tipo de modelos multidimensionales no es trivial por lo que, como paso intermedio, será necesario encontrar la conexión entre los modelos bidimensionales y los tridimensionales de matrices concatenadas.

5.2 Análisis de datos de tres vías asimétrico

Una de las principales ramas de la investigación actual trata sobre el análisis de bloques de matrices de datos que comparten entre ellas alguno de sus modos, bien individuos o bien variables. Estas matrices de datos comparten parte de la información, que no sería recogida por un análisis individualizado de cada una de ellas. Tratar cada tabla de datos de manera independiente en un contexto multivía sería ineficiente.

Al hablar de técnicas de datos de tres vías asimétricas se hace referencia a aquellas metodologías que analizan una matriz de tres vías perdiendo desde su inicio la estructura tridimensional de la matriz de datos al tratar uno de los modos de manera diferente (normalmente, las ocasiones). Este tipo de técnicas parten de la matriz desdoblada por alguna de sus dimensiones. Así, podemos encontrarnos con dos situaciones diferentes. En ocasiones, los datos estarán organizados en varias matrices en las que se dispone de información de un mismo conjunto de observaciones/individuos/objetos medidos sobre diferentes conjuntos de variables. En otros casos, un mismo grupo de variables serán medidas sobre distintos conjuntos de individuos.

Este tipo de técnicas pierden la estructura de cubo de los datos de tres vías, perdiendo la información de una de las vías de la matriz tridimensional, y trabajan posteriormente usando técnicas de descomposición de dos vías sobre la matriz

desdoblada para examinar el proceso latente compartido que hay entre las distintas tablas de datos.

Los métodos asimétricos más conocidos son los métodos FMA (Escoufier & Pagès, 1984), el análisis de componentes simultáneo (SCA) (Kiers & Berge, 1994) y los métodos de la familia STATIS (Abdi et al., 2012), como el STATIS y el STATIS Dual (L'Hermier des Plantes, 1976), el CANOSTATIS (Vallejo-Arboleda, Vicente-Villardón, & Galindo-Villardón, 2007), DISTATIS (Abdi, Valentin, Chollet, & Chrea, 2007), POWER STATIS (Benasseni & Bennani-Dosse, 2012), (K+1)-STATIS (Sauzay, Hanafi, Qannari, & Schlich, 2006), … Ahora bien, la mayoría de ellos se centran en explicar la varianza compartida (lo común) entre conjuntos de datos únicamente sin tener en cuenta la información individual de las matrices de datos y esto puede ser problemático (Lock et al., 2013). Incluso puede darse la situación de trabajar con matrices de datos que no sean directamente comparables. Por este motivo surgen en la literatura otro tipo de técnicas destinadas a reflejar en una descomposición matricial la variabilidad conjunta y la variabilidad individual (van der Kloet, Sebastián-León, Conesa, Smilde, & Westerhuis, 2016; Van Deun, Smilde, Thorrez, Kiers, & Van Mechelen, 2013). En este momento es importante decir que la interpretación de los datos no solo se mejora poniendo atención en lo que es común, sino teniendo en cuenta también las partes que son específicas de cada matriz y que las hace diferentes unas de otras. En este ámbito se han propuesto distintos métodos, que incluyen variantes del CCA (Hanafi & Kiers, 2006), el análisis de componentes simultáneo con rotación para la búsqueda de componentes comunes y distintas DISCO-SCA y la descomposición en valores singulares generalizada (GSVD) (van Deun et al., 2012). Trygg (2002) propone O2-PLS, un método para el análisis de bloques de datos con observaciones comunes con el objetivo de encontrar las fuentes de variación compartida entre los bloques e información específica de cada uno de ellos. En la misma línea de separar variabilidad común y específica surgen COBE (*Common Orthogonal Basis Extraction*) (Jere et al., 2014) y JIVE (*Joint and Invidivual Variation Explained*) (Lock et al., 2013). Ambas funciones trabajan para separar variabilidad entre y dentro de cada conjunto de datos, pero se diferencian entre ellas en las matrices

con las que trabajan y en su implementación computacional. JIVE es más costoso computacionalmente que COBE.

Estas dos últimas metodologías trabajan en la extracción de la estructura factorial compartida y específica a partir de la SVD. El reemplazamiento del método de factorización por otro, como la NMF ha abierto un amplio abanico para la propuesta de otras técnicas con los mismos fines. Entre otras, puede mencionarse la iNMF (*integrative* NMF) (Yang & Michailidis, 2015) que surge como una extensión de la jNMF (*joint* NMF) (Zhang et al., 2012) que integra múltiples conjuntos de datos con un conjunto común de observaciones, buscando sus patrones comunes mediante factores latentes no negativos.

En la próxima sección se explicará con detalle la descomposición matricial para múltiples conjuntos de datos JIVE, pues fue el punto de partida del proyecto "*Sparse principal component analysis in multi-way data context*".

5.2.1 Joint and Individual Variation Explained (JIVE)

Se presenta a continuación JIVE (Lock et al., 2013): una técnica de descomposición matricial, extensión del PCA, para un conjunto de matrices que comparten la información de uno de sus modos (observaciones/variables). Se trata de una descomposición general de la variación para el análisis integrado de S múltiples conjuntos de datos. JIVE separa las componentes compartidas entre las tablas de datos y las específicas de cada una de ellas mediante la descomposición de la matriz desconcatenada en tres partes diferentes:

- Matriz común o compartida: Una aproximación de bajo nivel que capta los componentes del espacio de variabilidad compartida entre tablas.
- Matrices específicas: aproximaciones de bajo rango que capturan las componentes latentes individuales de cada una de las matrices originales.
- Ruido residual que no ha sido explicado ni mediante la matriz común o compartida, ni mediante las matrices individuales o específicas.

Una idea de lo que pretende el método puede verse en la Figura 66, donde se muestran tres matrices originales que se corresponden con tres fotografías

de cuadros. Sobre cada una de las matrices originales se ha unificado la imagen de un cuadro específica más una imagen común del cuadro “El grito” sobre todas ellas. La descomposición JIVE trata de encontrar una estructura común a las submatrices originales; esto es, “El grito” y tres matrices específicas que contienen el cuadro de partida de cada submatriz. Describimos a continuación brevemente la definición matemática del método y su notación.

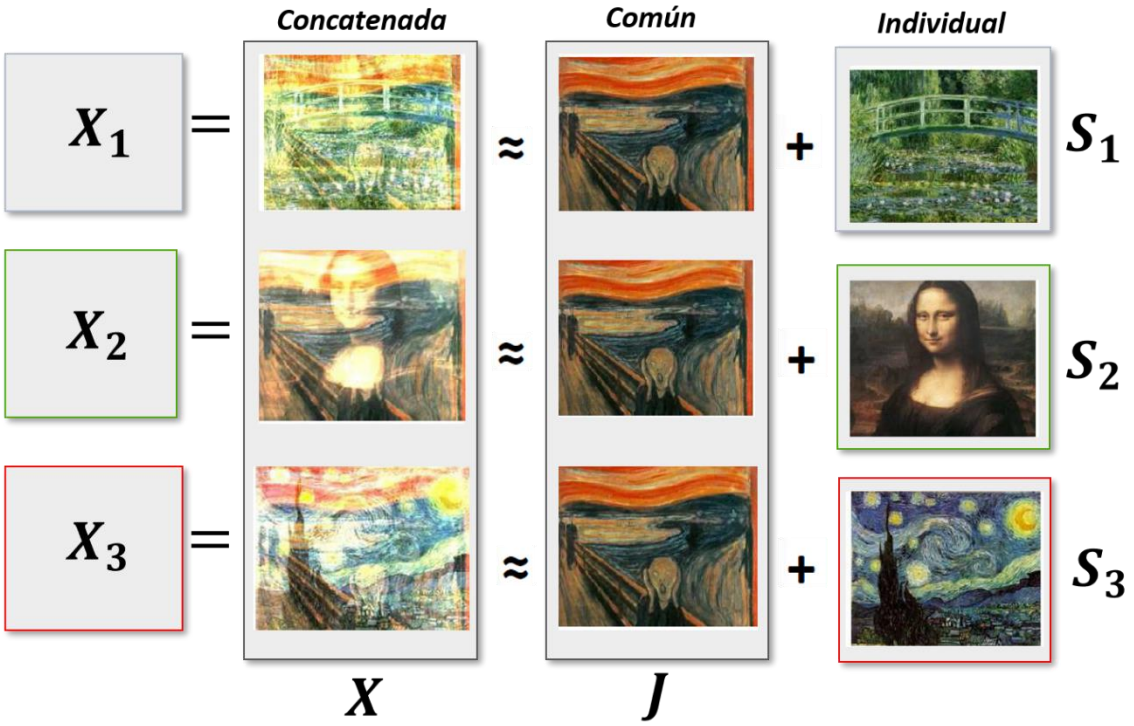


***Figura 66**. La descomposición JIVE para un conjunto mezclado de imágenes. Fuente: (Lock, 2012)*

Sea un conjunto de matrices $\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_T \in \mathbb{R}^{I_i x J}$, con las mismas J variables, pero distinto número de observaciones I_i en cada matriz $\boldsymbol{X}_i$. Se construye la matriz $\boldsymbol{X} \in \mathbb{R}^{I\,x\,J}$, con $I = \sum_{t=1}^{T} I_i$, concatenando por columnas cada una de las submatrices originales.

Es habitual que en las disciplinas en las que se trabaja con este tipo de descomposición, las matrices de partida contengan información de distintas fuentes o plataformas y no sean directamente comparables[3]. Esto hace que el preprocesamiento de las matrices de datos sea un punto trascendental que no

[3] Habitualmente esto ocurre al trabajar con datos ómicos, donde se dispone de un mismo conjunto de individuos sobre los que se han recogido datos de fuentes muy diversas: expresión génica, proteómica, …

puede pasar desapercibido. Frecuentemente se realiza la normalización individual de cada una de ellas mediante métodos como la normalización MFA, (Escofier & Pagès, 1983). Así, se asigna una misma variación total a la matriz $\boldsymbol{X}$, pues un análisis directo de ella podría ser problemático (Figura 67).

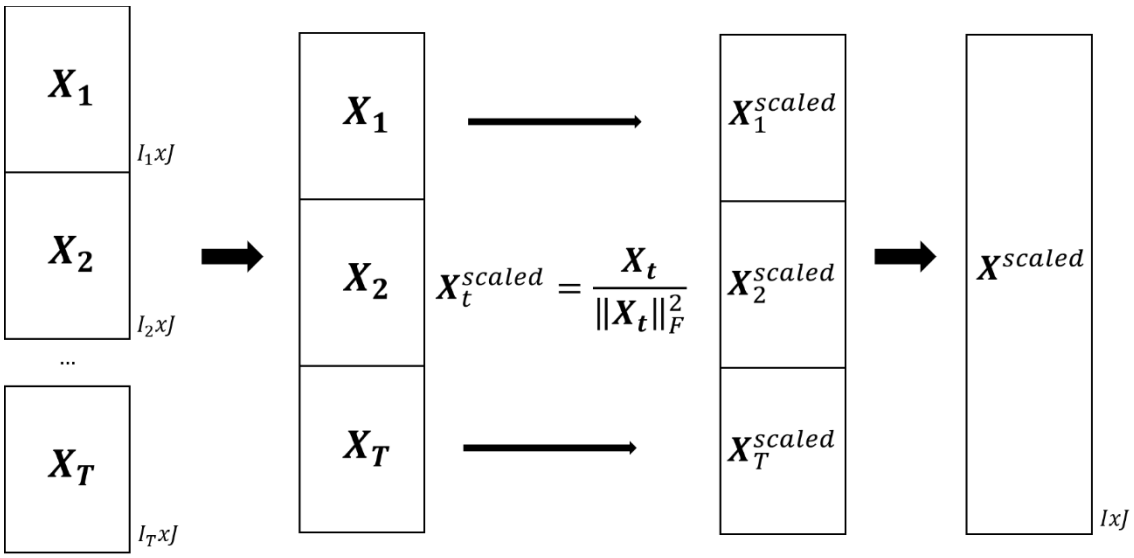


Figura 67. *Procesamiento Inicial de los datos*

JIVE plantea una descomposición general de la variabilidad de múltiples matrices integradas, compuesta por una aproximación de bajo rango de las tablas de datos originales que capture la **variación conjunta** entre matrices, aproximaciones de bajo rango para la **variación individual** de cada matriz y el **ruido residual** que aproxime de manera completa la matriz original. La estructura de variabilidad compartida es representada por una matriz $\boldsymbol{J}_{I\,x\,J}$ de rango $r < \min\{\text{rango}(\boldsymbol{X_1}), \ldots, \text{rango}(\boldsymbol{X_T})\}$ y $\boldsymbol{S}_{I\,x\,J}$ representa la estructura específica de cada matriz. El modelo de descomposición JIVE puede describirse como:

$$\boldsymbol{X}_t = \boldsymbol{J}_t + \boldsymbol{S}_t + \boldsymbol{E}_t$$

$$s.a.\ \boldsymbol{J}\boldsymbol{S}_t^T = \boldsymbol{O}$$

$\forall t = 1, \ldots, T$. La restricción de ortogonalidad asegura que las componentes latentes conjuntas y específicas estén no correlacionadas. Las matrices $\boldsymbol{J} = [\boldsymbol{J_1}, \ , \boldsymbol{J_T}]^T$ y $\boldsymbol{S} = [\boldsymbol{S_1}, \ , \boldsymbol{S_T}]^T$ representan la estructura conjunta e individual asociada a cada submatriz $\boldsymbol{X}_t$ recíprocamente. Aquí, $(\boldsymbol{J}_t)_{I_t\,x\,J}$es la submatriz de $\boldsymbol{J}$ correspondiente a $\boldsymbol{X}_t$ y $(\boldsymbol{S}_t)_{I_t\,x\,J}$ es la submatriz de rango $r_t < rango(\boldsymbol{X}_t)$ de la

estructura específica $\boldsymbol{S}$. Además, $\boldsymbol{E}_t$ representa la matrices residual para $t \in [1, T]$. Un esquema del modelo para dos matrices de partida puede verse en la Figura 68.

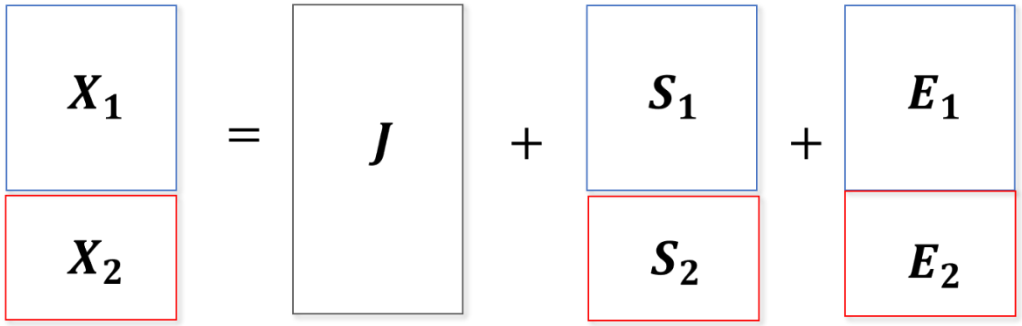


Figura 68. *Representación gráfica de la descomposición JIVE para la concatenación de dos matrices*

JIVE estima los patrones únicos y compartidos mediante un problema de optimización, minimizando la norma de Frobenius del error de reconstrucción:

$$\min \|\boldsymbol{E}\|_F^2 = \left\| \begin{matrix} \boldsymbol{E_1} \\ \boldsymbol{E_2} \\ \vdots \\ \boldsymbol{E_T} \end{matrix} \right\|_F^2 = \|\boldsymbol{X} - \boldsymbol{J} - \boldsymbol{S}\|_F^2 = \left\| \begin{matrix} \boldsymbol{X_1} - \boldsymbol{J_1} - \boldsymbol{S_1} \\ \boldsymbol{X_2} - \boldsymbol{J_2} - \boldsymbol{S_2} \\ \vdots \\ \boldsymbol{X_T} - \boldsymbol{J_T} - \boldsymbol{S_T} \end{matrix} \right\|_F^2$$

$$s.a.\ \boldsymbol{J}\boldsymbol{S}_t^T = \boldsymbol{O}$$

Este problema es resuelto mediante un algoritmo ALS, en un proceso iterativo a través de SVDs sucesivas (Lock et al., 2013).

$$\min \|\boldsymbol{E}\|_F^2 = \left\| \begin{bmatrix} \boldsymbol{X_1} \\ \boldsymbol{X_2} \\ \vdots \\ \boldsymbol{X_T} \end{bmatrix} - \boldsymbol{Q}\boldsymbol{D}\boldsymbol{U}^T - \begin{bmatrix} \boldsymbol{V_1}\boldsymbol{D_1}\boldsymbol{W_1^T} \\ \boldsymbol{V_2}\boldsymbol{D_2}\boldsymbol{W_2^T} \\ \vdots \\ \boldsymbol{V_T}\boldsymbol{D_T}\boldsymbol{W_T^T} \end{bmatrix} \right\|_F^2$$

$$s.a. \quad \boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{I}_{txt}, \boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}_{txt}, \boldsymbol{V}_t^T\boldsymbol{V}_t = \boldsymbol{I}_{txt}, \boldsymbol{Q}_t^T\boldsymbol{Q}_t = \boldsymbol{I}_{txt},$$

$$\boldsymbol{W}_1^T\boldsymbol{U} = \boldsymbol{W}_2^T\boldsymbol{U} = \cdots = \boldsymbol{W}_T^T\boldsymbol{U} = \boldsymbol{0}_{txt}$$

con $\boldsymbol{D}$, $\boldsymbol{D}_t$ matrices diagonales y r el rango de la aproximación, que puede ser diferente para I (variación individual) y J (variación conjunta). El algoritmo de optimización para el cálculo de las matrices $\boldsymbol{Q}, \boldsymbol{D}, \boldsymbol{U}, \boldsymbol{V}_t, \boldsymbol{D}_t, \boldsymbol{W}_t, t = 1, \ldots, T$, se basa en SVDs alternadas (Tabla 25).

***Tabla 25**. Algoritmo JIVE clásico (Lock et al., 2013)*

Algoritmo JIVE clásico: Joint and Individual Variation Explained	
Entrada:	$\boldsymbol{X} \in \mathbb{R}^{\boldsymbol{IxJ}}, r \in \mathbb{R}$ con $\boldsymbol{r} > 0$ rango de J y $r_t < rango(\boldsymbol{X}_t)$
Salida	$\boldsymbol{J}, \boldsymbol{S}$
Inicialización	itermax
1:	Cálculo de la SVD de **X**: $X \approx QDU^T$
2:	Definir**:** $J = QDU^T$
3:	**Mientras** $\|\boldsymbol{E}\|_F^2 > \varepsilon$ o $\text{niter} < \text{itermax}$
4:	Dividir $\boldsymbol{R} = \boldsymbol{X} - \boldsymbol{J}$ en T submatrices $\boldsymbol{R_t} \in \mathbb{R}^{I_i xJ}$
5:	**Para** $\boldsymbol{t} = 1, \dots, T$**,** cálculo de las SVDs de las matrices $\boldsymbol{R_t}(\boldsymbol{I} - \boldsymbol{UU}^T)$: $\boldsymbol{R_t}(\boldsymbol{I} - \boldsymbol{UU}^T) = \boldsymbol{V_t D_t W_t^T}$
6:	Construir $\boldsymbol{S} = [\boldsymbol{V_1 D_1 W_1^T}, \dots, \boldsymbol{V_T D_T W_T^T}]^T$
7:	$\boldsymbol{R} = \boldsymbol{R} - \boldsymbol{S}$
8:	Calcular la SVD de $\boldsymbol{X} - \boldsymbol{S} \approx \boldsymbol{QDU}^T$
9:	Definir $\boldsymbol{J} = \boldsymbol{QDU}^T$
	$\boldsymbol{J} = \boldsymbol{J}, \boldsymbol{S} = \boldsymbol{S}$
15:	**Fin**

Para ejecutar la descomposición JIVE, los autores proponen el paquete *r.jive* disponible en R (O'Connell & Lock, 2016).

Relación entre JIVE y PCA

El análisis de las matrices de variación compartida y específica puede ser abordado mediante técnicas multivariantes, las cuales reducen la dimensionalidad proyectando los datos sobre un subespacio óptimo, conservando los patrones de similaridad entre individuos y los patrones de covariación entre variables. A la hora de interpretar los resultados e identificar los patrones latentes que conforman la estructura común y cada una de las estructuras específicas, JIVE aparece combinado en la literatura con el Análisis de Componentes Principales.

En la descomposición JIVE de una matriz de tres vías concatenada se plantea la descomposición de cada una de las matrices iniciales como:

$$\boldsymbol{X}_1 \approx \boldsymbol{J}_1 + \boldsymbol{S}_1$$
$$\dots$$
$$\boldsymbol{X}_T \approx \boldsymbol{J}_T + \boldsymbol{S}_T$$

Cada una de las matrices de variabilidad conjunta $\boldsymbol{J}_1,\ldots,\boldsymbol{J}_T$ y de variabilidad individual $\boldsymbol{S}_1,\ldots,\boldsymbol{S}_T$ puede factorizarse mediante una técnica de descomposición de dos vías. La resolución del algoritmo mediante el cálculo de SVDs sucesivas permite fácilmente extender las matrices anteriores a partir de un modelo PCA. Conociendo sus vectores singulares a derecha, izquierda y los valores singulares (esto es inmediato puesto que en el algoritmo se calculan $\boldsymbol{J}_t$ y $\boldsymbol{S}_t$ a partir de SVDs) se puede escribir el modelo JIVE a partir de las cargas factoriales y puntuaciones factoriales obtenidas en las SVDs de la Tabla 25. Con ello, la descomposición JIVE puede reescribirse como:

$$\boldsymbol{X}_1 \approx \boldsymbol{J}_1 + \boldsymbol{S}_1 = \boldsymbol{F}_1\boldsymbol{M}^{\boldsymbol{T}} + \boldsymbol{W}_1\boldsymbol{Z}_1^T$$
$$\ldots$$
$$\boldsymbol{X}_T \approx \boldsymbol{J}_T + \boldsymbol{S}_T = \boldsymbol{F}_T\boldsymbol{M}^{\boldsymbol{T}} + \boldsymbol{W}_T\boldsymbol{Z}_T^T$$

donde $\boldsymbol{F}_t \in \mathbb{R}^{I_t x\, r}$ son matrices de puntuaciones factoriales de las I_t observaciones de la estructura de variabilidad común sobre las r dimensiones latentes retenidas en el PCA de $\boldsymbol{J}$ y $\boldsymbol{M} \in \mathbb{R}^{Jxr}$ matriz de cargas; es decir, de coeficientes de variables en las r componentes retenidas en el PCA sobre la matriz de variabilidad conjunta (Figura 69). Las matrices $\boldsymbol{W}_t \in \mathbb{R}^{I_t x\, r_t}$ y $\boldsymbol{Z}_t \in \mathbb{R}^{Jxr_t}$ representan las puntuaciones factoriales de las I_t observaciones y las cargas de las variables en la estructura específica sobre las r_t componentes retenidas en el PCA de las matrices de información específica $\boldsymbol{S}_t$ (Figura 69).

La relación entre JIVE y PCA permite establecer representaciones gráficas de los resultados obtenidos e interpretar las relaciones comunes y específicas visualmente sobre un espacio de dimensión latente menor.

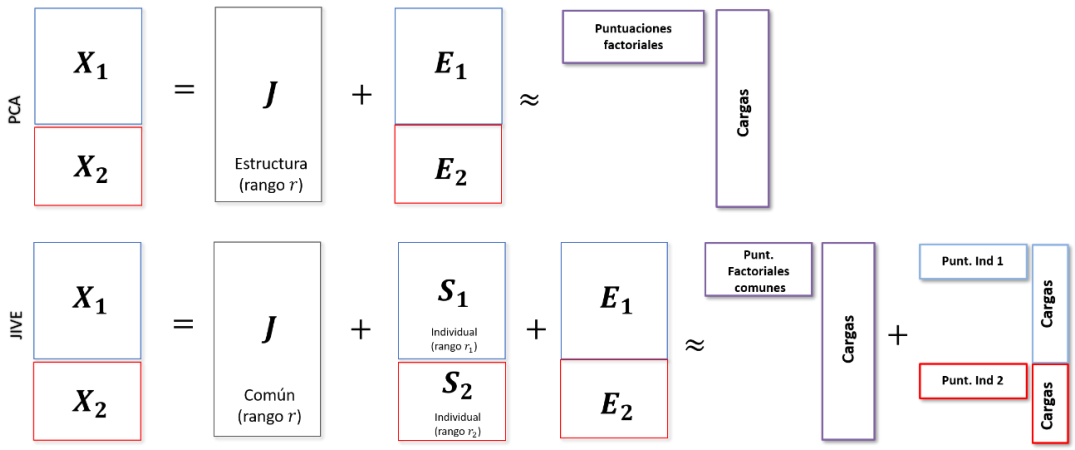


***Figura 69**. PCA vs JIVE. Fuente: (Lock, 2012)*

Una posible línea futura de esta metodología podría considerar el uso de otras técnicas de reducción de la dimensión y representación gráfica de observaciones y variables, como pueden ser los métodos Biplot.

5.2.2 Aplicación al análisis de datos reales. Contribuciones de JIVE

Una aplicación del método JIVE junto con la representación HJ-Biplot de los resultados en el contexto de contaminación química de ríos de Ecuador se presentó en el **VI Encuentro Iberoamericano de Biometría**, que tuvo lugar en noviembre de 2017 en Ecuador, con el trabajo "Caracterización de la contaminación geoquímica en ríos de Ecuador. Un análisis multivariante a través de JIVE y HJ-Biplot". Los datos que se han analizado en este estudio están relacionados con un problema clásico medioambiental: el análisis de la interacción entre elementos de un área fuertemente contaminada por compuestos metaloides.

El fin del estudio era exponer los principales aspectos de JIVE así como conocer su utilidad en el análisis de una matriz de datos químicos de una zona de ríos de Ecuador en relación con la posible contaminación debida a agentes externos. La contaminación hace referencia a la alteración del estado natural de un lugar, por ejemplo, ríos, como consecuencia de la introducción de un agente contaminante ajeno a este, como pueden ser desechos químicos provocados, por ejemplo, por una fábrica. Esta investigación se enfoca en el análisis de muestras recogidas en la zona Ponce Enriquez, una de las áreas más afectadas por la minería y la metalurgia en Ecuador, que ya desde 1980 presenta graves problemas medioambientales. Para el estudio se recogieron un total de 170 muestras procedentes de cuatro ríos de una zona cercana a Ponce Enriquez: 48 muestras del río Siete, 34 del río Fermín, 56 de Guanache y 32 de Villa. Para cada muestra se analizó la concentración de un total de 22 elementos químicos, Hierro o Plata, entre otros.

El objetivo de nuestro estudio era establecer una caracterización geoquímica de estos ríos teniendo en cuenta tanto su nivel de contaminación común, como el patrón de contaminación específico de cada río, haciendo uso

de las técnicas de estadística multivariante HJ-Biplot y método de descomposición JIVE.

Teniendo en cuenta las 4 matrices de datos a analizar, es lógico pensar que existirá una estructura química común entre los ríos y una estructura específica de cada uno de ellos. La aplicación del método JIVE permite identificar estas estructuras y, como observamos en la variabilidad recogida, en todos los ríos se observa una parte de variabilidad compartida (color azul) entorno al 30% y una variabilidad individual de cada uno de ellos que los diferencia (Figura 70). Lo primero a recalcar es que todos los ríos presentaban información propia a cada uno de ellos. Es decir, como es lógico, existe variabilidad propia del lugar de muestreo que no debía ser considerada como variabilidad compartida entre los diferentes ríos.

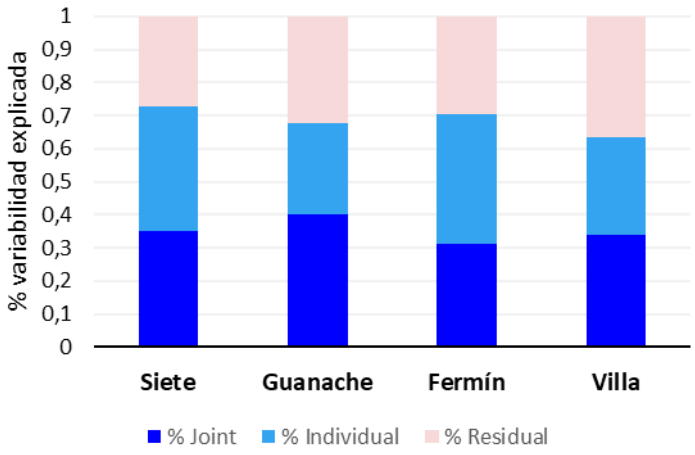


***Figura 70**. Porcentaje de varianza explicada por cada una de las componentes retenidas en la matriz de estructuras común y en las matrices de estructuras específicas*

Una vez obtenida la descomposición JIVE de la matriz original desglosada en las cuatro submatrices correspondientes a las zonas, se realizó un HJ-Biplot de las estructuras común y específicas para conocer el comportamiento de las muestras de cada río en base a una estructura geoquímica latente común y una estructura de patrones químicos individual. En la Figura 71 (izquierda) se puede observar lo que ocurre al analizar directamente la matriz de datos original concatenada mediante el HJ-Biplot, sin considerar los diferentes patrones de comportamiento entre ríos. Teniendo en cuenta estos resultados, se podría concluir que aproximadamente la mitad de las muestras presentan altos niveles

de sodio. Sin embargo, si se tiene en cuenta únicamente la variabilidad compartida entre los 4 ríos (derecha) todas las muestras presentan niveles cercanos al nivel medio en todos los contaminantes, sin verse afectados por la variabilidad de algunas variables. A diferencia del caso original, y como era lógico, el HJ-Biplot de la matriz conjunta proporciona un nivel de contaminación común entre todas las muestras analizadas. Esta matriz podría emplearse para definir unos niveles de contaminación media en la zona.

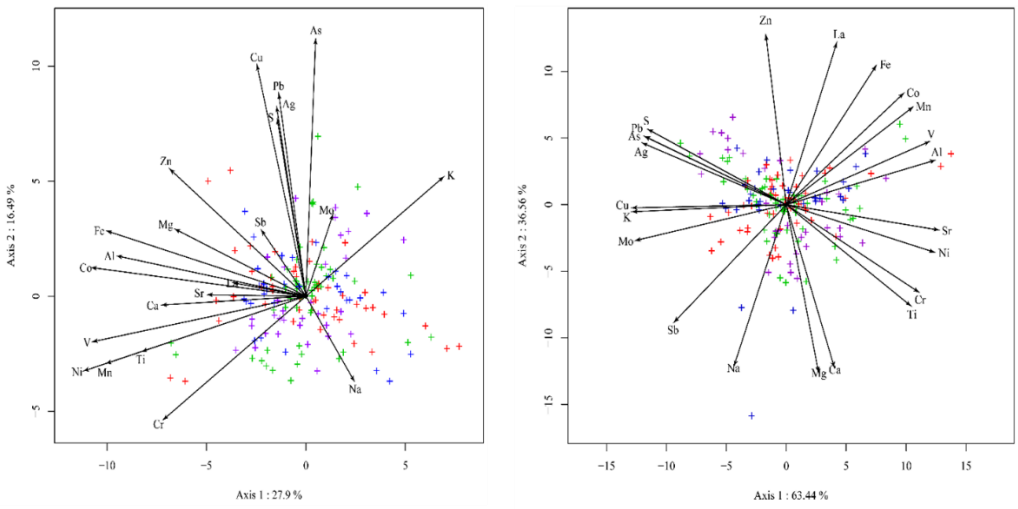


***Figura 71**. HJ-Biplot sobre la matriz concatenada original **X** (izquierda) y representación HJ-Biplot de la matriz de variabilidad común de JIVE (derecha) (rojo: Siete; azul: Fermín; verde: Guanache; morado: Villa)*

En lo respectivo a la parte específica, se puede analizar el comportamiento individual de las muestras en cualquiera de los cuatro ríos. En la Figura 72 se muestra la información de dos de ellos: el río Siete y el río Guanache. Puede observarse las diferencias que existen al analizar la matriz de datos del río siete original (izquierda) y lo que ocurre si únicamente tenemos en cuenta la variabilidad que no comparte con otras zonas de Ponce Enríquez (derecha). En el caso de la matriz inicial, puede verse una concentración mayoritaria en componentes químicos de eje 1 negativo, como pueden ser el calcio (Ca), la plata (Ag) o el magnesio (Mg) por un lado y por otro lado un alto número de muestras afectadas por altos niveles de potasio (K) y sodio (Na). Sin embargo, al analizar la variabilidad específica de este río, se vislumbra una alta correlación entre Na y K, y una concentración elevada de níquel (Ni), cobre (Cu) y cobalto (Co) de una parte de las muestras que antes no había sido diagnosticado. En el caso del río Guanache, cabe destacar en lo respectivo a la matriz de información individual una clara concentración de Ca, estroncio (Sr),

Na, Mg y azufre (S) en la mayor parte de las muestras, y bajas concentraciones de Cu, por ejemplo. Si se analiza la variabilidad específica únicamente de este río, cabe destacar correlaciones entre algunos elementos químicos que anteriormente estaban relacionados de manera inversa. Podría ser este el caso del S, K y Mg; que a diferencia del caso inicial, ahora son independientes del arsénico (As). Razonamientos similares se siguen en la interpretación de los otros dos ríos.

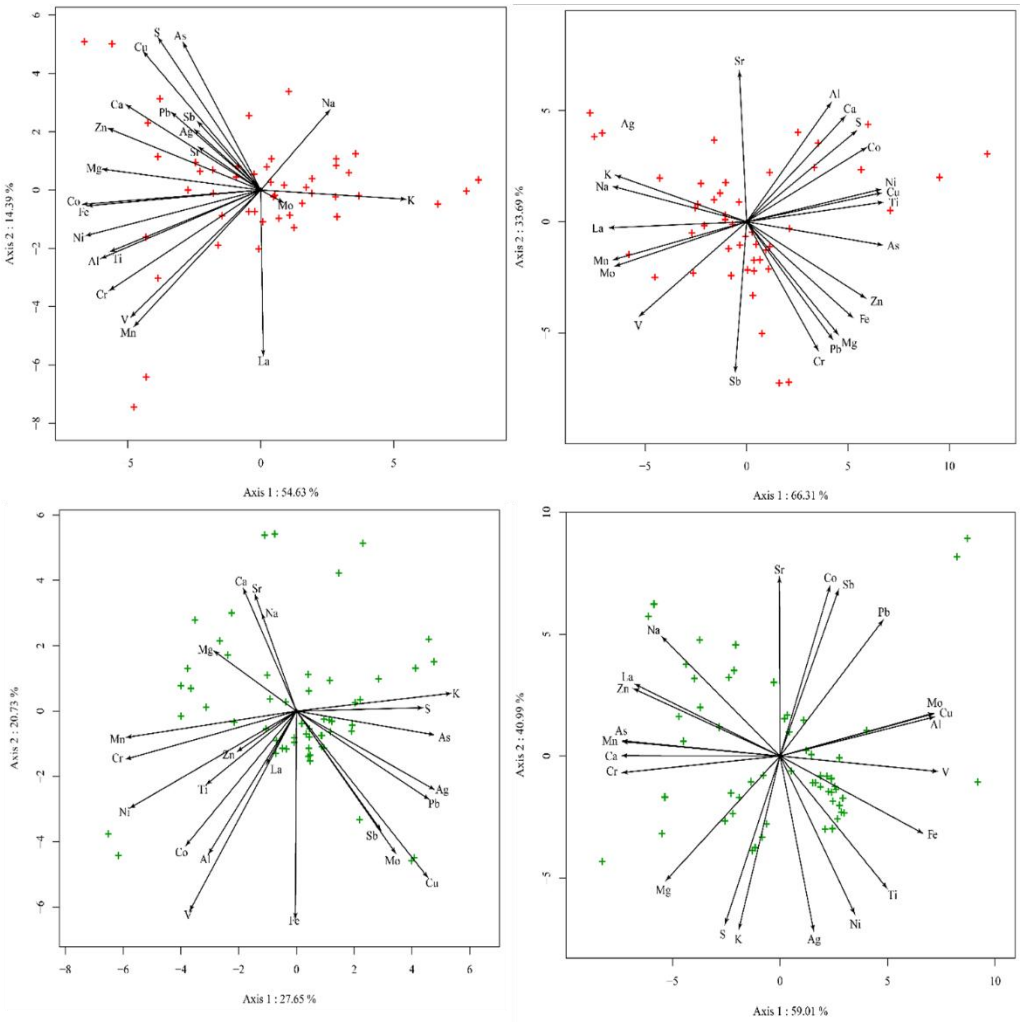


***Figura 72**. Representación HJ-Biplot de la estructura específica del río Siete (rojo) sobre la matriz de datos original (izquierda) y sobre la matriz de información específica (derecha) y del río Guanache (verde).*

JIVE permitió establecer la caracterización fisicoquímica de cuatro ríos de Ecuador, a partir de su nivel de composición química común, pudiendo servir como referencia para los estudios de línea base geo ambiental, y de la composición química **específica** de cada una de las zonas, que de otra manera

no serían detectables. La variabilidad específica es importante en términos de interpretación: cada zona puede verse afectada por diferentes agentes externos

5.2.3 Extensión de JIVE a los modelos sparse: C_{enet}JIVE

En algunas ocasiones, los patrones que explican las relaciones observaciones-variables dependen sólo de un subconjunto de las variables originales. En otras, la selección de variables con mayor contribución es deseable puesto que el análisis del modelo original es difícilmente manejable cuando el número de variables es alto. Esto motivó el uso de técnicas sparse mediante métodos de penalización en técnicas de dos vías, como el Sparse PCA (Zou et al., 2006; Shen & Huang, 2008), Sparse LDA (Clemmensen, Hastie, Witten, & Ersbøll, 2011), Sparse Biplot (Cubilla-Montilla et al., 2019), …

La importancia de los métodos sparse en dos vías ha dado lugar a su extensión también en las técnicas de tres vías. Por ello, Lock propone en su tesis doctoral la extensión de la penalización sparse a la descomposición JIVE (Lock, 2012). Formalmente, el problema se plantea como minimizar la suma cuadrada de residuales, modificando el problema original a un problema de optimización penalizado que trata de minimizar:

$$\|\boldsymbol{E}\|_F^2 + \lambda Pen(\boldsymbol{J}) + \sum_t \lambda_t Pen(\boldsymbol{S}_t)$$

La función Pen es una penalización diseñada para introducir *sparsity*. Introducir *sparsity* en el modelo quiere decir generar coeficientes nulos en los vectores de cargas de las matrices $\boldsymbol{J}$ **y** $\boldsymbol{S}_t$. Los escalares λ y λ_t toman valores positivos y controlan el grado de sparsity introducido en el modelo, bien en la estructura común o bien en las estructuras individuales. Como el modelo JIVE sparse se plantea desde el punto de vista de la optimización penalizada, cuanto mayores sean los parámetros de regularización λ y λ_t más coeficientes serán restringidos.

En el caso del sparse JIVE de Lock (2012), la función Pen se corresponde con la penalización de la norma L1 de los vectores, Lasso, ya enunciada anteriormente y cuya solución viene dada por el operador *soft-thresholding*

(Tibshirani, 1996). Pero podría ser sustituida por cualquier otra penalización que generase vectores sparse: Elastic net, hard-thresholding, SCAD, …

El procedimiento que propone Lock (2012) para encontrar la solución al problema de minimización penalizada sigue la línea del algoritmo de método JIVE ordinario: i) para $\boldsymbol{J}$ fija encontrar $\boldsymbol{S}_t$ que minimicen $\|\boldsymbol{E}_t\|_F^2 + \lambda_t Pen(\boldsymbol{S}_t), \forall t$; ii) para $\boldsymbol{S}_t$ fijas, encontrar $\boldsymbol{J}$ que minimice $\|\boldsymbol{E}\|_F^2 + \lambda Pen(\boldsymbol{J})$. Este procedimiento se repite hasta verificar el criterio de convergencia o alcanzar un número máximo de iteraciones posibles. Para implementarlo computacionalmente, Lock (2012) propone sustituir, en el algoritmo original ALS, la SVD por la sparse SVD de Lee, Shen, Huang y Marron (2010) que incorpora la penalización Lasso a los vectores singulares obtenidos, a la vez que estos son ortogonales.

En este trabajo proponemos la implementación de la C$_{enet}$SVD. Esta resuelve el mismo problema de optimización planteado en (Lock, 2012), en su versión restringida:

$$\min \|\boldsymbol{E}\|_F^2$$

$$s.\,a.: \mathrm{Pen}(\boldsymbol{J}) \leq \tau; \lambda_t \mathrm{Pen}(\boldsymbol{S}_t) \leq \tau_t \; \forall t \in [1, T]$$

con la diferencia de que no se penaliza la norma Lasso L1 de los vectores singulares de la SVD, sino que estos son restringidos a pertenecer al espacio Elastic net y ser a su vez ortogonales; es decir, pertenecer a la bola $\mathfrak{B}_{(\ell_1+\ell_2)\cap\ell_2}$. Tal y como se propone en (Lock, 2012), para llevar a cabo su programación basta con sustituir la SVD original por la SVD sparse y ortogonal, restringida al espacio Elastic net (Tabla 26). Los parámetros $\tau \in [1, (1-\alpha)\sqrt{r} + \alpha)]$ y $\tau_t \in [1, (1-\alpha_t)\sqrt{r_t} + \alpha_t)]$ (también α y α_t) pueden ser escogidos manualmente o mediante CV o algún criterio de información (AIC, BIC, GIC, …) como se ha mostrado en las metodologías de dos vías propuestas anteriormente restringidas a la región Elastic net. Si $\alpha = 0$ o $\alpha_t = 0$ para algún t, la restricción se convierte en la penalización de la norma L1.

***Tabla 26**. Algoritmo para implementar el modelo sparse C_{enet}JIVE*

Algoritmo JIVE clásico: Joint and Individual Variation Explained	
Entrada:	$\boldsymbol{X} \in \mathbb{R}^{\boldsymbol{IxJ}}, \boldsymbol{r} \in \mathbb{R}$ con $r > 0$ rango de J y $r_t <$ rango $(\boldsymbol{X}_t)$, $\alpha \in [0,1)$, $\alpha_t \in [0,1)$, $\tau \in [1, (1-\alpha)\sqrt{r} + \alpha)]$, $\tau_t \in [1, (1-\alpha_t)\sqrt{r_t} + \alpha_t)]$
Salida	$\boldsymbol{J}, \boldsymbol{S}$
Inicialización	itermax

1: Cálculo de la SVD de **X**: $X \approx QDU^T$

2: Definir**:** $\boldsymbol{J} = \boldsymbol{QDU^T}$

3: **Mientras** $\|\boldsymbol{E}\|_F^2 > \varepsilon$ o $niter < itermax$

4: Dividir $\boldsymbol{R} = \boldsymbol{X} - \boldsymbol{J}$ en T submatrices $\boldsymbol{R_t} \in \mathbb{R}^{I_i xJ}$

5: **Para** $\boldsymbol{t} = 1, \ldots, T$**,** cálculo de las C_{enet}SVDs de las matrices $\boldsymbol{R_t}(\boldsymbol{I} - \boldsymbol{UU}^T)$:

$$\boldsymbol{R_t}(\boldsymbol{I} - \boldsymbol{UU}^T) = \boldsymbol{V_t D_t W_t^T}$$

6: Construir $\boldsymbol{S} = [\boldsymbol{V_1 D_1 W_1^T}, \ldots, \boldsymbol{V_T D_T W_T^T}]^T$

7: $\boldsymbol{R} = \boldsymbol{R} - \boldsymbol{S}$

8: Calcular la C_{enet}SVD de $\boldsymbol{X} - \boldsymbol{S} \approx \boldsymbol{QDU}^T$

9: Definir $\boldsymbol{J} = \boldsymbol{QDU}^T$

$\boldsymbol{J} = \boldsymbol{J}, \boldsymbol{S} = \boldsymbol{S}$

15: **Fin**

Este desarrollo se encuentra aún en vías de programación.

5.3 Análisis de datos de tres vías simétrico

5.3.1 Conceptos introductorios

Todos los conceptos que se presentan a continuación han sido obtenidos de dos grandes referencias a nivel de análisis estadístico de tensores multidimensionales: (Cichocki et al., 2009; Kroonenberg, 2008).

Las matrices multivía son la generalización multilineal de matrices y vectores a espacios de mayor dimensión, donde los datos se organizan en tres o más direcciones. Así, podemos entender un vector de longitud I como un tensor de orden 1, perteneciente a $\mathbb{R}^I$. Una matriz de dimensión IxJ es un tensor de orden dos en $\mathbb{R}^{IxJ}$ y una matriz tridimensional (o array de tres vías) de tamaño $IxJxK$ es un tensor de orden tres perteneciente a $\mathbb{R}^{IxJxk}$. El término modo hace referencia a cada una de las dimensiones de la matriz de tres vías, de forma que el modo 1 o modo $\boldsymbol{A}$ hace referencia a la primera dimensión (filas), el modo 2 o modo $\boldsymbol{B}$ es el respectivo a la segunda dimensión (columnas) y el modo 3 o $\boldsymbol{C}$ se corresponde con el modo de las ocasiones.

Definición (Tensor). Sea $\underline{\mathbf{X}} \in \mathbb{R}^{I_1 x I_2 x \ldots x\, I_N}$ un tensor de orden N de dimensión $I_1 x I_2 x \ldots x\, I_N$, cuyos elementos se denotan por $x_{i_1,i_2,\ldots,i_n}$, con $i_1 \in I_1, i_2 \in I_2$ e $i_n \in I_N$.

En el caso particular de matrices de tres vías $\underline{\mathbf{X}} = (x_{ijk}) \in \mathbb{R}^{IxJxK}$, con $i \in \{1,\ldots,I\}, j \in \{1,\ldots,J\}$ y $k \in \{1,\ldots,K\}$ (Figura 73). Por ejemplo, el elemento x_{122} denota la información del individuo 1 en la variable 2 y condición 2.

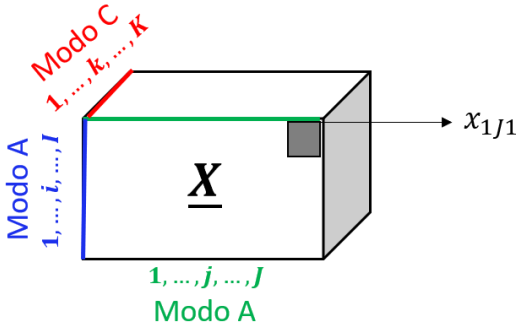


***Figura 73**. Tensor tridimensional de modos $\boldsymbol{I}, \boldsymbol{J}, \boldsymbol{K}$*

Llamaremos **fibra** de un tensor a cada uno de los fragmentos de una sola dimensión obtenidos al fijar todos los modos menos 1 (Figura 74, panel A) (dando lugar a vectores). Cada una de las columnas de una matriz se corresponde con una fibra de modo 1 (pues se fijan todos los modos a excepción de este), cada una de las filas con una fibra de modo 2 y cada uno de los tubos con una fibra de modo 3. Denotaremos por $\boldsymbol{x}_{:\,3\,1} \in \underline{\mathbf{X}}$ a la fibra referente al tercer vector columna de la primera condición (Figura 74 panel A, izquierda), por $\boldsymbol{x}_{2\,:\,1} \in \underline{\mathbf{X}}$ al vector correspondiente a la segunda fila y primera condición (Figura 74 panel A, medio) y se llama tubo al fragmento $\boldsymbol{x}_{1\,2:} \in \underline{\mathbf{X}}$, formado tras la fijación de la fila 1 y columna 2 y variación del modo 3 (Figura 74 panel A, derecha).

Por otro lado se define la **cara** de un tensor como cada uno de los fragmentos de dos dimensiones (matrices) obtenidos al fijar uno de los modos y variar los otros dos. Las distintas situaciones dan lugar a las caras horizontales, en las que se fija la primera dimensión, pero varían la segunda y la tercera; fibras laterales en las que se fija la segunda dimensión y varían la primera (filas) y la tercera (condiciones) y por último las fibras frontales, en las que se fija la tercera dimensión, pero varían la primera y la segunda. Denotaremos así por $\boldsymbol{X}_{1::} \in \underline{\mathbf{X}}$ a la cara horizontal obtenida al fijar la primera dimensión en la fila 1, por $\boldsymbol{X}_{:2:} \in \underline{\mathbf{X}}$ al fijar el segundo elemento del modo 2 y variar los elementos de los modos 1 y 3 y, por último, denotaremos la cara frontal como $\boldsymbol{X}_{::1} \in \underline{\mathbf{X}}$, la matriz obtenida al fijar la condición 1 ($k = 1$) y variar los elementos de los dos primeros modos (véase el panel B de la Figura 74).

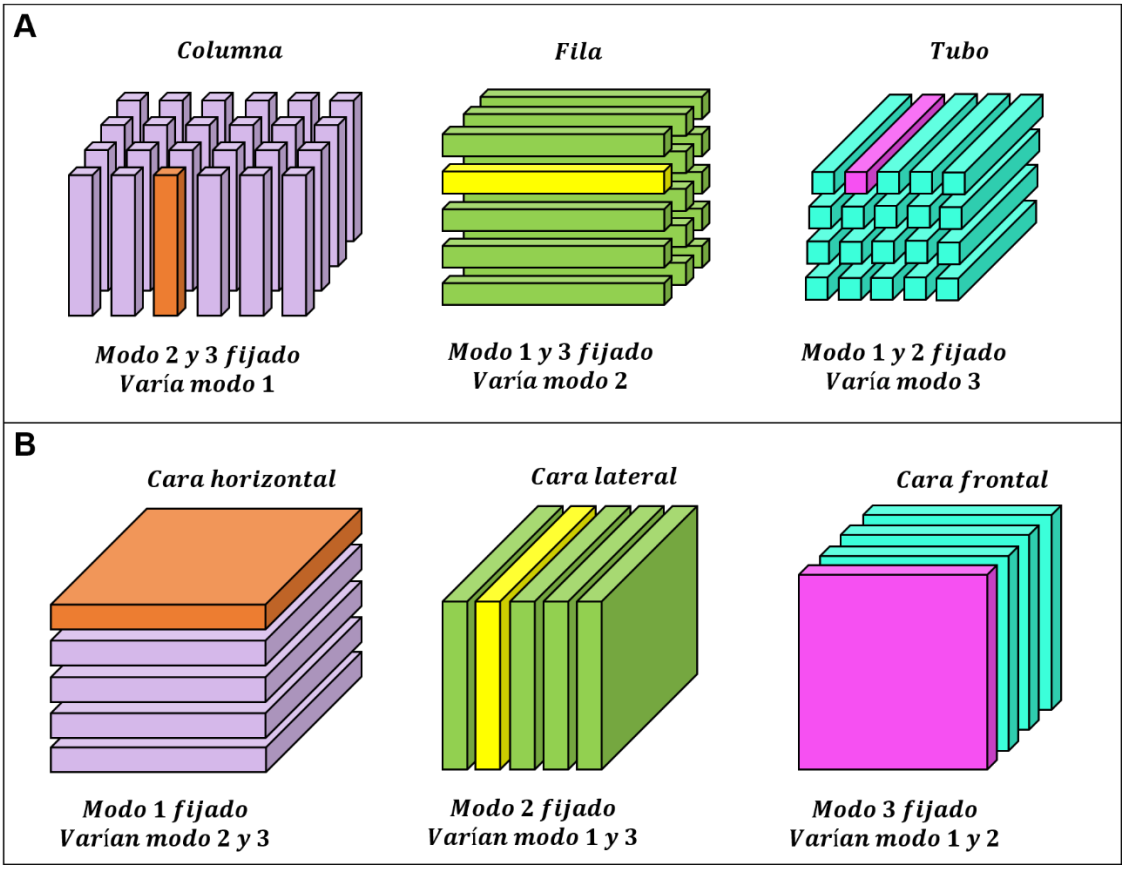


***Figura 74**. Fibras de un tensor de orden 3 (panel A) y caras de un tensor de orden 3 (panel B). Representado a partir de: (Cichocki et al., 2009)*

Operaciones con tensores.

A continuación se presentan algunas de las operaciones más útiles a la hora de trabajar con datos almacenados en tensores. Serán útiles para entender la representación de las relaciones multidimensionales entre los datos a partir de los modelos de descomposición tensorial que se presentarán más tarde.

Desdoblado o matriciación (unfolding, flattening). En diversas situaciones es conveniente representar la información de tensores mediante matrices. Esta operación de transformación de una matriz de tres vías en una de dos recibe el nombre de desdoblado o, más formalmente, **matriciación**. La matriciación de un tensor $\underline{\mathbf{X}}$ por el modo n de denota como $\boldsymbol{X}_n$ (Figura 75). El desdoblado de $\underline{\mathbf{X}}$ por su modo A se denotará como $\boldsymbol{X}_a$ y dará lugar a una matriz de tamaño $I\ x\ (JK)$, en la que se mantienen las observaciones del primer modo en filas y se concatenan por columnas cada una de las condiciones. La matriciación de $\underline{\mathbf{X}}$ por

el modo B da lugar a $\boldsymbol{X}_b \in \mathbb{R}^{J\ x\ (IK)}$, una matriz con las variables del modo 2 colocadas en filas y con los elementos de los modos 1 y 3 colocados de manera combinada en columnas. Finalmente la matriciación de $\underline{\mathbf{X}}$ por el modo C genera una matriz $\boldsymbol{X}_c \in \mathbb{R}^{K\ x\ (IJ)}$ con las condiciones colocadas en filas y las observaciones y variables en columnas.

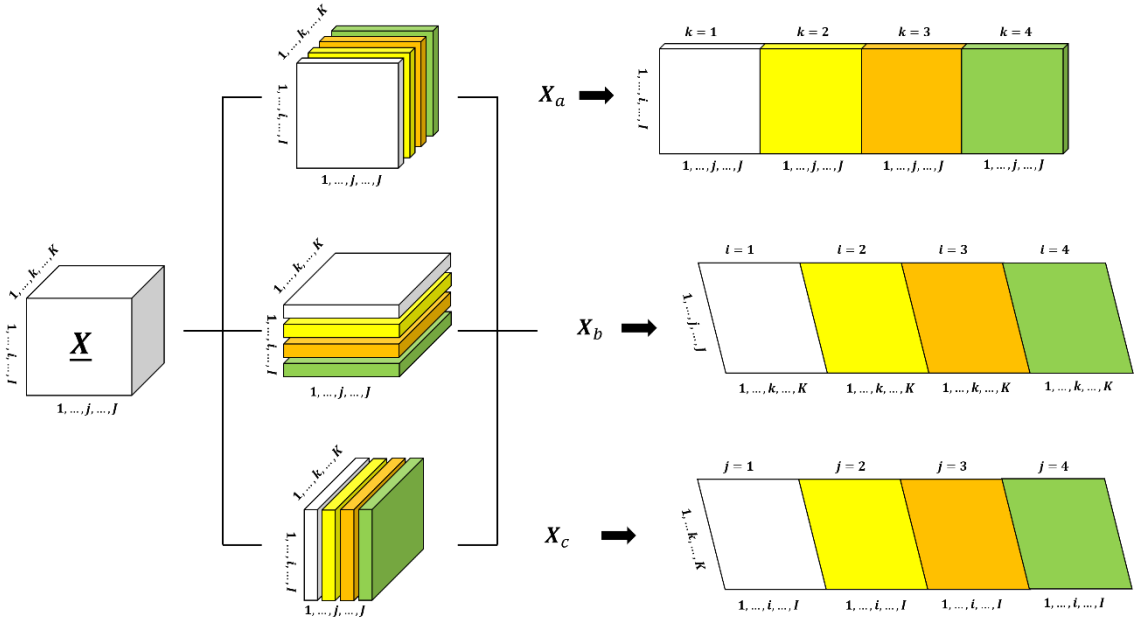


***Figura 75**. Proceso de matriciación de un tensor $\underline{\mathbf{X}}$ por cada uno de sus modos*

A continuación se muestra un ejemplo de matriciación de un tensor $\underline{\mathbf{X}} \in \mathbb{R}^{2\ x\ 3\ x\ 2}$ (Figura 76).

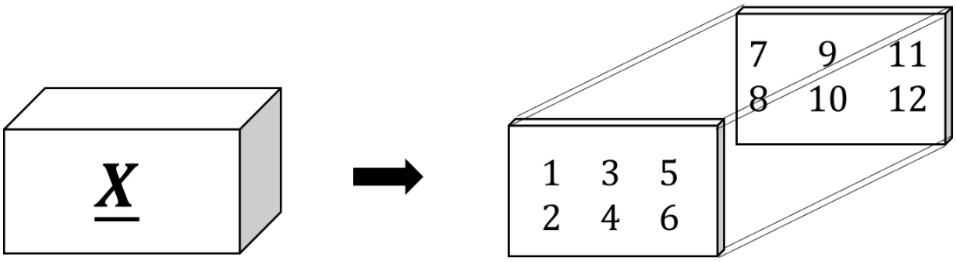


Figura 76 Ejemplo de tensor de dimensión $\boldsymbol{2\ x\ 3\ x\ 2}$

El proceso de matriciación del tensor $\underline{\mathbf{X}}$ da lugar a las matrices $\boldsymbol{X}_a, \boldsymbol{X}_b$ y $\boldsymbol{X}_c$:

$$\boldsymbol{X}_a = \begin{pmatrix} 1 & 3 & 5 & 7 & 9 & 11 \\ 2 & 4 & 6 & 8 & 10 & 12 \end{pmatrix} \in \mathbb{R}^{2\ x\ 6}$$

$$\boldsymbol{X_b} = \begin{pmatrix} 1 & 7 & 2 & 8 \\ 3 & 9 & 4 & 10 \\ 5 & 11 & 6 & 12 \end{pmatrix} \in \mathbb{R}^{3\,x\,4}$$

$$\boldsymbol{X_c} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 7 & 8 & 9 & 10 & 11 & 12 \end{pmatrix} \in \mathbb{R}^{2x6}$$

Producto de Hadamard. Dadas dos matrices $\boldsymbol{X} \in \mathbb{R}^{IxJ}, \boldsymbol{Y} \in \mathbb{R}^{IxJ}$ se define el producto de Hadamard $\boldsymbol{X} * \boldsymbol{Y}$ a partir del producto elemento a elemento de ambas matrices:

$$\boldsymbol{X} * \boldsymbol{Y} = \begin{pmatrix} \boldsymbol{x_{11} \cdot y_{11}} & \boldsymbol{x_{12} \cdot y_{12}} & \cdots & \boldsymbol{x_{1J} \cdot y_{1J}} \\ \boldsymbol{x_{21} \cdot y_{21}} & \boldsymbol{x_{22} \cdot y_{22}} & \cdots & \boldsymbol{x_{2J} \cdot y_{2J}} \\ \cdots & \cdots & \cdots & \cdots \\ \boldsymbol{x_{I1} \cdot y_{I1}} & \boldsymbol{x_{I2} \cdot y_{I2}} & \cdots & \boldsymbol{x_{IJ} \cdot y_{IJ}} \end{pmatrix}$$

Por ejemplo,

$$\boldsymbol{X} * \boldsymbol{Y} = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} * \begin{pmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 9 \end{pmatrix} = \begin{pmatrix} 1 & 8 & 21 \\ 8 & 25 & 48 \\ 21 & 48 & 81 \end{pmatrix}$$

Producto de Kronecker. Dadas dos matrices $\boldsymbol{X} \in \mathbb{R}^{IxJ}, \boldsymbol{Y} \in \mathbb{R}^{KxL}$ se define el producto de Kronecker $(\boldsymbol{X} \otimes \boldsymbol{Y})_{(IK)\,x\,(JL)}$ como:

$$\boldsymbol{X} \otimes \boldsymbol{Y} = \begin{pmatrix} \boldsymbol{x_{11} \cdot Y} & \boldsymbol{x_{12} \cdot Y} & \cdots & \boldsymbol{x_{1J} \cdot Y} \\ \boldsymbol{x_{21} \cdot Y} & \boldsymbol{x_{22} \cdot Y} & \cdots & \boldsymbol{x_{2J} \cdot Y} \\ \cdots & \cdots & \cdots & \cdots \\ \boldsymbol{x_{I1} \cdot Y} & \boldsymbol{x_{I2} \cdot Y} & \cdots & \boldsymbol{x_{IJ} \cdot Y} \end{pmatrix}$$

Por ejemplo,

$$\boldsymbol{X} \otimes \boldsymbol{Y} = \begin{pmatrix} 1 & 4 & 4 \\ 2 & 0 & 2 \\ 1 & 3 & 2 \end{pmatrix} * \begin{pmatrix} -1 & 2 & 1 \\ 3 & 1 & 0 \\ 0 & 3 & 1 \end{pmatrix} = \begin{pmatrix} -1 & 2 & 1 & -4 & 8 & 4 & -4 & 8 & 4 \\ 3 & 1 & 0 & 12 & 4 & 0 & 12 & 4 & 0 \\ 0 & 3 & 1 & 0 & 12 & 4 & 0 & 12 & 4 \\ -2 & 4 & 2 & 0 & 0 & 0 & -2 & 4 & 2 \\ 6 & 2 & 0 & 0 & 0 & 0 & 6 & 2 & 0 \\ 0 & 6 & 2 & 0 & 0 & 0 & 0 & 6 & 2 \\ -1 & 2 & 1 & -3 & 6 & 3 & -2 & 0 & 2 \\ 3 & 1 & 0 & 9 & 3 & 0 & 6 & -2 & 0 \\ 0 & 3 & 1 & 0 & 9 & 3 & 0 & 6 & 2 \end{pmatrix}$$

Producto de Khatri-Rao. El producto de Khatri-Rao se define como el producto de Kronecker por columnas. Sean $\boldsymbol{X} = (\boldsymbol{x_1}, \ldots, \boldsymbol{x_J}) \in \mathbb{R}^{IxJ}, \boldsymbol{Y} = (\boldsymbol{y_1}, \ldots, \boldsymbol{y_J}) \in \mathbb{R}^{KxJ}$ dos matrices, se define el producto $\boldsymbol{X} \odot \boldsymbol{Y}$ como

$$\boldsymbol{X} \odot \boldsymbol{Y} = (\boldsymbol{x_1} \otimes \boldsymbol{y_1}, \ldots, \boldsymbol{x_J} \otimes \boldsymbol{y_J}) \in \mathbb{R}^{ITxJ}$$

Por ejemplo,

$$\boldsymbol{X} \odot \boldsymbol{Y} = \begin{pmatrix} 1 & 4 & 4 \\ 2 & 2 & 2 \\ 1 & 3 & 2 \end{pmatrix} * \begin{pmatrix} 1 & 2 & 1 \\ 3 & 1 & 4 \\ 6 & 3 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 8 & 4 \\ 3 & 4 & 16 \\ 6 & 12 & 4 \\ 2 & 4 & 2 \\ 6 & 2 & 8 \\ 12 & 6 & 2 \\ 1 & 6 & 2 \\ 3 & 3 & 8 \\ 0 & 9 & 2 \end{pmatrix}$$

Producto de modo n de un tensor por una matriz. Para que el producto entre un tensor y una matriz pueda realizarse es necesario especificar el modo del tensor por el que se quiere multiplicar. De manera general, dado $\underline{\mathbf{X}} = (x_{ijk}) \in \mathbb{R}^{I_1 x I_2 x \ldots x\, I_N}$ un tensor de orden N e $\boldsymbol{Y} \in \mathbb{R}^{J_n x I_n}$, el producto por el modo n del tensor $\underline{\mathbf{X}}$ y la matriz $\boldsymbol{Y}$ es también un tensor $\underline{\mathbf{T}} = \underline{\mathbf{X}}\ \mathrm{X}_N\ \boldsymbol{Y}$, cuyos elementos vienen dados por el producto de los elementos:

$$t_{i_1, i_2, \ldots, i_{n-1}, j_n, i_{n+1}, \ldots, i_N} = \sum_{i_n}^{I_n} x_{i_1, i_2, \ldots, i_N} y_{j_n, i_n}$$

El lector puede encontrar un ejemplo en la Figura 77.

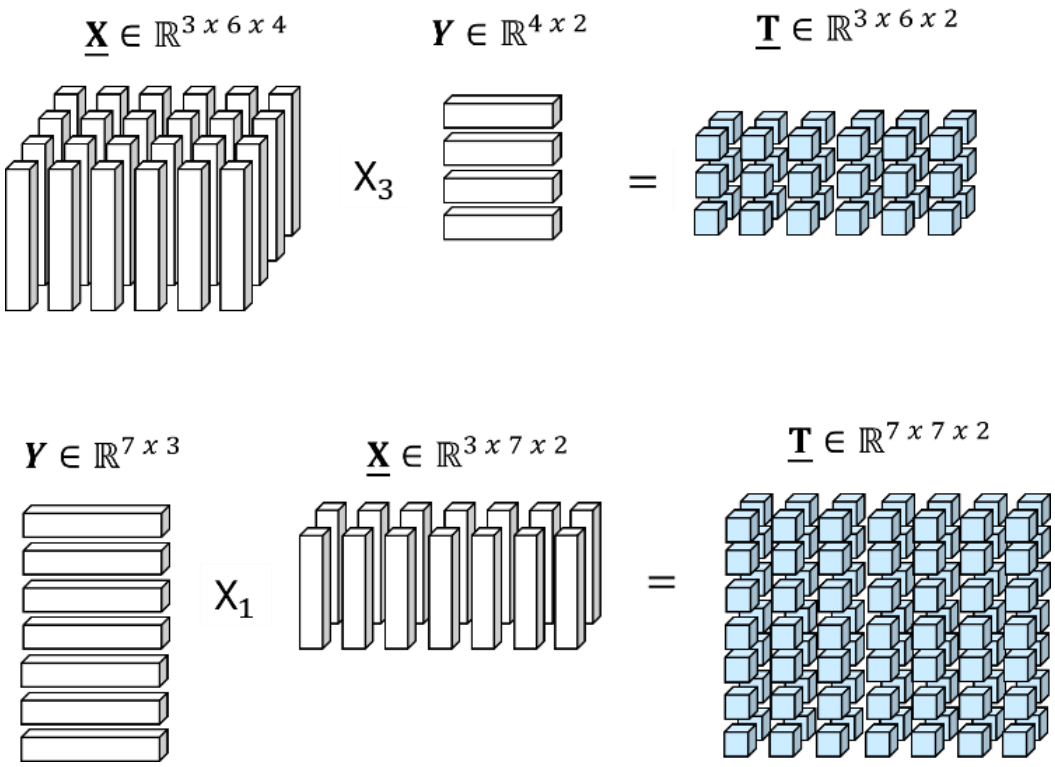


***Figura 77**. Ejemplo ilustrativo del producto tensorial de modo $\boldsymbol{n}$ de un tensor por una matriz*

Rango de un tensor. El rango R de un tensor $\underline{\mathbf{X}}$ está definido por el número de tensores de rango uno cuya combinación da lugar a $\underline{\mathbf{X}}$. Se dice que un tensor de orden 3 es de rango 1 cuando puede escribirse a partir del producto de 3 vectores (Figura 78):

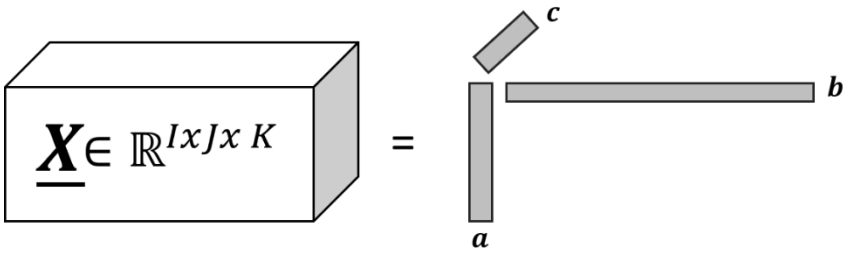


***Figura 78.** Tensor tridimensional de rango 1*

Así, el rango de un tensor $\underline{\mathbf{X}} \in \mathbb{R}^{I\,x\,J\,x\,K}$ se define como el número de tensores de rango uno, $\underline{\mathbf{X}}_1, \underline{\mathbf{X}}_2, \ldots, \underline{\mathbf{X}}_R$, tales que $\underline{\mathbf{X}} = \underline{\mathbf{X}}_1 + \underline{\mathbf{X}}_2 + \ldots + \underline{\mathbf{X}}_R$ (Kruskal, 1977). Esto es similar a lo que ocurría con la SVD de una matriz de dos vías. Por este motivo, los métodos CANDECOMP/PARAFAC son conocidos como la generalización de la SVD a datos multidimensionales. El concepto de rango de un tensor es similar al de una matriz, con la principal diferencia de que no existe un método para su obtención como ocurría en el caso matricial.

Preprocesamiento de datos

A la hora de realizar un análisis, lo más común y necesario suele ser realizar un preprocesado de los datos para unificarlos y que sean comparables, normalmente mediante el centrado y la estandarización de las variables consideradas. En el caso del análisis de 3 vías, lo mismos tipos de preprocesados deben ser tenidos en cuenta, con la puntualización de que el centrado y el escalado deben realizarse con respecto a algún modo en particular (Elisa Frutos, 2015). Así, si se desea centrar los datos a través del modo B, por ejemplo, deberá restarse a todos los datos la media obtenida para los elementos del modo B:

$$\tilde{x}_{ijk} = \tilde{x}_{ijk} - \frac{\sum_{j=1}^{J} x_{ijk}}{J}$$

Los datos pueden ser centrados por varios modos. Para ello, los datos centrados para un modo serán posteriormente centrados con respecto a otro. Este tipo de preprocesado se conoce con el nombre de doble centrado.

Por otro lado, para estandarizar las variables y eliminar diferencias debidas a las escalas de partida los datos pueden ser estandarizados. Los datos de partida serán divididos por un factor de escala relacionado con el modo con respecto al cual se quiera estandarizar. Por ejemplo, para estandarizar con respecto al modo B, la transformación a realizar sería:

$$\tilde{x}_{ijk} = \frac{\tilde{x}_{ijk}}{\sqrt{\sum_{i=1}^{I}\sum_{k=1}^{K} x_{ijk}^2 / IK}}$$

5.3.2 Modelos

A continuación se presenta la metodología asociada a dos de los modelos simétricos más relevantes e importantes: PARAFAC/CANDECOMP y modelos Tucker. Ambos pueden encontrarse en R en el paquete *ThreeWay* (Giordani, Kiers, & Del Ferraro, 2014) y en el paquete *rrcov3way* (Galloa, Todorovb, & Palmaa, 2017). Este último implementa también versiones del Tucker3 robustas (Gallo, 2015) y para datos composicionales (Engelen & Hubert, 2011).

EL método PARAFAC/CANDECOMP ha sido aplicado con éxito en diversas disciplinas. Ricci, De Gemmis y Semeraro (2012) lo emplean en los sistemas de recomendación, sistemas de filtrado de información muy utilizados en comercio. PARAFAC ha sido utilizado también con éxito para la clasificación de comida (Sádecká, Uríčková, Hroboňová, & Májek, 2015) y con datos de espectroscopia de fluorescencia para la caracterización de harinas de cereal (Lenhardt et al., 2017). Su uso en sistemas biológicos ha dado lugar a la posible caracterización nutricional de algas utilizadas como combustible renovable (Van Benthem, Lane, Davis, Lane, & Keenan, 2011) o para la identificación de impurezas orgánicas en aguas debidas a contaminantes petrolíferos (Li, Lv, & Zhang, 2013). Los métodos CP también han sido empleados para examinar la tasa de deficiencia estructural de puentes (Adarkwa, Schumacher, & Attoh-Okine, 2015). En el análisis de texto ha sido de gran utilidad para el estudio de conversaciones vía e-mail (Bader, Berry, & Browne, 2008). A nivel matemático, ha sido utilizado para la resolución de ecuaciones parciales (Zander & Matthies, 2007).

En cuanto a los modelos Tucker, estos han sido utilizados para valorar la calidad de los servicios hospitalarios (Giordani & Kiers, 2018), para el estudio de indicadores medioambientales en combinación con otras metodologías (Rodríguez-Rosa, Gallego-Álvarez, & Galindo-Villardón, 2019), calidad de agua de ríos (Singh, Malik, Singh, Basant, & Sinha, 2006) o incluso en aplicaciones de datos toxico-genómicos (Conesa, Prats-Montalbán, Tarazona, Nueda, & Ferrer, 2010). Una de las áreas en las que más beneficio ha aportado el uso de estos modelos es en el estudio de imágenes, sobre todo en el campo de la neurociencia para valorar la actividad del cerebro (Cichocki, 2013) mediante imágenes de resonancia magnética (Miyoshi et al., 2019), electroencefalografía (Cong et al., 2015)…

Otros métodos de descomposición tensorial simétrico y que no son foco de atención de la tesis incluyen INDSCAL (Carroll & Chang, 1970), CANDELINC (Carroll, Pruzansky, & Kruskal, 1980), DEDICOM (Harshman, 1978), PARAFAC2 (Harshman & Lundy, 1984b), …

PARAFAC/CANDECOMP

En el año 1970 y de manera independiente Carroll y Chang (1970) y Harshman (1970) propusieron un mismo método de descomposición tensorial que tenía su base en la SVD para matrices de dos vías, denominados CANDECOMP (*Canonical Decomposition*) y PARAFAC (*Parallel Factor Analysis*) respectivamente. Fue posteriormente Kiers, en el año 2000 (Kiers, 2000), quien quiso estandarizar la terminología y propuso CP (CANDECOMP/PARAFAC).

El método CP plantea la descomposición de un tensor $\underline{\mathbf{X}} \in \mathbb{R}^{IxJxK}$ como la suma de R tensores de rango uno. Formalmente, dado $\underline{\mathbf{X}} \in \mathbb{R}^{IxJxK}$ CP trata de definir tres matrices $\boldsymbol{A} \in \mathbb{R}^{IxR}, \boldsymbol{B} \in \mathbb{R}^{JxR}, \boldsymbol{C} \in \mathbb{R}^{KxR}$ de componentes (matrices de cargas) de manera que $\underline{\mathbf{X}}$ se factorice como:

$$\underline{\mathbf{X}} \approx \sum_{r=1}^{R} \boldsymbol{a}_r \circ \boldsymbol{b}_r \circ \boldsymbol{c}_r \equiv [[\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}]]$$

donde $\boldsymbol{A} = (\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_R), \boldsymbol{B} = (\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_R)$ y $\boldsymbol{C} = (\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_R)$ son las matrices formadas por los factores de carga de cada modo en las R componentes retenidas. En la Figura 79 puede encontrar una representación gráfica del modelo.

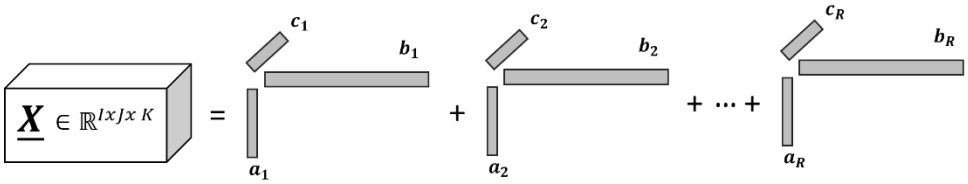


Figura 79. *Representación gráfica de la descomposición PARAFAC/TUCKER de un tensor* $\underline{\boldsymbol{X}} \in \mathbb{R}^{IxJxK}$ *como suma de* $\boldsymbol{R}$ *tensores de rango uno*

Cada uno de los elementos $x_{ijk} \in \underline{\mathbf{X}}$ viene dado por:

$$x_{ijk} = \sum_{r=1}^{R} a_{ir}\, b_{jr} c_{kr} + e_{ijk}$$

Matricialmente el modelo puede expresarse a partir del producto de modo n como:

$$\underline{\mathbf{X}} = \underline{\mathbf{\Lambda}} \times_1 \boldsymbol{A} \times_2 \boldsymbol{B} \times_3 \boldsymbol{C} + \underline{\mathbf{E}}$$

con $\underline{\mathbf{\Lambda}} \in \mathbb{R}^{RxRxR}$ es un tensor cúbico diagonal o como:

$$\boldsymbol{X}_a = \boldsymbol{A}(\boldsymbol{C} \odot \boldsymbol{B})^{\boldsymbol{T}} + \boldsymbol{E}_a$$

La solución a este problema se obtiene minimizando la función de pérdida:

$$\min_{\widehat{\underline{\boldsymbol{X}}}} \left\|\underline{\mathbf{E}}\right\|_F^2 = \left\|\underline{\mathbf{X}} - \underline{\widehat{\mathbf{X}}}\right\|_F^2$$

En la minimización de los errores es habitual utilizar el desdoblado de matrices para su resolución, de manera que, a partir de la ecuación el problema de optimización queda escrito equivalentemente como:

$$\min_{\boldsymbol{A},\boldsymbol{B},\boldsymbol{C}} \|\boldsymbol{E}_a\|_F^2 = \|\boldsymbol{X}_a - \boldsymbol{A}(\boldsymbol{C} \odot \boldsymbol{B})^{\boldsymbol{T}}\|_F^2$$

donde $\boldsymbol{X}_a$ denota el tensor $\underline{\mathbf{X}}$ desdoblado por el modo $\boldsymbol{A}$.

A pesar de que han sido propuestas diversas alternativas para su resolución, la mayoría siguen el enfoque de los algoritmos de mínimos cuadrados alternados (ALS, por sus siglas en inglés) (Tomasi & Bro, 2006), donde cada una de las matrices de componentes se calcula iterativamente. Esto quiere decir que las matrices $\boldsymbol{A}$, $\boldsymbol{B}$ o $\boldsymbol{C}$ son obtenidas tras haber fijado las dos restantes; es decir, el algoritmo encuentra la matriz $\boldsymbol{A}$ fijando $\boldsymbol{B}$ y $\boldsymbol{C}$; posteriormente, se fijarán las matrices $\boldsymbol{A}$ y $\boldsymbol{C}$ para encontrar $\boldsymbol{B}$ y, por último, la matriz $\boldsymbol{C}$ se definirá tras dejar fijas las matrices $\boldsymbol{A}$ y $\boldsymbol{B}$. El pseudocódigo del algoritmo puede verse en la Tabla 27.

***Tabla 27**. Pseudo-código del algoritmo ALS para la implementación del PARAFAC/CANDECOMP*

Algoritmo: PARAFAC/CANDECOMP	
Entrada:	$\underline{\mathbf{X}} \in \mathbb{R}^{IxJxK}$, rango R
Salida:	$\boldsymbol{A} \in \mathbb{R}^{IxR}, \boldsymbol{B} \in \mathbb{R}^{JxR}, \boldsymbol{C} \in \mathbb{R}^{KxR}$
Inicialización:	$\boldsymbol{B_0}, \boldsymbol{C_0}$ (a partir de la SVD, aleatoriamente, …)
1:	**Para** i en 1: $itermax$ hacer:
2:	$t = 0$
3:	**Mientras** $\|\boldsymbol{A_{t+1}} - \boldsymbol{A_t}\|_F^2 \geq \varepsilon$ & $\|\boldsymbol{B_{t+1}} - \boldsymbol{B_t}\|_F^2 \geq \varepsilon$ & $\|\boldsymbol{C_{t+1}} - \boldsymbol{C_t}\|_F^2 \geq \varepsilon$ hacer:
4:	$\boldsymbol{A}_{t+1} = \mathbf{X}_a[(\boldsymbol{C} \odot \boldsymbol{B})^{\boldsymbol{T}}]^{\dagger}$
5:	$B_{t+1} = \mathbf{X}_b[(\boldsymbol{C} \odot \boldsymbol{A})^{\boldsymbol{T}}]^{\dagger}$
6:	$C_{t+1} = \mathbf{X}_a[\boldsymbol{B}(\boldsymbol{C} \odot \boldsymbol{A})^{\boldsymbol{T}}]^{\dagger}$
7:	$t = t + 1$
8:	**Fin**
9:	$\boldsymbol{A} = \boldsymbol{A}_{t+1}$
10:	$\boldsymbol{B} = \boldsymbol{B}_{t+1}$
11:	$\boldsymbol{C} = \boldsymbol{C}_{t+1}$
12:	**Fin**

La descomposición CP genera la mejor aproximación de rango R de $\underline{\mathbf{X}}$, tal y como ocurría con la SVD de Eckart y Young (1936). Ahora bien, a diferencia de lo que ocurría con la SVD de dos vías, el método CANDECOMP/PARAFAC no impone restricciones de ortogonalidad sobre las matrices de cargas de cada uno de los modos $\boldsymbol{A}$, $\boldsymbol{B}$ y $\boldsymbol{C}$. Este punto es importante pues es una de las cuestiones en las que se centra nuestra posterior contribución teórica. Por otro lado, los modelos de descomposición tensorial presentan el inconveniente de que no siempre la solución que generan es única (Stegeman, Berge, & Lathauwer, 2006). En el caso del CP, esto se entiende como la única combinación de tensores de rango uno cuya suma genera el tensor original. En este caso, la unicidad de la descomposición existe bajo ciertas condiciones. Según Kruskal (Kruskal, 1989) la unicidad del modelo está íntimamente relacionada con el rango de la matriz y asegura que la condición suficiente para que exista esta es que la suma de rangos de las matrices de factores para cada modo sea menor que:

$$r_a + r_b + r_c \leq 2R + 2$$

con r_n denotando el rango de la matriz de componentes n $(n = \{a, b, c\})$. La generalización de la condición suficiente para tensores de orden n fue propuesta por (Sidiropoulos & Bro, 2000). Aportaciones adicionales y alternativas en cuanto

a la unicidad de las soluciones pueden encontrarse en (De Lathauwer, 2006; Jiang & Sidiropoulos, 2004; Stegeman & Ten Berge, 2006; Ten Berge & Sidiropoulos, 2002).

A pesar del uso exitoso de la técnica en diversas áreas del conocimiento, existen múltiples investigaciones que enuncian la dificultad del algoritmo para encontrar soluciones válidas que dan lugar a soluciones degeneradas: soluciones no interpretables y altamente correlacionadas (Giordani & Rocci, 2016). Para intentar paliar esta problemática se ha demostrado que los modelos deben ser implementados con un menor número de componentes, plantear diferentes formas de preprocesamiento de los datos o imponer diversos tipos de restricciones. Como se decía anteriormente, el CP no impone ortogonalidad sobre sus matrices de factores $\boldsymbol{A}, \boldsymbol{B}$ y $\boldsymbol{C}$ y, sin embargo, este tipo de restricción puede facilitar la obtención de soluciones no degeneradas (Harshman & Lundy, 1984a; Rocci & Giordani, 2010; Stegeman, 2007). (Harshman & Lundy, 1984a) proponen el modelo CANDECOMP/PARAFAC con restricciones de ortogonalidad (CP-Orth), resolviendo el problema de optimización:

$$\min_{\boldsymbol{A},\boldsymbol{B},\boldsymbol{C}} \|\boldsymbol{X}_a - \boldsymbol{A}(\boldsymbol{C} \odot \boldsymbol{B})^{\boldsymbol{T}}\|_F^2$$
$$s.a.\ \boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{I}$$

La restricción de ortogonalidad garantiza que el problema siempre tiene un mínimo (Krijnen, Dijkstra, & Stegeman, 2008). Al igual que esta, también se plantea la opción de imponer la restricción de no negatividad de alguno de los modos, sobre todo en casos en que datos negativos no son interpretables por el contexto tratado. Esto ha hecho que algunos investigadores hayan propuesto alternativas al CP en este sentido. Carrol, De Soete y Pruzansky (1989) proponen una versión no negativa del PARAFAC, imponiendo la restricción de no negatividad sobre las matrices de cargas; Paatero y Tapper (1994) proponen la extensión de la factorización matricial positiva (PMF, por sus siglas en inglés) de dos a tres vías y posteriormente Kim, Park, & Eldén (2007) proponen una misma metodología de factorización no negativa, pero basándose en el enfoque NLS (*Nonnegative Least Squares*). Lock y Li (2018), desde una perspectiva diferente, proponen el CP probabilístico incorporando información de covariables auxiliares que pueden ser utilizadas en la reducción de la dimensión para capturar factores latentes más precisos. Muy recientemente, en noviembre de

2019, Kim, Bismeijer, Zwart, Wessels y Vis (2019) desarrollan WON-PARAFAC, incorporando al modelo PARAFAC la restricción de ortogonalidad y no negatividad ponderada, produciendo factores sparse y mejorando la interpretabilidad de los factores latentes.

Modelos TUCKER

Los modelos de descomposición tensorial Tucker fueron introducidos por (Tucker, 1966). Al igual que el CP es entendido con la SVD generalizada a matrices de orden superior, la factorización Tucker es conocida como el PCA generalizado de orden superior. Es por este motivo que a este tipo de modelos se les conoce bajo diversos nombres: *Three-mode factor analysis* (Tucker, 1966), Three-mode PCA (Kroonenberg & Leeuw, 1980), *Higher-order SVD* (HOSVD, (De Lathauwer, De Moor, & Vandewalle, 2000))… Además, los modelos Tucker engloban principalmente tres modelos distintos, en base al número de matrices factoriales para los modos que utilizan: Tucker1, Tucker2 y Tucker3. La potencialidad de los modelos Tucker reside en la existencia de un tensor Core que permitirá examinar las relaciones resultantes de la interacción de los n modos de un array de orden n.

Dado $\underline{\mathbf{X}} \in \mathbb{R}^{IxJxK}$ y tres índices $P < I$, $Q < J$ y $R < K$ la descomposición Tucker3 trata de encontrar tres matrices de componentes $\boldsymbol{A} \in \mathbb{R}^{IxP}$, $\boldsymbol{B} \in \mathbb{R}^{JxQ}$, $\boldsymbol{C} \in \mathbb{R}^{KxR}$ y un tensor $\underline{\mathbf{G}} \in \mathbb{R}^{PxQxR}$, de manera que su producto aproxime $\underline{\mathbf{X}}$ de la mejor forma posible (Figura 80).

$$\underline{\mathbf{X}} = \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{r=1}^{R} g_{pqr}\,(\boldsymbol{a}_p \circ \boldsymbol{b}_q \circ \boldsymbol{c}_r) + \underline{\mathbf{E}} \equiv [[\underline{\mathbf{G}};\, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}]]$$

donde $\boldsymbol{A} = (\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_P), \boldsymbol{B} = (\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_Q)$, $\boldsymbol{C} = (\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_R)$ son las matrices de cargas para cada uno de los modos en sus componentes retenidas $\boldsymbol{P}, \boldsymbol{Q}$ y $\boldsymbol{R}$ respectivamente y g_{pqr} son los elementos del tensor $\underline{\mathbf{G}} = (g_{pqr})$. Equivalentemente, elemento a elemento, la descomposición puede reescribirse como:

$$x_{ijk} = \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{r=1}^{R} g_{pqr}\, a_{ip}\, b_{jq} c_{kr} + e_{ijk}$$

con $i = 1, \dots, I, j = 1, \dots, J, k = 1, \dots, K$. Y de forma matricial mediante el producto tensorial de modo n:

$$\underline{\mathbf{X}} = \underline{\mathbf{G}} \times_1 \boldsymbol{A} \times_2 \boldsymbol{B} \times_3 \boldsymbol{C} + \underline{\mathbf{E}}$$

O en notación matricial:

$$\boldsymbol{X}_a = \boldsymbol{A}\boldsymbol{G}_a(\boldsymbol{C}^T \otimes \boldsymbol{B}^T) + \boldsymbol{E}_a$$

con $\otimes$ denotando el producto de Kronecker, $\boldsymbol{X}_a \in \mathbb{R}^{IxJK}$, $\boldsymbol{G}_a \in \mathbb{R}^{PxQR}$ y $\boldsymbol{E}_a \in \mathbb{R}^{IxJK}$ las matrices desdobladas por el modo 1 asociadas a $\underline{\mathbf{X}}$, $\underline{\mathbf{G}}$ y $\underline{\mathbf{E}}$.

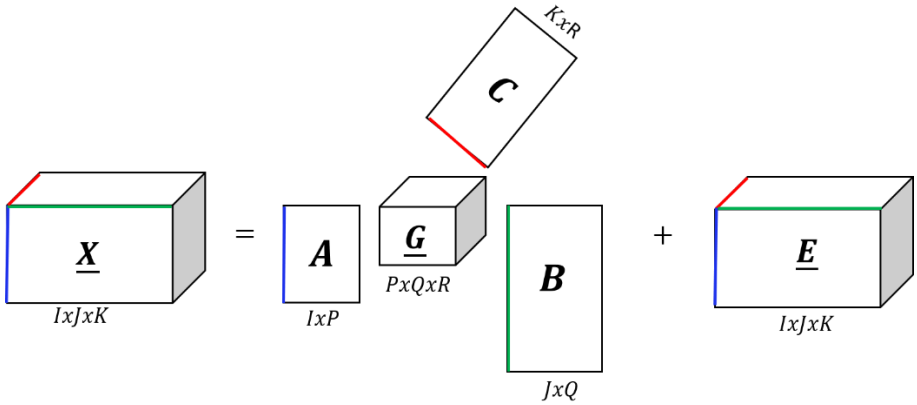


Figura 80. *Modelo de descomposición tensorial Tucker3*

A diferencia del CP, el modelo Tucker original sí impone restricciones de ortogonalidad sobre las matrices factoriales $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$, hecho que será de gran interés para nuestro trabajo. Otros autores ignoran este tipo de restricciones e imponen restricciones de no negatividad en las matrices de factores de carga, generando lo que se conoce como la descomposición Tucker no negativa (NTD, por sus siglas en inglés) (Kiers & Smilde, 1998; Kim & Choi, 2007) y que no será de interés en este documento.

El modelo Tucker1 (Figura 81, panel A) busca la descomposición del tensor $\underline{\mathbf{X}}$ a partir del producto del tensor Core $\underline{\mathbf{G}}$ y una sola matriz de factores para el modo **A**:

$$\underline{\mathbf{X}} = \underline{\mathbf{G}} \times_1 \boldsymbol{A} + \underline{\mathbf{E}} \equiv [[\underline{\mathbf{G}}; \boldsymbol{A}, \boldsymbol{I}, \boldsymbol{I}]]$$

Y en el caso del Tucker2 (Figura 81, panel B), se establecen dos matrices de componentes para los modos A y B, reescribiendo el tensor original como el producto:

$$\underline{\mathbf{X}} = \underline{\mathbf{G}} \times_1 \boldsymbol{A} \times_2 \boldsymbol{B} + \underline{\mathbf{E}} \equiv [[\underline{\mathbf{G}}; \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{I}]]$$

Los modelos Tucker1 y Tucker2 siguen la misma formulación que el modelo general Tucker3, con la diferencia de que las matrices del modo $\boldsymbol{B}$ y del modo $\boldsymbol{C}$ son consideradas la matriz identidad.

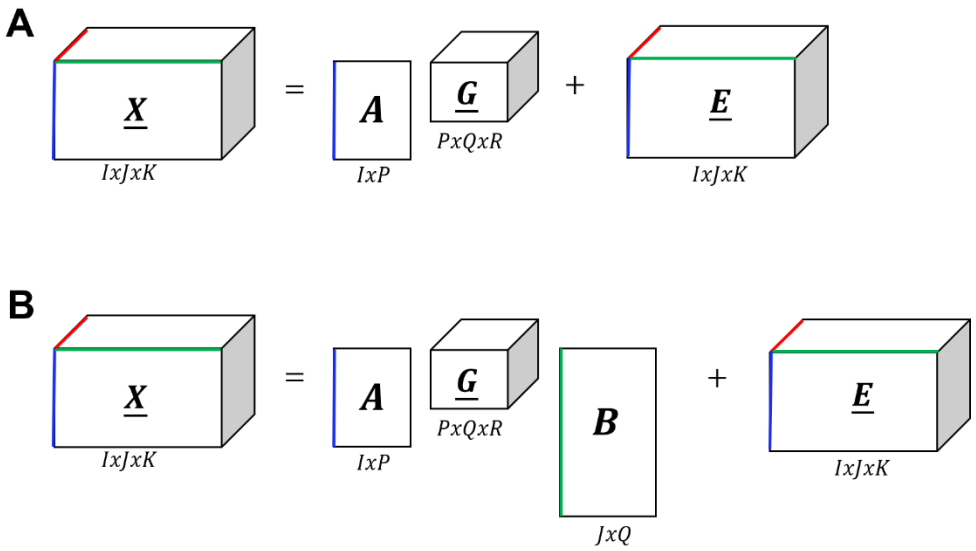


***Figura 81**. Modelo de descomposición tensorial Tucker1 (panel A) y Tucker2 (panel B).*

Por último, y a diferencia de lo que ocurría con el CP, la solución no es única; es decir, la descomposición Tucker no es única (Frutos, 2015).

ALGORITMO TUCKALS. Las soluciones propuestas por Tucker (1966) para las matrices $\boldsymbol{A}$, $\boldsymbol{B}$ y $\boldsymbol{C}$ no son estimadores mínimo cuadráticos y es por este motivo que las aproximaciones propuestas por Tucker (1966) fueron sustituidas por el algoritmo Tuckals para la obtención de $\boldsymbol{A}$,$\boldsymbol{B}$,$\boldsymbol{C}$ y $\underline{\boldsymbol{G}}$ propuesto por Kroonenberg y Leeuw (1980). Se trata de un algoritmo de mínimos cuadrados alternados nuevamente en el que la solución se obtiene minimizando la función de pérdida:

$$\min_{\hat{\underline{\mathbf{X}}}} \left\|\underline{\mathbf{E}}\right\|_F^2 = \left\|\underline{\mathbf{X}} - \hat{\underline{\mathbf{X}}}\right\|_F^2 = \|\boldsymbol{X}_a - \boldsymbol{A}\boldsymbol{G}_a(\boldsymbol{C}^T \otimes \boldsymbol{B}^T)\|_F^2$$

Partiendo de la inicialización de las matrices $\boldsymbol{A}$, $\boldsymbol{B}$ y $\boldsymbol{C}$ de manera aleatoria o mediante algún tipo de técnica de factorización matricial (Kroonenberg, 2008), en cada uno de los pasos de este algoritmo se realiza la descomposición de una matriz de dos vías mediante técnicas conocidas de análisis clásico para el

cómputo de las matrices de componentes $\boldsymbol{A}$, $\boldsymbol{B}$ y $\boldsymbol{C}$; normalmente, la descomposición en valores singulares. Una vez que el algoritmo converge y se han encontrado las matrices $\boldsymbol{A}$, $\boldsymbol{B}$ y $\boldsymbol{C}$ que proporcionan la mejor aproximación de $\underline{\mathbf{X}}$ según el modelo Tucker3 (Tabla 28), se calcula el tensor Core $\underline{\mathbf{G}}$ tal que:

$$\boldsymbol{G}_a = \boldsymbol{A}^{\boldsymbol{T}}\mathbf{X}_a(\boldsymbol{C} \otimes \boldsymbol{B})$$

En el algoritmo mostrado en la Tabla 28 para la implementación de la descomposición Tucker3, la notación $SVD(\boldsymbol{Y}, \text{rango } = R)\$\boldsymbol{Z}$ hace referencia al almacenamiento de la matriz $\boldsymbol{Z}$ obtenida en la SVD de $\boldsymbol{Y}$ de rango R, siendo la SVD de $\boldsymbol{Y}$:

$$\boldsymbol{Y}_{IxJ} = \boldsymbol{M}_{IxR}\boldsymbol{D}_{RxR}\boldsymbol{Z}^{T}_{RxJ}$$

***Tabla 28**. Pseudo-código del algoritmo Tuckals3 para la implementación del Tucker3*

Algoritmo: TUCKALS3 – TUCKER3	
Entrada:	$\underline{\mathbf{X}} \in \mathbb{R}^{IxJxK}$, rango P, Q, R, $\varepsilon \approx 0$
Salida:	$\boldsymbol{A} \in \mathbb{R}^{IxP}, \boldsymbol{B} \in \mathbb{R}^{\boldsymbol{JxQ}}, \boldsymbol{C} \in \mathbb{R}^{KxR}, \underline{\mathbf{G}} \in \mathbb{R}^{PxQxR}$
Inicialización:	$\boldsymbol{A_0}, \boldsymbol{B_0}, \boldsymbol{C_0}$ (a partir de la SVD, aleatoriamente,…)

```
1:  Para i en 1:itermax hacer:
2:      t = 0
3:      Mientras ||A_{t+1} − A_t||_F^2 ≥ ε & ||B_{t+1} − B_t||_F^2 ≥ ε & ||C_{t+1} − C_t||_F^2 ≥ ε  o
        ||X_a − (X̂_a)_{t+1}||_F^2 ≥ ε  hacer:
4:          A_{t+1} = SVD(X_a(C ⊗ B), rango = P)$U
5:          B_{t+1} = SVD(X_b(A ⊗ C), rango = Q)$U
6:          C_{t+1} = SVD(X_c(B ⊗ A), rango = R)$U
7:          t = t + 1
8:      Fin
9:  A = A_{t+1}
10: B = B_{t+1}
11: C = C_{t+1}
12: G_a = A^T X_a (C ⊗ B)
13: Fin
```

Conocido el algoritmo Tuckals para la obtención de la descomposición Tucker3, es inmediato extenderlo para la obtención de las matrices de componentes del modelo Tucker2. En el caso del Tucker2 bastaría con eliminar la línea 6 del código mostrado en la Tabla 28 y sustituir la matriz $\boldsymbol{C}$ por la matriz identidad $\boldsymbol{I}_{kxk}$ (Tabla 29).

***Tabla 29**. Pseudo-código del algoritmo Tuckals2 para la implementación del Tucker2*

Algoritmo: TUCKALS2 – TUCKER2	
Entrada:	$\underline{\mathbf{X}} \in \mathbb{R}^{IxJxK}$, rango P, Q, R, $\varepsilon \approx 0$
Salida:	$\boldsymbol{A} \in \mathbb{R}^{IxP}, \boldsymbol{B} \in \mathbb{R}^{JxQ}, \boldsymbol{C} \in \mathbb{R}^{KxR}, \underline{\mathbf{G}} \in \mathbb{R}^{PxQxR}$
Inicialización:	$\boldsymbol{A_0}, \boldsymbol{B_0}$ (a partir de la SVD, aleatoriamente,…) $\boldsymbol{C} = \boldsymbol{I}$
1:	**Para** i en 1: itermax hacer:
2:	$t = 0$
3:	**Mientras** $\|\boldsymbol{A_{t+1}} - \boldsymbol{A_t}\|_F^2 \geq \varepsilon$ & $\|\boldsymbol{B_{t+1}} - \boldsymbol{B_t}\|_F^2 \geq \varepsilon$ o $\left\|\mathbf{X}_a - (\widehat{\boldsymbol{X}}_a)_{\boldsymbol{t+1}}\right\|_F^2 \geq \varepsilon$ hacer:
4:	$\boldsymbol{A}_{t+1} = SVD(\mathbf{X}_a(\boldsymbol{C} \otimes \boldsymbol{B}), \text{rango } = P)\$\boldsymbol{U}$
5:	$\boldsymbol{B}_{t+1} = SVD(\mathbf{X}_b(\boldsymbol{A} \otimes \boldsymbol{C}), \text{rango } = Q)\$\boldsymbol{U}$
6:	$t = t + 1$
7:	**Fin**
8:	$\boldsymbol{A} = \boldsymbol{A}_{t+1}$
9:	$\boldsymbol{B} = \boldsymbol{B}_{t+1}$
10:	$\boldsymbol{G}_a = \boldsymbol{A}^{\boldsymbol{T}}\mathbf{X}_a(\boldsymbol{C} \otimes \boldsymbol{B})$
11:	**Fin**

5.3.3 Métodos de descomposición sparse de tensores

Cuando los métodos clásicos producen soluciones degeneradas, no interpretables, esta problemática puede ser resuelta imponiendo algún tipo de restricción sobre las componentes generadas, como la ortogonalidad y la no negatividad (Harshman & Lundy, 1984a; Rocci & Giordani, 2010; Stegeman, 2007). Lundy et al. (1989) evidencia que este tipo de propuestas generan soluciones con sentido. Al igual que ocurrió en las técnicas de dos vías, la ortogonalización o los métodos de rotación son dos propuestas para la clarificación de los resultados, pero en la actualidad este campo de investigación ha abierto nuevas vías, como el caso de los métodos de regularización.

A continuación se muestran los resultados más relevantes de la revisión bibliográfica de las principales técnicas de descomposición matricial tradicionales, así como de los principales desarrollos encontrados en la literatura en los últimos años de este tipo de técnicas adaptadas a las necesidades del análisis de datos de altas dimensiones, en torno al ámbito de las técnicas de regularización. La aparición en la literatura de métodos de Descomposición/Factorización Sparse de tensores es reciente y muy pocos autores han estudiado la inclusión de la penalización sparse en las matrices de componentes. Aunque el *Sparse* PCA para matrices de orden dos, así como

otras técnicas sparse, ha supuesto muchos beneficios en distintas disciplinas, se ha prestado menos atención a su equivalente tensorial. Quizá, en parte a que la técnica bidimensional sigue en proceso de desarrollo o porque el problema está cambiando continuamente debido a la falta de definición de una descomposición en valores singulares única para tensores.

Si es cierto que en el contexto de la factorización tensorial no negativa existen algunas investigaciones que plantean la opción de agregar la penalización sparse a los modelos (Liu, Liu, Wonka, & Ye, 2012; Mørup, Hansen, & Arnfred, 2008) y otros que tratan de unificar ambas penalizaciones (no negatividad y sparsity) en algunos de los factores latentes resultantes en la descomposición (Allen, 2012; Cichocki et al., 2009). Este tipo de técnicas se utilizan con propósitos de clustering en datos tensoriales.

En el contexto de descomposiciones tensoriales puramente sparse, esta propiedad ha sido reconocida como necesaria para: i) comprimir conjuntos de datos de múltiples dimensiones (desde el enfoque del almacenamiento de datos) (Kolda & Bader, 2009), ii) desechar las variables irrelevantes de un modelo de altas dimensiones y iii) facilitar la visualización de los resultados. Además, la inconsistencia del PCA a nivel asintótico en bases de datos de altas dimensiones (Johnstone & Lu, 2009a), que en el ámbito de dos vías es paliada por las técnicas sparse (Amini & Wainwright, 2008; Johnstone & Lu, 2009a), se extiende también al PCA de orden superior.

Varios autores han realizado propuestas teóricas de incorporación de Lasso a la descomposición PARAFAC/CANDECOMP. En el año 2012, Allen propone los métodos *Sparse Higher-Order DVS* (HOSVD) y s*parse CANDECOMP/PARAFAC Decomposition* (Allen, 2012). Ambos métodos, siguen las directrices de los modelos Tucker y PARAFAC respectivamente. El HODVS trata de penalizar las matrices de cada uno de los modos resultantes de la descomposición Tucker (matrices $\boldsymbol{A}, \boldsymbol{B}$ y $\boldsymbol{C}$) aplicando en su cálculo el *Sparse* PCA en lugar del PCA o SVD clásico (Tabla 30) para obtener componentes sparse en cada una de las matrices de componentes. Sin embargo, además de que Allen (2012) asegura que esta estrategia no es computacionalmente eficiente, presenta la desventaja de que las componentes sparse generadas en

el sparse HOSVD no son ortonormales puesto que no lo son en la mayor parte de los métodos sparse de dos vías (Johnstone & Lu, 2009b; Shen & Huang, 2008; Zou et al., 2006).

Tabla 30. *Algoritmo de sparse HOSVD (Allen, 2012)*

Algoritmo: Sparse HOSVD (Allen, 2012)	
Entrada:	$\underline{\mathbf{X}} \in \mathbb{R}^{IxJxK}$, rango P, Q, R
Salida:	$\boldsymbol{A} \in \mathbb{R}^{IxP}, \boldsymbol{B} \in \mathbb{R}^{JxQ}, \boldsymbol{C} \in \mathbb{R}^{KxR}, \underline{\mathbf{G}} \in \mathbb{R}^{PxQxR}$
1:	$\boldsymbol{A} \leftarrow$ Primeras P sparse PCs resultantes en el sparse PCA de $\boldsymbol{X_a}$
2:	$\boldsymbol{B} \leftarrow$ Primeras Q sparse PCs resultantes en el sparse PCA de $\boldsymbol{X_b}$
3:	$\boldsymbol{C} \leftarrow$ Primeras R sparse PCs resultantes en el sparse PCA de $\boldsymbol{X_c}$
4:	$\underline{\boldsymbol{G}} = \underline{\mathbf{X}} \times_1 \boldsymbol{A} \times_2 \boldsymbol{B} \times_3 \boldsymbol{C}$

Allen plantea un segundo método para incorporar sobre la descomposición CP la penalización sparsity. Para ello, parte del problema de optimización del modelo PARAFAC (5.1) y de su formulación equivalente (Kolda & Bader, 2009)(5.2):

$$\min_{\boldsymbol{a},\boldsymbol{b},\boldsymbol{c},d} \left\| \underline{\mathbf{X}} - d \circ \boldsymbol{a} \circ \boldsymbol{b} \circ \boldsymbol{c} \right\|_F^2 \qquad (5.1)$$
$$s.a. \quad \boldsymbol{a}^T\boldsymbol{a} = 1, \boldsymbol{b}^T\boldsymbol{b} = 1, \boldsymbol{c}^T\boldsymbol{c} = 1$$

$$\max_{\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}} \underline{\mathbf{X}} \times_1 \boldsymbol{a} \times_2 \boldsymbol{b} \times_3 \boldsymbol{c} \qquad (5.2)$$
$$s.a. \quad \boldsymbol{a}^T\boldsymbol{a} = 1, \boldsymbol{b}^T\boldsymbol{b} = 1, \boldsymbol{c}^T\boldsymbol{c} = 1$$

con $\lambda_a, \lambda_b, \lambda_c > 0$ parámetros que controlan la cantidad de penalización sparse introducida en el modelo. Uno de los algoritmos utilizados para definir la solución a este problema de optimización es el algoritmo “Tensor Power Iteration”, adaptado del algoritmo “Power Iteration” para la SVD clásica (Tabla 31). Este puede modificarse para lograr ortogonalidad de las componentes utilizando el método de Graham-Schmidt, que ortogonaliza los vectores de una matriz $\boldsymbol{A} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_P)$ modificando cada vector como sigue:

$$\boldsymbol{a}_p = \frac{(\mathrm{I} - \mathbf{A}\mathbf{A}^{\mathbf{T}})\underline{\mathbf{X}} \times_2 \mathbf{b} \times_3 \mathbf{c}}{\left\|(\mathrm{I} - \mathbf{A}\mathbf{A}^{\mathbf{T}})\underline{\mathbf{X}} \times_2 \mathbf{b} \times_3 \mathbf{c}\right\|_F^2}$$

***Tabla 31**. Método "Power Iteration" para la descomposición CP (Allen, 2012)*

Algoritmo: Power iteration para CP (Allen, 2012)	
Entrada:	$\underline{\mathbf{X}} \in \mathbb{R}^{IxJxK}$, rango R
Salida:	$\boldsymbol{A} \in \mathbb{R}^{IxR}, \boldsymbol{B} \in \mathbb{R}^{JxR}, \boldsymbol{C} \in \mathbb{R}^{KxR}$
Inicialización:	$\widehat{\underline{\mathbf{X}}} = \underline{\mathbf{X}}$
1: **Para** $r = 1, \dots, R$ **hacer:**	
Hasta que se verifique el criterio de convergencia **hacer**:	
$\boldsymbol{a}_r = \widehat{\underline{\mathbf{X}}} \times_2 \boldsymbol{b}_r \times_3 \boldsymbol{c}_r \,/ \left\| \underline{\mathbf{X}} \times_2 \boldsymbol{b}_r \times_3 \boldsymbol{c}_r \right\|_F^2$	
$\boldsymbol{b}_r = \widehat{\underline{\mathbf{X}}} \times_1 \boldsymbol{a}_r \times_3 \boldsymbol{c}_r \,/ \left\| \underline{\mathbf{X}} \times_1 \boldsymbol{a}_r \times_3 \boldsymbol{c}_r \right\|_F^2$	
$\boldsymbol{c}_r = \widehat{\underline{\mathbf{X}}} \times_1 \boldsymbol{a}_r \times_2 \boldsymbol{b}_r \,/ \left\| \underline{\mathbf{X}} \times_1 \boldsymbol{a}_r \times_2 \boldsymbol{b}_r \right\|_F^2$	
$\boldsymbol{d}_r = \underline{\mathbf{X}} \times_1 \boldsymbol{a}_r \times_2 \boldsymbol{b}_r \times_3 \boldsymbol{c}_r$	
13: $\widehat{\underline{\mathbf{X}}} = \widehat{\underline{\boldsymbol{X}}} - \boldsymbol{d}_r \circ \boldsymbol{a}_r \circ \boldsymbol{b}_r \circ \boldsymbol{c}_r$	

Allen (2012) propone modificar el problema de optimización a un problema de optimización restringido en el que penaliza la norma Lasso de la componente de cada modo:

$$\max_{\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}} \underline{\mathbf{X}} \times_1 \boldsymbol{a} \times_2 \boldsymbol{b} \times_3 \boldsymbol{c} - \lambda_a \|\boldsymbol{a}\|_1 - \lambda_b \|\boldsymbol{b}\|_1 - \lambda_c \|\boldsymbol{c}\|_1 \qquad (5.3)$$

$$s.a. \;\; \boldsymbol{a}^T\boldsymbol{a} \leq 1, \boldsymbol{b}^T\boldsymbol{b} \leq 1 \;\; \& \;\; \boldsymbol{c}^T\boldsymbol{c} \leq 1$$

Al igual que se propuso en el capítulo 4 en el planteamiento de C$_{enet}$SVD, las restricciones de igualdad se relajan para simplificar el problema de optimización. Aun relajando las restricciones, la solución de cada componente tendrá norma 1 o 0. Dado que el problema (5.1) es convexo, el problema (5.3) es cóncavo para cada componente, fijadas el resto, y por ello la función objetivo se incrementa con cada iteración. Por eso, se puede obtener su solución mediante sucesivas iteraciones hasta que se dé la convergencia a su máximo local (Allen, 2012) (Tabla 32) haciendo uso del operador *soft-thresholding* $S_\lambda(\cdot) = \text{sign}(\cdot)(|\cdot| - \lambda)_+$.

***Tabla 32**. Descomposición CP sparse (Allen, 2012)*

Algoritmo: Descomposición CP sparse (Allen, 2012)	
Entrada:	$\underline{\mathbf{X}} \in \mathbb{R}^{IxJxK}$, rango R
Salida:	$\boldsymbol{A} \in \mathbb{R}^{IxR}, \boldsymbol{B} \in \mathbb{R}^{JxR}, \boldsymbol{C} \in \mathbb{R}^{KxR}$
Inicialización:	$\widehat{\underline{\mathbf{X}}} = \underline{\mathbf{X}}$
1:	**Para** $r = 1, \dots, R$ **hacer:**
	Hasta que se verifique el criterio de convergencia **hacer**:
	$\widehat{\boldsymbol{a}}_r = \boldsymbol{S}_{\lambda_a}(\widehat{\underline{\mathbf{X}}} \times_2 \boldsymbol{b}_r \times_3 \boldsymbol{c}_r)$
	$\boldsymbol{a}_r = \begin{cases} \widehat{\boldsymbol{a}}_r / \lVert \widehat{\boldsymbol{a}}_r \rVert_F^2, & si \ \lVert \widehat{\boldsymbol{a}}_r \rVert_F^2 > 0 \\ 0, & en\ caso\ contrario \end{cases}$
	$\widehat{\boldsymbol{b}}_r = \boldsymbol{S}_{\lambda_b}(\widehat{\underline{\mathbf{X}}} \times_1 \boldsymbol{a}_r \times_3 \boldsymbol{c}_r)$
	$\boldsymbol{b}_r = \begin{cases} \widehat{\boldsymbol{b}}_r / \lVert \widehat{\boldsymbol{b}}_r \rVert_F^2, & si \ \lVert \widehat{\boldsymbol{b}}_r \rVert_F^2 > 0 \\ 0, & en\ caso\ contrario \end{cases}$
	$\widehat{\boldsymbol{c}}_r = \boldsymbol{S}_{\lambda_c}(\widehat{\underline{\mathbf{X}}} \times_1 \boldsymbol{a}_r \times_2 \boldsymbol{b}_r)$
	$\boldsymbol{c}_r = \begin{cases} \widehat{\boldsymbol{c}}_r / \lVert \widehat{\boldsymbol{c}}_r \rVert_F^2, & si \ \lVert \widehat{\boldsymbol{c}}_r \rVert_F^2 > 0 \\ 0, & en\ caso\ contrario \end{cases}$
	$\boldsymbol{d}_r = \underline{\mathbf{X}} \times_1 \boldsymbol{a}_r \times_2 \boldsymbol{b}_r \times_3 \boldsymbol{c}_r$
13:	$\widehat{\underline{\mathbf{X}}} = \widehat{\underline{\boldsymbol{X}}} - \boldsymbol{d}_r \circ \boldsymbol{a}_r \circ \boldsymbol{b}_r \circ \boldsymbol{c}_r$ #Proceso de deflación para obtener múltiples componentes

Si el objetivo es generalizar la descomposición CP sparse a otro tipo de penalizaciones, el lector puede revisar (Allen, 2012), pues se propone una generalización de la Tabla 32 a otro tipo de penalizaciones, como la penalización Group Lasso (Yuan & Lin, 2006) para incorporar la penalización sparse por grupos de variables o la restricción no-negativa, sustituyendo la función *soft-thresholding* por el operador positive-thresholding $P(\boldsymbol{x}, \lambda) = (\boldsymbol{x} - \lambda)_+$ (Allen & Maletić-Savatic, 2011).

En la misma línea, Brink-Jensen (2014) propone en su trabajo de tesis doctoral la incorporación de la restricción de la norma Lasso al modelo PARAFAC. Dado que las variables seleccionadas por Lasso podrían no ser las más importantes dentro de un conjunto de variables, Brink-Jensen (2014) propone el uso de pruebas de significación basadas en permutación para escoger las variables más relevantes. Otros como Kim, Ollila y Koivunen (2013) incorporan la penalización Lasso a las matrices de componentes. Proponen el uso del criterio BIC para la selección del parámetro de regularización. Su mayor contribución se basa en la apuesta que realizan en la inicialización de las matrices. Habitualmente, en el proceso algorítmico son generadas aleatoriamente o mediante la SVD; sin embargo, para producir unas buenas estimaciones de los factores en la descomposición tensorial proponen utilizar la

solución a un problema de CP resuelto por mínimos cuadrados alternados en el que se añade la penalización Ridge. Demuestran como su proposición origina buenos resultados en los casos en los que la estructura de los datos demanda una contracción de los coeficientes, en vez de dar lugar muchos coeficientes nulos (Kim, Ollila, & Koivunen, 2013). La búsqueda de las matrices de componentes penalizadas mediante Lasso emplean el algoritmo LARS (*Least Angle Regression*) (Efron, Hastie, Johnstone, & Tibshirani, 2004). Asimismo, en el año 2017 aparece *Tensor Truncated Power (TTP),* un método de descomposición tensorial sparse que incorpora la selección de variables en las matrices de componentes, incorporando un paso de truncado (preservar los coeficientes con las mayores magnitudes) en las etapas del método Power Iteration (Sun, Lu, Liu, & Cheng, 2017). Otros trabajos similares incluyen el de (Martínez-Montes, Sánchez-Bornot, & Valdés-Sosa, 2008) o el de Giordani y Rocci (2016) que proponen añadir la penalización Lasso sobre el problema de optimización original de la factorización CP para lograr soluciones interpretables.

Cuando el interés reside en obtener componentes sparse en presencia de datos con outliers el punto de partida se ubica en desarrollar modelos que integren simultáneamente las propiedades de sparsity y robustez frente a outliers. Kim, Ollila, Koivunen y Croux (2013) realizan el primer aporte esencial en este sentido desarrollando *CP Alternating LAD-LASSO*, que combina la regresión LAD (*Least Absolute Deviation*) y la restricción Lasso en el contexto de la descomposición CP. Su objetivo es minimizar la función objetivo que incorpora la penalización de la norma L1 como sigue:

$$\min \sum_{i=1}^{N} \left\{ \sum_{j=1}^{J} \left| x_{ij} - \boldsymbol{z}_j^T \boldsymbol{a}_i \right| + \lambda_1 \|\boldsymbol{a}_i\|_1 \right\} + \lambda_2 \|\boldsymbol{B}\|_1 + \lambda_3 \|\boldsymbol{C}\|_1$$

Desde el punto de vista de los modelos Tucker la mayor parte de las propuestas, de entre las pocas existentes, se centran en la penalización del tensor Core para clarificar las relaciones existentes entre componentes de los distintos modos. En la práctica los resultados de los modelos Tucker conllevan la interpretación de los elementos no nulos de la matriz Core sobre los que suele aplicarse el proceso de umbralización. El investigador es el que decide el umbral a partir del cual considerar los coeficientes como nulos. Con el fin de automatizar

la generación de coeficientes nulos en la Core surgen *Sparse core Tucker2* (ScTucker2) (Ikemoto & Adachi, 2016) y el algoritmo de Zubair y Wang (2013) y su método Tensor OMP para generar una descomposición Tucker sparse. En el caso del ScTucker2, se añade una penalización sobre la cardinalidad de la matriz Core.

Recientemente Ahmed, Raja y Bajwa (2019) estudian los modelos de regresión lineal en tensores estructurados haciendo uso de la descomposición Tucker sparse. Para ello, proponen un método de descomposición Tucker Sparse, planteado como un problema de optimización no convexa que resuelven mediante una variante del algoritmo del gradiente descendiente proyectado (*tensor projected gradient descent*, TPGD).

Cuando los datos presentan una estructura de grupos, es interesante utilizar penalizaciones sparse que mantengan estos patrones. Por eso Chen, He, Yokoya, & Huang (2019) presentan LRTDGS, un método de descomposición de tensores en bajo rango ponderada, regularizada por una penalización sparse con estructura de grupos, útil en el estudio de imágenes hiperespectrales. Además, diseñan un algoritmo ALM (*augmented lagrange multiplier*) para encontrar la solución del modelo de bajo rango.

La extensión en las técnicas de regularización en modelos de datos tensoriales no solo aparece en el campo de las componentes principales o la descomposición en valores singulares generalizada a orden superior. Se han propuesto técnicas sparse para descomposición de datos de tensores también en el ámbito del análisis discriminante, con la propuesta de *sparse tensor discriminant analysis* (STDA) (Lai, Xu, Yang, Tang, & Zhang, 2013). Para ello, Lai et al. (2013) transforman el problema de optimización del análisis discriminante multilineal a un problema de optimización penalizado, imponiendo restricciones sobre las normas L1 y L2 para generar subespacios discriminantes. Li, Xu, Zhou y Li (2018) extienden los modelos de regresión a la descomposición Tucker en el campo de las neuroimágenes, e imponen la penalización Lasso pero únicamente sobre el tensor Core. Por otro lado, la reducción de la dimensión está íntimamente asociada con las técnicas de agrupamiento o clustering. Papalexakis, Sidiropoulos y Bro (2013) implementan una técnica de biclustering para datos tensoriales, formulada como un método de descomposición

multidimensional restringido a partir de componentes latentes sparse obtenidas mediante la penalización Lasso. Las técnicas de biclustering tratan de buscar grupos de variables correlacionadas entre sí con un solo conjunto de filas; es decir tienden a agrupar filas y columnas de una matriz de datos simultáneamente entre sí.

5.3.4 Extensión de C$_{enet}$SVD a los modelos Tucker: Sparse&Ortogonal C$_{enet}$Tucker

Habitualmente, las matrices de componentes obtenidas en el modelo Tucker clásico, así como la matriz Core son sometidas a procesos de rotación para facilitar su interpretación. Los métodos de rotación, junto con la umbralización, son una de las alternativas más empleadas en la práctica para mejorar la interpretación de los resultados al igual que ocurría en el análisis de datos de dos dimensiones. En capítulos anteriores se ha plasmado como las técnicas de penalización han supuesto un enfoque moderno y complementario a los métodos de rotación en la mejora de la interpretación de resultados en dos vías. Nuestro interés particular aquí es extender esto mismo al análisis de datos de tres vías, proponiendo una alternativa a los métodos de rotación cuyo fin último también sea la mejora de la interpretación y/o selección de observaciones/variables/condiciones. Por ello se proponen a continuación los modelos de descomposición C$_{enet}$Tucker, cuyo objetivo principal es incorporar la penalización Elastic net sobre los vectores de cargas de las matrices de cada modo, logrando así coeficientes exactamente nulos. Hasta dónde llega nuestro conocimiento, este es uno de los primeros métodos de descomposición para datos tensoriales que implementa la restricción enet.

Dado un tensor $\underline{\mathbf{X}} \in \mathbb{R}^{IxJxK}$ de orden 3, la descomposición *sparse* Tucker3 restringida al espacio Elastic net trata de encontrar tres matrices de componentes $\boldsymbol{A_{enet}} \in \mathbb{R}^{IxP}, \boldsymbol{B_{enet}} \in \mathbb{R}^{JxQ}, \boldsymbol{C_{enet}} \in \mathbb{R}^{KxR}$ penalizadas y un tensor $\underline{\mathbf{G}} \in \mathbb{R}^{PxQxR}$, de manera que su producto aproxime $\underline{\mathbf{X}}$ de la mejor forma posible:

$$\underline{\mathbf{X}} = \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{r=1}^{R} g_{pqr}\,(\boldsymbol{a}_p \circ \boldsymbol{b}_q \circ \boldsymbol{c}_r) + \underline{\mathbf{E}} \equiv [[\underline{\mathbf{G}};\ \boldsymbol{A_{enet}}, \boldsymbol{B_{enet}}, \boldsymbol{C_{enet}}]]$$

donde $\boldsymbol{A_{enet}} = (\boldsymbol{a}_1, \boldsymbol{a}_2, \dots, \boldsymbol{a}_P), \boldsymbol{B_{enet}} = \left(\boldsymbol{b}_1, \boldsymbol{b}_2, \dots, \boldsymbol{b}_Q\right),\ \boldsymbol{C_{enet}} = (\boldsymbol{c}_1, \boldsymbol{c}_2, \dots, \boldsymbol{c}_R)$ son las matrices de cargas penalizadas para cada uno de los modos en sus componentes retenidas $\boldsymbol{P}, \boldsymbol{Q}$ y $\boldsymbol{R}$ (véase la Figura 82). La matriz $\boldsymbol{A}$ ($\boldsymbol{B}$ y $\boldsymbol{C}$, respectivamente) contiene en filas la información de las observaciones (variables, condiciones) del modo 1 del tensor $\underline{\mathbf{X}}$ (modos 2 y 3) en cada una de las P componentes *sparse* (Q, R respectivamente) retenidas para este espacio. Esto significa que cada p componente del modo 1 es una combinación de solo un subconjunto de las observaciones, en base a la cantidad de penalización introducida en este modo. De manera similar para $\boldsymbol{B}$ y $\boldsymbol{C}$, estas serán matrices con algunas de las cargas exactamente nulas, lo cual facilitará posteriormente la interpretación de su información. La penalización sobre los vectores $\boldsymbol{a}_p, \boldsymbol{b}_q, \boldsymbol{c}_r$ se introduce en el problema imponiendo una restricción sobre sus normas ℓ_2 y $\ell_1 + \ell_2$, de manera que

$$\begin{gathered}
\boldsymbol{a_p}^T \boldsymbol{a}_p = 1, \boldsymbol{a}_p{}^T \boldsymbol{a}_{p'} = 0 \quad \forall p \neq p' \\
(1-\alpha_A)\left\|\boldsymbol{a}_p\right\|_1 + \alpha_A \left\|\boldsymbol{a}_p\right\|_2^2 \leq \tau_{A,p} \\
\boldsymbol{b_q}^T \boldsymbol{b}_q = 1, \boldsymbol{b}_q{}^T \boldsymbol{b}_{q'} = 0 \quad \forall q \neq q' \\
(1-\alpha_B)\left\|\boldsymbol{b}_q\right\|_1 + \alpha_B \left\|\boldsymbol{b}_q\right\|_2^2 \leq \tau_{B,q} \\
\boldsymbol{c}_r{}^T \boldsymbol{c}_r = 1, \boldsymbol{c}_r{}^T \boldsymbol{c}_{r'} = 0 \quad \forall r \neq r' \\
(1-\alpha_C)\|\boldsymbol{c}_r\|_1 + \alpha_C \|\boldsymbol{c}_r\|_2^2 \leq \tau_{C,r}
\end{gathered} \tag{5.4}$$

No es necesario que las matrices de los modos sean restringidas simultáneamente; es decir, no es necesario penalizar las tres matrices en un mismo modelo. Esto dependerá del interés del investigador y la situación particular de cada análisis. En caso de que una matriz de componentes no sea penalizada, esta coincidirá con su respectiva en el modelo Tucker clásico sin penalización.

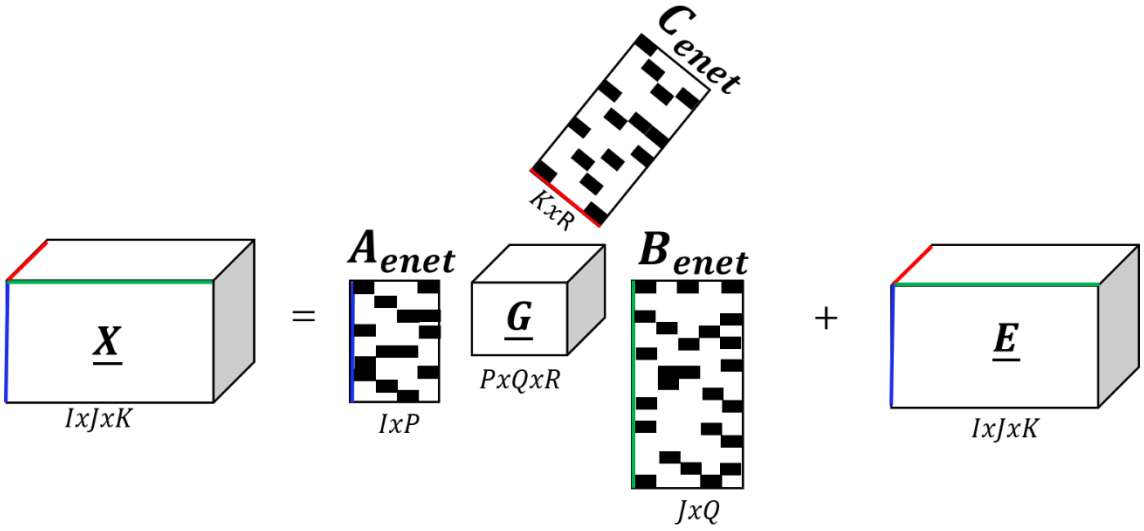


***Figura 82**. Modelo de descomposición tensorial C$_{enet}$Tucker3.*

Para el desarrollo del software, bastará con modificar las líneas 4-6 de pseudocódigo del algoritmo Tuckals3 (Tabla 28), sustituyendo el cálculo de las matrices de componentes $\boldsymbol{A}$, $\boldsymbol{B}$ y $\boldsymbol{C}$ mediante la SVD clásica por la SVD restringida C$_{enet}$SVD, sparse y ortogonal (Tabla 33). La solución se buscará nuevamente mediante un proceso iterativo. El algoritmo comienza con la inicialización de las matrices $\boldsymbol{A_0}, \boldsymbol{B_0}, \boldsymbol{C_0}$ de manera aleatoria o mediante la descomposición en valores singulares según como especifique el usuario. Así mismo, al inicio de este punto, es necesario definir los parámetros de regularización $\tau_A \in [1, (1-\alpha_A)\sqrt{I}+\alpha_A]$, $\tau_B \in [1, (1-\alpha_B)\sqrt{J}+\alpha_B]$ y $\tau_C \in [1, (1-\alpha_C)\sqrt{K}+\alpha_C]$ que controlarán el grado de *sparsity* introducido en el modelo, para algún $\alpha_A \in [0,1), \alpha_B \in [0,1)$ y $\alpha_C \in [0,1)$ definiendo el grado de penalización de Lasso y Ridge en la restricción Elastic net. Por defecto, $\alpha_A = 0{,}5 = \alpha_B = \alpha_C$. Cuanto más bajo sea el valor de α la penalización Elastic net estará conformada en mayor medida por la restricción Lasso y, consecuentemente, más cargas de la matrices de modos serán contraídas a 0. El usuario debe recordar que a menor valor de τ más coeficientes se harán nulos o casi nulos. En caso de no querer añadir ningún tipo de penalización a alguna de las matrices de carga, esto se logrará definiendo $\alpha_A = 0$ y $\tau_A = \sqrt{I}$ para el modo $\boldsymbol{A}$, $\alpha_B = 0$ y $\tau_B = \sqrt{J}$ para $\boldsymbol{B}$ y $\alpha_C = 0$ y $\tau_C = \sqrt{K}$ para $\boldsymbol{C}$. Posteriormente, y hasta que se cumpla el criterio de convergencia, se ejecutan los pasos 4-5-6 para el cálculo de $\boldsymbol{A}$, $\boldsymbol{B}$ y $\boldsymbol{C}$ mediante

la C$_{enet}$SVD para parámetros de regularización τ_A, τ_B y τ_C respectivamente. Un punto importante es que, a diferencia de lo que ocurría en las técnicas de dos vías donde el interés radicaba en penalizar la matriz de vectores singulares a izquierda y/o derecha, en este caso tan solo se penalizará la matriz $\boldsymbol{U}$ de la descomposición restringida, de vectores pseudo-singulares a izquierda.

Por último, una vez obtenidas las matrices $\boldsymbol{A}$, $\boldsymbol{B}$ y $\boldsymbol{C}$ de pseudo-vectores singulares a izquierda restringidos a la bola $\mathfrak{B}_{(\ell_1+\ell_2)\cap\ell_2}$ se calcula la matriz Core con las matrices de componentes de los modos fijadas.

Tabla 33. *Adaptación del algoritmo TUCKALS3 para la implementación del modelo C$_{enet}$Tucker3*

Algoritmo: C$_{enet}$Tucker3	
Entrada:	$\underline{\mathbf{X}} \in \mathbb{R}^{IxJxK}$, rango P, Q, R, $\varepsilon \approx 0$, $\tau_A, \alpha_A, \tau_B, \alpha_B, \tau_C, \alpha_C$
Salida:	$\boldsymbol{A} \in \mathbb{R}^{IxP}, \boldsymbol{B} \in \mathbb{R}^{JxQ}, \boldsymbol{C} \in \mathbb{R}^{KxR}, \underline{\mathbf{G}} \in \mathbb{R}^{PxQxR}$
Inicialización:	$\boldsymbol{A_0}, \boldsymbol{B_0}, \boldsymbol{C_0}$ (a partir de la SVD, aleatoriamente,…)
1:	**Para** i en 1: itermax hacer:
2:	$t = 0$
3:	**Mientras** $\|\boldsymbol{A_{t+1}} - \boldsymbol{A_t}\|_F^2 \geq \varepsilon$ & $\|\boldsymbol{B_{t+1}} - \boldsymbol{B_t}\|_F^2 \geq \varepsilon$ & $\|\boldsymbol{C_{t+1}} - \boldsymbol{C_t}\|_F^2 \geq \varepsilon$ o $\left\|\mathbf{X}_a - (\widehat{\boldsymbol{X}}_a)_{\boldsymbol{t+1}}\right\|_F^2 \geq \varepsilon$ hacer:
4:	$\boldsymbol{A}_{t+1} = C_{enet}SVD(\mathbf{X}_a(\boldsymbol{C} \otimes \boldsymbol{B}), \tau_A, \alpha_A, P)\$\boldsymbol{U}_{enet}$
5:	$\boldsymbol{B}_{t+1} = C_{enet}SVD(\mathbf{X}_b(\boldsymbol{A} \otimes \boldsymbol{C}), \tau_B, \alpha_B, Q)\$\boldsymbol{U}_{enet}$
6:	$\boldsymbol{C}_{t+1} = C_{enet}SVD(\mathbf{X}_c(\boldsymbol{B} \otimes \boldsymbol{A}), \tau_C, \alpha_C, R)\$\boldsymbol{U}_{enet}$
7:	$t = t + 1$
8:	**Fin**
9:	$\boldsymbol{A} = \boldsymbol{A}_{t+1}$
10:	$\boldsymbol{B} = \boldsymbol{B}_{t+1}$
11:	$\boldsymbol{C} = \boldsymbol{C}_{t+1}$
12:	$\boldsymbol{G}_a = \boldsymbol{A}^T\mathbf{X}_a(\boldsymbol{C} \otimes \boldsymbol{B})$
13:	**Fin**

5.3.5 Interpretación de resultados

La interpretación de los resultados se realiza de la misma manera que en los modelos Tucker clásicos. Básicamente, hay cuatro formas de presentar los resultados del análisis de tres modos: i) matrices de cargas para cada uno de los modos; ii) gráficos factoriales de pares de componentes para las matrices de componentes de cada uno de los modos; iii) Biplot interactivo, en el que se representan en un mismo gráfico las componentes de los tres modos y iv) Biplot conjunto, en el que se fijan las componentes de uno de los modos y para cada

una de ellas se representan los gráficos factoriales de los otros dos modos no fijados. A continuación se muestran algunas nociones básicas acerca de ello.

Matriz Core (G)

A diferencia de otro tipo de modelos de análisis de datos de tres vías, la potencialidad de los modelos Tucker reside en la existencia de una matriz $\boldsymbol{G}$ que permite analizar la interacción entre los tres modos del tensor simultáneamente. Sus elementos denotan los niveles de interacción entre las componentes de cada uno de los modos; es decir, las relaciones entre observaciones, variables y condiciones. Esto hace de su interpretación una cuestión fundamental. La información de la matriz Core es utilizada para analizar las relaciones entre las componentes de cada uno de los modos $(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$. Cada uno de los elementos $g_{pqr} \in \boldsymbol{G}$ contienen la información de la variabilidad explicada en conjunto por el eje p del primer modo, el eje q del segundo modo y el eje r del tercer modo. De esta manera, aquellos elementos de la matriz core cuyo valor absoluto sea elevado, equivaldrán a las componentes que más variabilidad explican conjuntamente.

Por un lado, dado un coeficiente $g_{pqr} \in \underline{\mathbf{G}}$ este es una medida del grado de correlación entre la componente p retenida del primer modo, la componente q del segundo y la componente r del tercero. La variabilidad de los datos explicada por dichas componentes viene dada por:

$$\frac{g_{pqr}^2}{\sum_{pqr} g_{pqr}^2}$$

Y la variabilidad específica explicada por cada una las componentes retenidas se calcula a partir de la suma de cuadrados de los elementos de la matriz Core obtenidos para la componente de un modo en particular, variando los índices de los otros dos modos.

A la hora de interpretar la matriz $\boldsymbol{G}$ hay que tener en cuenta dos cuestiones diferentes (Figura 83). Por un lado, el **valor absoluto de los elementos** es sinónimo de grado de variabilidad explicada por los tres factores simultáneamente y, por otro, el **signo** de cada uno de estos coeficientes, medida de la interacción entre las componentes de los distintos modos.

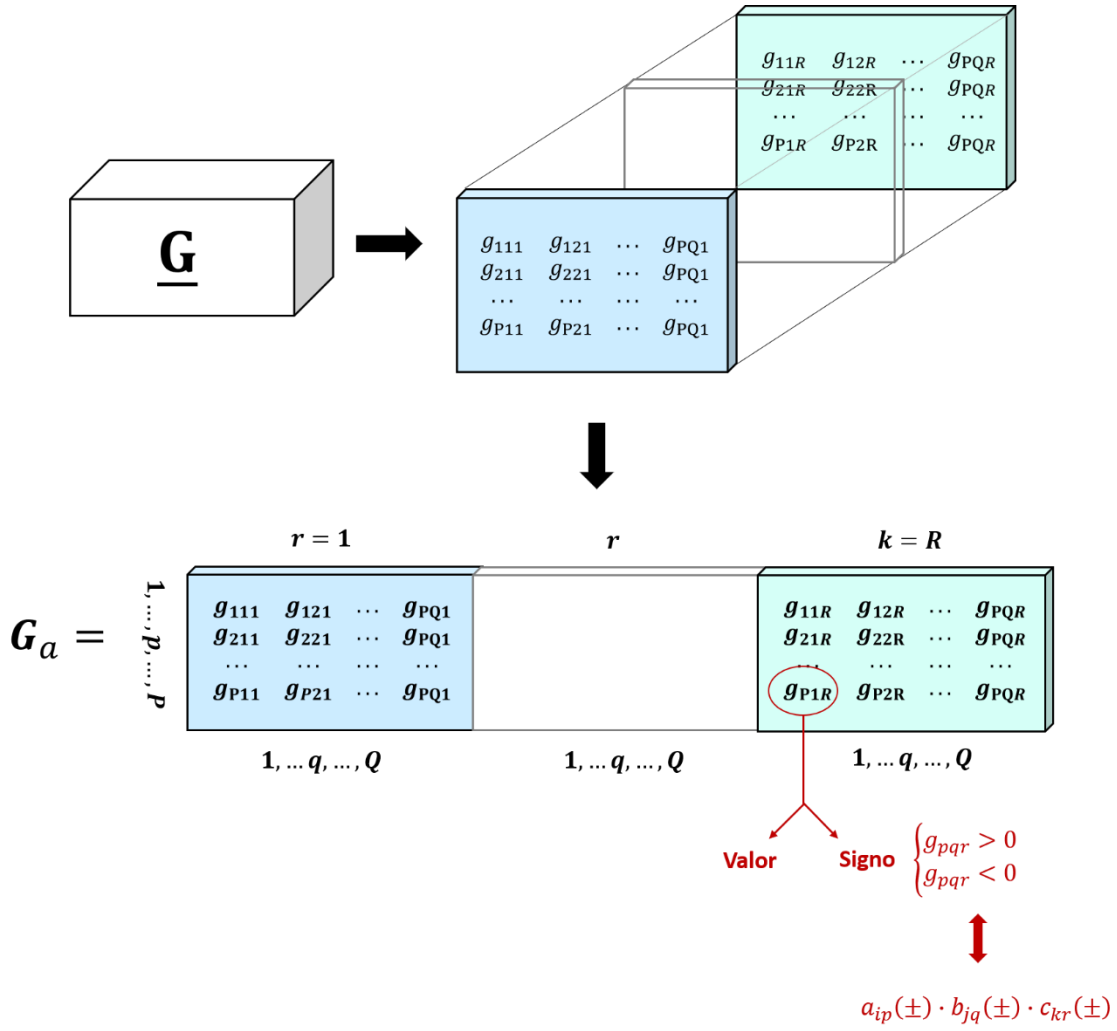


***Figura 83**. Interpretación de los elementos de la matriz Core: valor y signo*

El efecto de las componentes no depende solo de los tamaños de los coeficientes sino también de las combinaciones de signos de cada uno de los cuatro términos: $a_{ip}, b_{jq}, c_{kr}, g_{pqr}$. De la interpretación de los signos de la matriz Core hay que tener en cuenta la siguiente regla del producto de los signos de los cuatro términos anteriores para conocer el efecto global.

$$[\text{Modo } \boldsymbol{A}\ (\pm)\ \cdot\ \text{Modo } \boldsymbol{B}\ (\pm)\ \cdot\ \text{Modo } \boldsymbol{C}\ (\pm)]\ \cdot\ \text{Core}\ (\pm)\ =\ \text{Efecto (sign)}$$

El signo del coeficiente g_{pqr} en la componente p de la matriz Core de su primera dimensión, componente q de la segunda y r de la tercera, se combinará con los signos de los elementos $a_{ip} \in \boldsymbol{A}, b_{jq} \in \boldsymbol{B}$ y $c_{kr} \in \boldsymbol{C}, \forall i, j, k.$ Al escoger uno de los términos $g_{pqr} \in \underline{\mathbf{G}}$ pueden darse dos situaciones atendiendo a su signo: ser positivo o negativo.

- $g_{pqr} > 0$. Una contribución positiva global puede ser el resultado de cuatro posibles combinaciones de signos de a_{ip}, b_{jq} y c_{kr}: $\{(+,+,+),(+,-,-),(-,-,+),(-,+,-)\}$ junto con el elemento positivo de la matriz Core (Tabla 34). El signo $\pm$ se refiere al signo positivo (o negativo) de los coeficientes de la componente p del modo $\boldsymbol{A}$, de la componente q del modo $\boldsymbol{B}$ y de la componente r del modo $\boldsymbol{C}$.
- $g_{pqr} < 0$. Una interacción negativa global en el modelo puede ser el resultado de cuatro posibles combinaciones de signos de a_{ip}, b_{jq} y c_{kr} con el elemento negativo de la matriz Core: $\{(-,-,-),(+,-,+),(-,+,+),(+,+,-)\}$ (Tabla 34). El signo $\pm$ de las ternas anteriores se refiere al signo positivo (o negativo) de los coeficientes de la componente p del modo $\boldsymbol{A}$, de la componente q del modo $\boldsymbol{B}$ y de la componente r del modo $\boldsymbol{C}$.

***Tabla 34**. Interpretación de los signos de los elementos de la matriz Core*

Elemento	**Modo A -** a_{ip}	**Modo B -** b_{jq}	**Modo C -** c_{kr}	**Efecto / Contribución**
	Sign	Sign	Sign	
$g_{pqr} > 0$	+	+	+	Positivo
	+	-	-	
	-	-	+	Los tres modos (A,B,C) están
	-	+	-	relacionados de forma positiva
$g_{pqr} < 0$	-	-	-	Negativo
	-	+	+	
	+	-	+	Los tres modos están relacionados
	+	+	-	negativamente

Para facilitar la interpretación de los niveles de interacción es habitual aplicar métodos de rotación sobre la matriz Core resultante. Estos tienen el objetivo de transformar la estructura factorial de la misma para simplificar su configuración. Frecuentemente, esto hace que muchos de sus elementos sean eliminados y por lo tanto la interacción entre dichas componentes nula. En caso de no ser así, en la práctica es habitual obtener una interpretación de aquellos elementos de la matriz Core superiores a un determinado umbral e interpretar únicamente aquellos con mayores pesos; es decir, los que más contribuyen a la

relación entre las tres componentes. Esta forma de actuar coincide con la umbralización aplicada sobre los factores de carga de los métodos de dos vías.

En nuestro caso en particular, a diferencia de otros métodos Sparse, la matriz Core no es sometida a penalización directamente; pero si indirectamente al penalizar las matrices de componentes que la conforman. Es por este motivo, y como se verá en la sección posterior, que muchos de sus elementos se anularán de manera automática o serán prácticamente cero y, en este trabajo, no será de interés el post-procesamiento de las matrices mediante técnicas de rotación.

Representaciones Biplot: Biplot interactivo y Biplots conjuntos

Los métodos Biplot permiten representar gráficamente una matriz de dos vías en un plano bidimensional mediante el uso de dos matrices de marcadores (una para los elementos fila u observaciones y otra para los elementos columna o variables). En el caso de la descomposición Tucker los datos de partida se descomponen a partir de tres matrices de marcadores, ya que la matriz inicial es una matriz tridimensional (observaciones, variables y condiciones). De manera directa, los métodos Biplot no son aplicables. Por este motivo, Carlier y Kroonenberg (1996) proponen dos alternativas para que los datos se ajusten a un método Biplot: Biplot interactivo y Biplots conjuntos. Su objetivo es la representación simultánea de las tres matrices de marcadores $\boldsymbol{A}, \boldsymbol{B}$ y $\boldsymbol{C}$.

Biplot interactivo. Consiste en obtener dos matrices de marcadores $\boldsymbol{A}$ y $\boldsymbol{D}$ a partir de las tres del modelo Tucker mediante la combinación de dos de ellas (J, K) como sigue:

$$x_{ijk} \approx \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{r=1}^{R} g_{pqr}\, a_{ip}\, b_{jq} c_{kr} = \sum_{r=1}^{R} a_{ip} \left(\sum_{q=1}^{Q}\sum_{r=1}^{R} g_{pqr}\, b_{jq} c_{kr} \right) = \sum_{p=1}^{P} a_{ip} d_{(jq)p}$$

De forma matricial:

$$\boldsymbol{D} = \boldsymbol{G}_a (\boldsymbol{C} \otimes \boldsymbol{B})^T$$

Una vez obtenidas las dos matrices de marcadores, ambas se representan en un plano bidimensional donde una de las matrices de marcadores aparecerá representada mediante puntos y la respectiva a la combinación de los dos modos

mediante vectores (Figura 84). El número de ejes retenidos para la representación dependerá del número de componentes retenidas en el primer modo.

Los Biplot interactivos son útiles cuando la tercera dimensión, respectiva a las condiciones, hace referencia a distintas medidas de tiempo, pues entonces tendría sentido realizar trayectorias correspondientes a la posible evolución. También son aconsejables al trabajar con matrices en las que no haya un número muy grande de elementos en los modos que se concatenan (Carlier & Kroonenberg, 1996). En la aplicación de datos reales que se mostrará más adelante este Biplot no tiene sentido por esta segunda razón.

Biplot conjunto. A diferencia del Biplot interactivo donde se representa la información de las componentes de todos los modos de forma conjunta en un solo gráfico, en el caso del Biplot conjunto se realiza un Biplot condicionado a uno de los modos. En caso de fijar el modo 3 (condiciones), se realizará una representación Biplot para cada una de las componentes retenidas en el modo $\boldsymbol{C}$. Esto es, en caso de que la matriz de componentes del modo 3 se haya calculado reteniendo tres componentes, se obtendrán tres Biplots conjuntos (uno para cada componente retenida en el modo $\boldsymbol{C}$). Las observaciones y las variables son representadas en un mismo gráfico interpretable, gracias a la factorización Biplot, proyectadas sobre la información de tan solo una de las componentes del modo $\boldsymbol{C}$ (condiciones) (Figura 84).

Para ello, en primer lugar se construyen las matrices $\boldsymbol{D}_r$ con $r = 1, \dots R$:

$$\boldsymbol{D}_r = \boldsymbol{A}\boldsymbol{G}_r\boldsymbol{B}^T \in \mathbb{R}^{IxJ}$$

con $\boldsymbol{G}_r$ la capa del tensor $\underline{\mathbf{G}}$ correspondiente a la componente r del modo $\boldsymbol{C}$. Sobre las matrices $\boldsymbol{D}_r$ se calcula la factorización Biplot haciendo uso de la factorización matricial SVD clásica:

$$\boldsymbol{D}_r = \boldsymbol{M}\boldsymbol{\Lambda}\boldsymbol{P}^T$$

y se definen las matrices de marcadores fila y columna que posteriormente son representadas en el correspondiente plano factorial. A la hora de interpretar los resultados en cada uno de los r Biplots conjuntos es muy importante tener en cuenta que, en el estudio simultáneo de las relaciones de los tres modos se hará

uso de las cargas obtenidas en la matriz C de la descomposición Tucker. Esto es así porque en cada Biplot conjunto r se obtendrán conclusiones de las relaciones entre observaciones, variables y condiciones, pero las últimas tan solo se interpretarán aquellas con cargas de coeficientes altos en la componente r del modo $\boldsymbol{C}$. Además, será de especial importancia el signo de las cargas de $\boldsymbol{c}_r$ en la interpretación de las relaciones de los tres modos.

Por ejemplo, si la categoría k del tercer modo tiene en la componente r una carga de signo positivo, entonces en la representación Biplot aquellos marcadores fila y columna que estén próximos estarán directamente relacionados con la condición k. Sin embargo, si la carga c_{kr} de la categoría k en la componente r es negativa, entonces aquellos marcadores fila y columna que en el gráfico se encuentren próximos interactuarán de manera negativa con la condición k del tercer modo.

A pesar de que hay más representaciones gráficas que en el caso anterior, estas son más fáciles de interpretar en aquellos casos en que las categorías de los modos son elevadas y cuyas conclusiones no se podían obtener a través del Biplot interactivo.

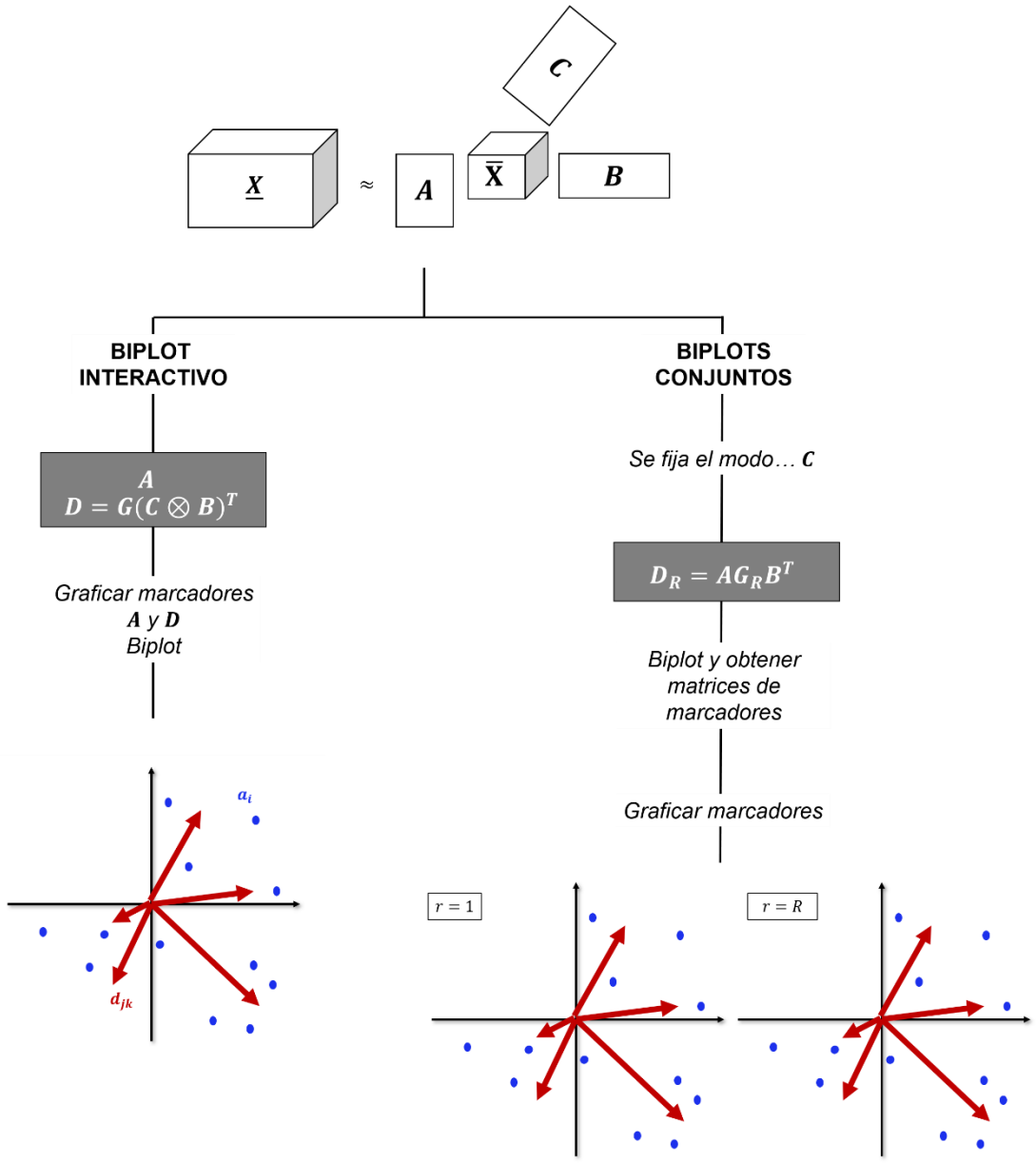


Figura 84. *Figura esquematizada de los gráficos de representación de resultados en el modelo Tucker3: Biplot interactivo y Biplots conjuntos*

5.3.6 Implementación en R

Para la programación de C_{enet}Tucker en R se han implementado las funciones de la Tabla 35:

***Tabla 35**. Funciones implementadas en R para la descomposición Tucker3 sparse (C_{enet}Tucker 3) y gráficos asociados*

Función	Argumentos	
tucker.enet	X, tucker.type,I,J,K,p,q,r, tau.uA = 1.4, tau.uB = 1.4, tau.uC=1.4, alpha.uA=1e-16, alpha.uB=1e-16, alpha.uC=1e-16, tau.vA = sqrt(I), tau.vB = sqrt(J), tau.vC=sqrt(K), alpha.vA=1e-16, alpha.vB=1e-16,alpha.vC=1e-16, itermax.pi=100, itermax.pocs=100, itermax.tucker=100, eps.pi=1e-16, eps.pocs=1e-16, eps.tucker=1e-16, init.svd="svd"	Descomposición tensorial Tucker (tucker3: tucker.type=3, tucker2: tucker.type=2)
factorplots.enet	A, B, C, Qa.plot=c(1,2), Qb.plot=c(1,2), Qc.plot=c(1,2)	Gráficos de puntuaciones factoriales para cada uno de los modos (*Qa.plot* hace referencia a los ejes factoriales a graficar para el modo A).
intBiplot.enet	A,B,C,G,Q.plot=c(1,2)	Biplot interactivo
jointBiplot.enet	A,B,C,G, Q.plot=c(1,2)	Biplots conjuntos

5.4 Análisis de datos reales: aplicación de C_{enet}Tucker3

5.4.1 Base de datos

Ilustraremos la aplicación de la técnica central de este capítulo, C_{enet}Tucker3, analizando la percepción del comportamiento paternal de los padres y sus hijos (150 niños y 153 niñas) en Japón (Kojima, 1975). La base de datos Kojima (Kojima, 1975; Kroonenberg, 2008) está disponible en el paquete *ThreeWay* de R. El estudio planteado por Kojima tenía el objetivo de validar la estructura factorial del cuestionario de conducta parental percibida CRBPI (*Child's Report of Parental Behavior Inventory*) (Schaefer, 1965), la escala más utilizada para medir el comportamiento de los padres percibido por los hijos.

La percepción que tienen los adolescentes sobre la manera en que sus padres les tratan (cómo les controlan, la disciplina que hay en casa, comunicación, cariño, …) se utilizan para examinar el modo en que los padres educan a sus hijos (Valiente, Magaz, Chorot, & Sandín, 2016). Estas conductas son de interés porque pueden influir en el desarrollo psicológico de niños/adolescentes e incluso causar problemas psicopatológicos como ansiedad y depresión, problemas de conducta, … (Muris, 2010; Valiente et al., 2016)

La solución de la aplicación C_{enet}Tucker3 analiza la respuesta de las 153 niñas japonesas y sus respectivos padres a la versión japonesa del CRPBI, para estudiar la comparación de sus reacciones. Se recogieron las respuestas de padres y niñas al cuestionario CRBPI; en concreto, a dos versiones del mismo: una para hijos y otra para padres (conocida como PR-PBI) que desarrolló Kojima (1975) de forma paralela.

El instrumento está compuesto por 18 escalas de tipo Likert con tres categorías (*nunca o casi nunca, solo algunas veces, muchas veces*) en las que padres y niños contestan a una serie de enunciados diseñados para medir la percepción de los niños con respecto a la aceptación que sienten de sus padres, autonomía psicológica y control parental (Kroonenberg, Harshman, & Murakami, 2009). La escala original desarrollado por Schaefer (1965) estaba formado por 260 ítems agrupados en torno a 26 escalas factoriales (10 ítems cada una). Sin

embargo, dada su poca utilidad práctica, el cuestionario fue posteriormente acortado en una variedad de formas (Cross, 1969; Margolies & Weintraub, 1977). De todas ellas, la más utilizada a nivel de investigación y también en el diseño de los datos de este análisis es la versión de Schludermann & Schludermann (1970) formada por 108 ítems agrupados en 18 subescalas en torno a 3 constructos latentes: Aceptación (AC) (vs rechazo), control firme (FC) (vs control permisivo) y control psicológico (PC) (autonomía psicológica).

La base de datos está formada por la opinión de los padres con respecto al comportamiento de sus hijas y de las hijas con respecto al de sus padres. La base de datos de estructura tridimensional está formada por $I = 153$ chicas (modo 1, filas), $J = 18$ escalas de comportamiento respectivas al CRBPI (columnas, modo 2) y $K = 4$ condiciones (el juicio de los padres hacia su propio comportamiento (F-F), el de las madres hacia su propio comportamiento (M-M), el de las hijas hacia el comportamiento de sus padres (G-F) y el de las hijas hacia el comportamiento de sus madres (G-M). Así, el array consiste en una tabla de datos de dimensión $153\ x\ 18\ x\ 4$.

Los objetivos que plantean Kroonenberg et al (2009) en su investigación hacen referencia a conocer en qué medida la estructura factorial de la escala es independiente de quién juzga el comportamiento de los padres (padres o hijos) y al análisis de las posibles diferencias individuales entre quienes lo valoran. Para mejorar la interpretación de la validez factorial de la versión corta del cuestionario han sido utilizados métodos de rotación (Kawash & Clewes, 1988; Kroonenberg et al., 2009). En este caso, además de los objetivos planteados Kroonenberg et al (2009), nuestro interés se centra en comparar los resultados del modelo clásico de tres vías Tucker 3 para analizar las hipótesis planteadas con respecto al uso de métodos de selección de variables como Lasso y Elastic net.

5.4.2 Análisis

Los datos fueron analizados mediante el modelo C_{enet}Tucker3. Uno de los puntos más investigados en los modelos Tucker es la selección del número de componentes a retener para cada uno de los modos, pues puede variar enormemente a diferencia de lo que ocurre en la descomposición CP. En este

sentido se hace uso del análisis realizado por Kroonenberg et al (2009) para la selección del número de componentes a retener (Figura 85). En estos gráficos se representan todos los modelos posibles para $P, Q, R \leq 4$ según la suma del número de componentes retenidas y la suma de cuadrados residual. A partir de estos resultados, el modelo escogido como óptimo es el que retiene $P = 4$ componentes para el primer modo, $Q = 3$ para el segundo y $R = 2$ para el tercero. Los modelos posteriores a este presentan una reducción de la suma de cuadrados residual mínima.

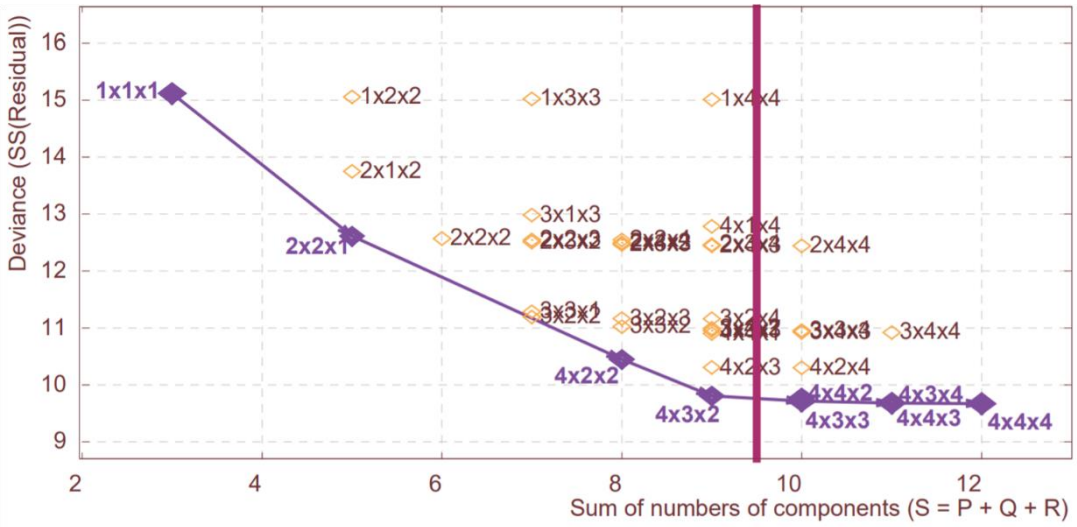


Figura 85. *Suma de cuadrados residual del Tucker3 según el número de componentes retenidas en cada uno de los modos. Fuente:* (Kroonenberg et al., 2009)

El C$_{enet}$Tucker3 fue ejecutado 250 veces (con 500 iteraciones para cada C$_{enet}$SVD) para aumentar la posibilidad de encontrar la solución óptima (las soluciones se mostraron estables con otros números de iteraciones). La tolerancia para el criterio de convergencia fue igual a 10^{-16}. Se utilizó la SVD clásica para la inicialización de las matrices. El preprocesamiento de los datos se llevó a cabo con el centrado de los mismos por el modo $\boldsymbol{A}$ y estandarización por el modo B. Según los niveles de sparsity para el parámetro de regularización propuestos por (Guillemot et al., 2019), se escogen parámetros de regularización que incluyan un nivel medio de penalización sparse en los modos A y B ($\tau_a = \text{sqrt}(153) * (2/3)$; $\tau_b = \text{sqrt}(18) * (2/3)$). En el caso del modo $\boldsymbol{C}$, como está

compuesto por un número menor de elementos ($K = 4$), la penalización incluida en el modelo es menor ($\tau_c = \text{sqrt}(4) * (3/4)$). En cuanto a la medida α que define la penalización Elastic net incluida en el modelo, α es escogido manualmente como $\alpha = 0{,}2$. Como el objetivo principal de la aplicación se centra en la estructura factorial del cuestionario, en la sección de resultados se mostrará la comparación de la selección de distintos valores para el parámetro α; desde el modelo Tucker3 clásico sin penalización, el modelo Tucker3 penalizado con Elastic net $\alpha = 0{,}5$ y $\alpha = 0{,}2$ y el modelo sparse Tucker3 con penalización Lasso ($\alpha = 0$). Se observará así la influencia de este parámetro en la selección de variables final.

5.4.3 Resultados

En primer lugar se muestran las matrices de componentes para cada uno de los modos.

Subescalas de percepción de comportamiento CRBPI (modo 2). A pesar de que la estructura trifactorial del cuestionario es la más aceptada por la comunidad, varios autores discuten este hecho y proponen otro tipo de soluciones. Esto en parte es debido a que los métodos clásicos como el Tucker y el CP no recogen de manera clara la configuración del cuestionario en sus resultados iniciales. Por este motivo, los métodos de rotación han sido siempre la opción escogida para mejorar la interpretación de los resultados (Kroonenberg et al., 2009). En nuestro caso, los métodos de rotación, aplicados como un postprocesamiento sobre los resultados del método de descomposición, son sustituidos por técnicas de selección de variables. En este caso, mediante la adición de términos de penalización como es el caso de C_{enet}Tucker3 con la penalización Elastic net. Con el objetivo de comparar los resultados de distintos grados de sparsity, se presentan en la Figura 86 las matrices de componentes para el modo $\boldsymbol{B}$ obtenidas bajo distintos niveles del parámetro α. Se observa que para $\alpha = 0.5$ el modelo clásico no restringido y el modelo restringido presentan soluciones similares; mientras que para $\alpha = 0$ (enet es transformado en Lasso) las componentes evidencian soluciones sparse con factores disjuntos (las escalas solo cargan a la formación de una sola componente). A nivel práctico, el objetivo no es producir factores disjuntos puesto que las diferentes escalas presentan

correlaciones que quedarían así perdidas, se considera el método para $\alpha = 0{,}2$, que aclara la formación de la PC1 con respecto a la técnica tradicional y genera matrices de componentes para el resto de los modos más claras que con el Tucker3 (estos datos no se muestran aquí).

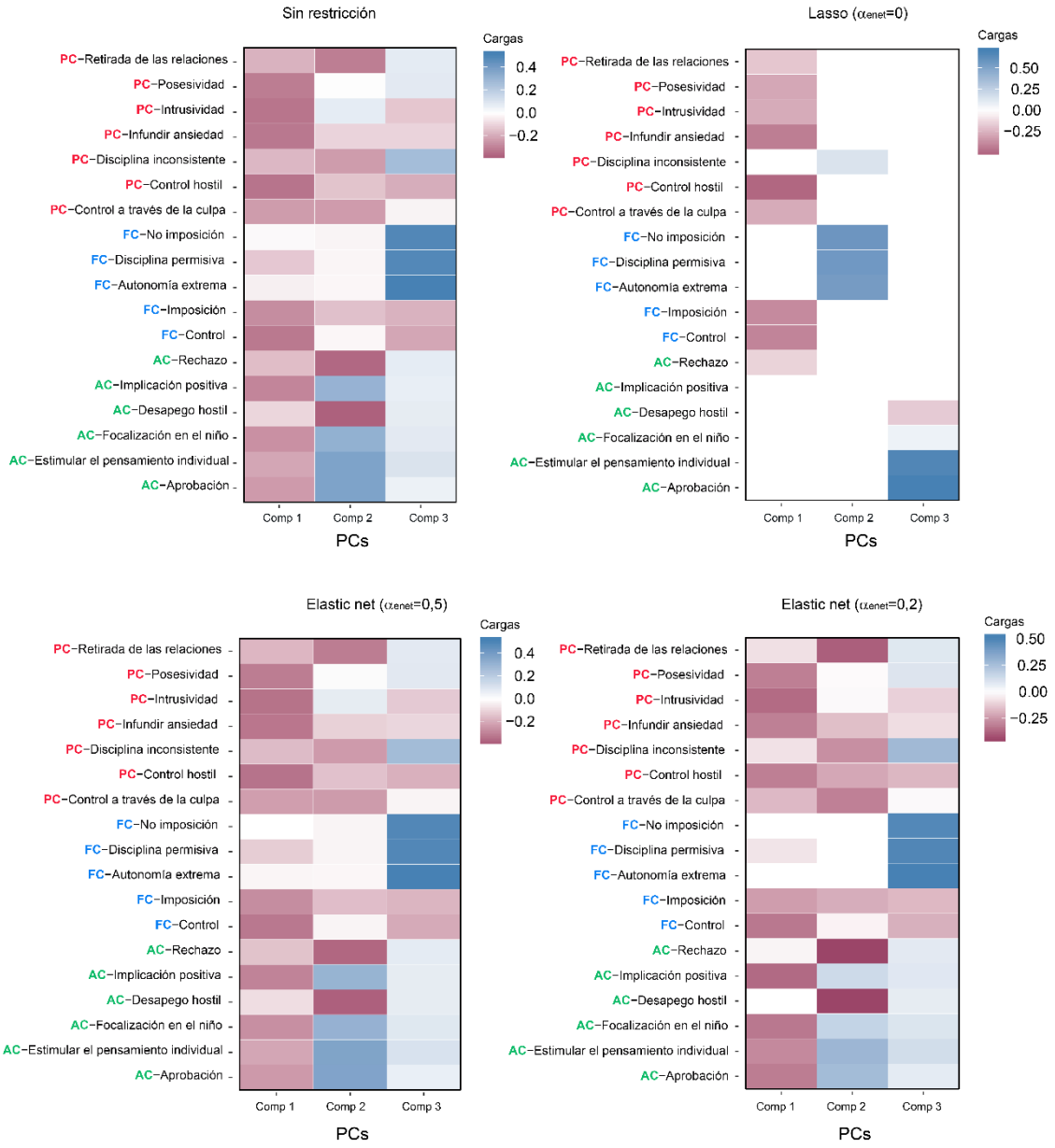


***Figura 86**. Matriz de cargas para las componentes del modo B según el tipo de penalización incluido en el modelo*

La solución para la matriz de cargas de las escalas del cuestionario CRBPI para la penalización enet seleccionada ($\alpha = 0{,}2$) se recogen en la Tabla 36. La primera componente sparse restringida está constituida por ítems de las escalas teóricas AC y PC en referencia al apoyo parental o la aprobación (Aprobación, Estimular el pensamiento individual, Focalización en el niño, Implicación positiva,

Posesividad, Infundir ansiedad, Intrusividad, Control hostil, Imposición y Control), frente a la segunda componente formada por ítems de las escalas PC y AC relacionados con el rechazo (Control a través de la culpa, Retirada de las relaciones, Rechazo, Desapego hostil) y la tercera componente relacionada con una disciplina permisiva. En este último caso, los ítems no se relacionan con los oficiales definidos en su escala teórica (FC). Este hecho ya ha sido evidenciado por otros autores y la matriz de configuración, que no representa la estructura factorial teórica a nivel preciso apoya los resultados de (Kroonenberg et al., 2009).

Tabla 36. *Matriz de marcadores* $\boldsymbol{B}$ *resultante en el* C_{enet}*Tucker3 y en el Tucker3 clásico.*

Subescala	**Escala teórica**	**C_{enet}C1**	**C_{enet}C2**	**C_{enet}C3**
Aprobación	AC	**-0,31**	0,27	0,07
Estimular el pensamiento individual	AC	**-0,28**	0,28	0,13
Focalización en el niño	AC	**-0,32**	0,17	0,10
Implicación positiva	AC	**-0,35**	0,14	0,09
Posesividad	PC	**-0,30**		0,10
Infundir ansiedad	PC	**-0,30**	-0,15	-0,08
Intrusividad	PC	**-0,35**	-0,01	-0,11
Control hostil	PC	**-0,30**	-0,21	-0,17
Imposición	FC	**-0,23**	-0,19	-0,16
Control	FC	**-0,32**		-0,18
Control a través de la culpa	PC	-0,16	**-0,30**	-0,01
Retirada de las relaciones	PC	-0,07	**-0,38**	0,09
Rechazo	AC		**-0,44**	0,08
Desapego hostil	AC		**-0,44**	0,07
Disciplina inconsistente	PC	-0,07	-0,26	**0,27**
Disciplina permisiva	FC	-0,06		**0,50**
Autonomía extrema	FC			**0,52**
No imposición	FC			**0,49**

Condiciones-Juicio del comportamiento (modo 3). El tercer modo es el referente a la opinión que tienen las hijas de sus padres y madres, respectivamente (girl-father (G.F), girl-mother (G.M)) y la opinión de los padres sobre sí mismos (father-father (F.F), mother-mother (M.M)). En el análisis, se retienen $R = 2$ factores. De la matriz de cargas de C_{enet}Tucker3 en la Tabla 37 se sigue que las chicas tienen opiniones similares a sus dos padres y también que los padres/madres hacen juicios parecidos de su propio comportamiento. Puede hablarse así de un comportamiento de los padres a nivel general sin necesidad de diferenciación. La matriz de configuración en el caso del Tucker clásico no evidencia las relaciones entre opiniones de hijas y opiniones de padres que le Tucker3 restringido si es capaz de detectar.

***Tabla 37**. Matriz de marcadores **C** resultante en el C_{enet}Tucker3 y en el Tucker3 clásico.*

	C_{enet}Tucker3		**Tucker3**	
Juicio	**C_{enet}C1**	**C_{enet}C2**	**C1**	**C2**
G.F	**-0,64**	0,21	-0,53	0,43
F.F	-0,18	**-0,71**	-0,46	-0,58
G.M	-**0,68**	0,26	-0,53	0,50
M.M	-0,30	**-0,62**	-0,47	-0,47

Familias (modo 1). Dada la dificultad para interpretar la matriz de cargas del primer modo de dimensión 153 x 4, los valores de las puntuaciones factoriales de las familias en las cuatro dimensiones latentes se recogen de manera gráfica en la Figura 87. La matriz de cargas para el modo A puede consultarse en los anexos del trabajo (Tabla S3). Se encuentran cuatro grupos de chicas diferentes.

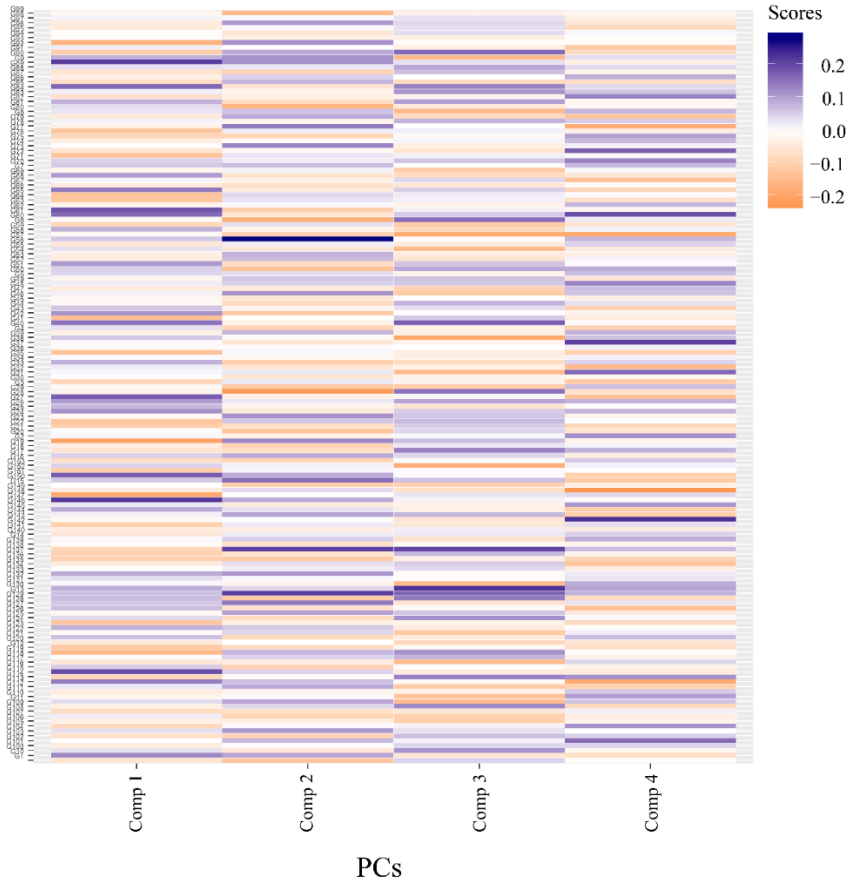


***Figura 87**. Gráfico de cargas factoriales para la matriz de componentes del modo A (puntuaciones de las familias (girls) en las cuatro componentes retenidas)*

Interacción entre componentes. Matriz Core.

La matriz Core (Tabla 38) obtenida tras la aplicación del CenetTucker3 contiene los elementos para valorar las relaciones entre las componentes de cada uno de los modos. Cada uno de sus elementos contiene la información de la variabilidad explicada por la combinación de las respectivas componentes del modo $\boldsymbol{A}$, $\boldsymbol{B}$ y $\boldsymbol{C}$. Así, por ejemplo el elemento $g_{111} = 1{,}19$ contiene la cantidad de inercia explicada conjuntamente por la primera componente del modo $\boldsymbol{A}$ y por las primeras componentes restringidas de los modos $\boldsymbol{B}$ y $\boldsymbol{C}$, respectivamente. La combinación de las componentes 111 explican un 17,5% de la varianza explicada por el análisis. Nótese que los coeficientes que se interpretan de la matriz Core son los de coeficientes más altos; en este caso se interpretarán los coeficientes $g_{111} = 1{,}19$, $g_{221} = 1{,}16$ y $g_{112} = -0{,}77$. Pero igualmente se podrían interpretar otros como $g_{211} = -0{,}94$, pues no hay gran diferencia entre la varianza explicada entre ellos.

Tabla 38. *Matriz Core resultante en el C_{enet}Tucker3*

			Elementos			Varianza explicada		
			Modo 2					
			C_{enet}C1	C_{enet}C2	C_{enet}C3	C_{enet}C1	C_{enet}C2	C_{enet}C3
Modo 3	C_{enet}C1	Modo 1 C_{enet}C1	1,19	0,98	-0,07	17,53%	11,95%	0,07%
		Modo 1 C_{enet}C2	-0,94	1,16	0,12	11,05%	16,63%	0,17%
		Modo 1 C_{enet}C3	0,38	0,21	0,69	1,83%	0,52%	5,94%
		Modo 1 C_{enet}C4	-0,23	-0,05	0,72	0,66%	0,04%	6,50%
	C_{enet}C2	Modo 1 C_{enet}C1	0,62	0,28	-0,17	4,81%	0,95%	0,36%
		Modo 1 C_{enet}C2	-0,23	0,39	-0,09	0,66%	1,86%	0,10%
		Modo 1 C_{enet}C3	-0,77	-0,28	0,39	7,43%	0,99%	1,88%
		Modo 1 C_{enet}C4	0,72	-0,08	0,35	6,50%	0,07%	1,51%

La interpretación de los signos de los elementos de la matriz Core con una mayor contribución se realiza a continuación. Sabemos que $g_{111} = 1{,}19$ representa el grado de relación entre las primeras componentes de cada modo. Ahora, se estudian los elementos de las matrices $\boldsymbol{A}$ (estudiantes), $\boldsymbol{B}$ (escalas) y $\boldsymbol{C}$ (programas) que ofrecen los mayores pesos en la primera componente retenida de cada modo, seleccionando aquellos con una mayor contribución tras aplicar la umbralización en los estudios de dos vías. Los gráficos factoriales de las puntuaciones de las matrices para cada uno de los modos; es decir, de las familias, escalas y juicios aparecen en la Figura 88 (planos 1-2) y en la Figura 89 (planos 2-3 de los modos 1 y 2). Estos, junto con los resultados de la Tabla 38 servirán para interpretar las relaciones entre familias, subescalas de comportamiento y juicios.

Como el elemento $1x1x1$ de la matriz Core (Tabla 38) es positivo, las familias (niñas) que tengan coordenadas positivas en la primera componente (Figura 88A, cuadrantes azul y verde) interaccionarán de manera positiva con las escalas que tienen coordenadas negativas en la primera componente (Aprobación, Estimular el pensamiento individual, Focalización en el niño, Implicación positiva, Posesividad, Infundir ansiedad, Intrusividad, Control hostil, Imposición, Control) en cualquiera de los juicios realizados (todos los elementos de la matriz $\boldsymbol{C}$ son negativos en la primera componente).

Ahora, las familias que se encuentran en los cuadrantes rosa y morado (coordenadas negativas en su primera componente) interactúan de forma

negativa con las escalas de coordenadas negativas de la primera componente en cualquiera de las 4 opiniones.

De la misma forma, analizando el elemento $2x2x1$ de la Core (de signo positivo), se sigue que las siguientes combinaciones tendrán una interacción positiva:

Familias2 (+) x Subescalas2 (-) x Juicios1 (-) x Core (+) = Interacción (+)

Familias2 (-) x Subescalas2 (+) x Juicios1 (-) x Core (+) = Interacción (+)

De donde se sigue que:

- Las familias con coordenadas positivas en la segunda componente (cuadrantes rosa y azul) interaccionan de manera positiva con las escalas que tienen coordenadas negativas en la segunda componente (Infundir ansiedad, Control hostil, Imposición, Control a través de la culpa, Retirada de las relaciones, Rechazo, Desapego hostil, Disciplina inconsistente) en las cuatro opiniones consideradas (coordenadas negativas en la primera componente del modo C).
- Las familias con coordenadas negativas en la segunda componente (cuadrantes morado y verde) interaccionan de manera positiva con las escalas que tienen coordenadas positivas en la segunda componente (Aprobación, Estimular el pensamiento individual, Focalización en el niño, Implicación positiva) en las cuatro opiniones consideradas (coordenadas negativas en la primera componente del modo C).

Además,

Familias2 (+) x Subescalas2 (+) x Juicios1 (-) x Core (+) = Interacción (-)

Familias2 (-) x Subescalas2 (-) x Juicios1 (-) x Core (+) = Interacción (-)

Esto quiere decir que:

- Las familias con coordenadas positivas en la segunda componente (cuadrantes rosa y azul, Figura 88) interaccionan de manera negativa con las escalas que tienen coordenadas positivas en la segunda componente (Aprobación, Estimular el pensamiento individual, Focalización en el niño,

Implicación positiva) en las cuatro opiniones consideradas (coordenadas negativas en la primera componente del modo C).

- Las familias con coordenadas negativas en la segunda componente (cuadrantes morado y verde, Figura 88) interaccionan de manera negativa con las escalas que tienen coordenadas negativas en la segunda componente (Infundir ansiedad, Control hostil, Imposición, Control a través de la culpa, Retirada de las relaciones, Rechazo, Desapego hostil, Disciplina inconsistente) en las cuatro opiniones consideradas (coordenadas negativas en la primera componente del modo C).

Por último, la interpretación del elemento $1x1x2$ de la Core (de signo negativo) y los signos de las componentes $P = 1, Q = 1, R = 2$:

- Las familias con coordenadas positivas en la primera componente (cuadrantes azul y verde, Figura 88) se relacionan de manera directa con las escalas que tienen coordenadas negativas en la primera componente para las opiniones de las niñas acerca de su padre y de su madre (coordenadas positivas en el tercer modo de condiciones).

Familias1 (+) x Subescalas1 (-) x Juicios2 (+) x Core (-) = Interacción (+)

- Las familias con coordenadas negativas en la primera componente (cuadrantes rosa y morado, Figura 88) se relacionan de manera directa con las escalas que tienen coordenadas negativas en la primera componente para las opiniones de los padres/madres acerca de ellos mismos.

Familias1 (-) x Subescalas1 (-) x Juicios2 (-) x Core (-) = Interacción (+)

- Las familias con coordenadas positivas en la primera componente (cuadrantes azul y verde, Figura 88) interactúan negativamente con las escalas que tienen coordenadas negativas en la primera componente para las opiniones de los padres/madres acerca de ellos mismos.

Familias1 (+) x Subescalas1 (-) x Juicios2 (-) x Core (-) = Interacción (-)

- Las familias con coordenadas negativas en la primera componente (cuadrantes rosa y morado, Figura 88) interactúan de manera inversa con

las escalas que tienen coordenadas negativas en la primera componente para las opiniones de las niñas acerca de su padre y de su madre

Familias1 (-) x Subescalas1 (-) x Juicios2 (+) x Core (-) = Interacción (-)

Los planos factoriales 1-2 para el modo A, B y C y los planos de ejes factoriales 2 y 3 para los modos A y B se recogen en la Figura 88 y Figura 89.

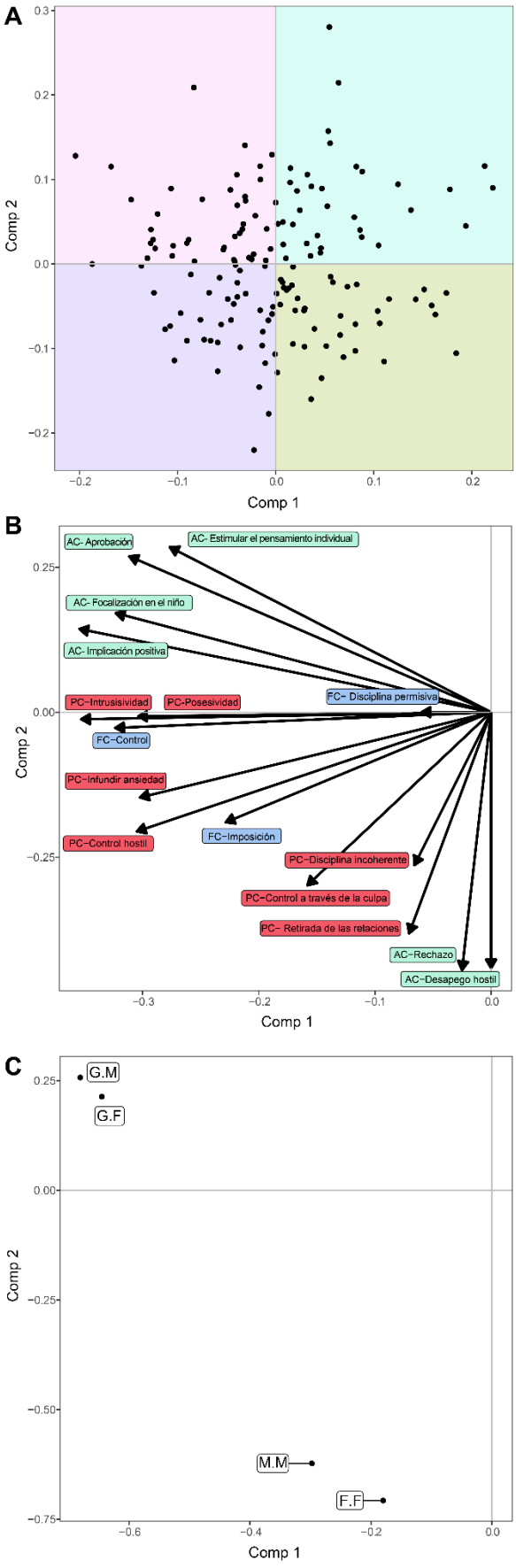


***Figura 88**. Planos factoriales 1-2 para los modos A, B, C*

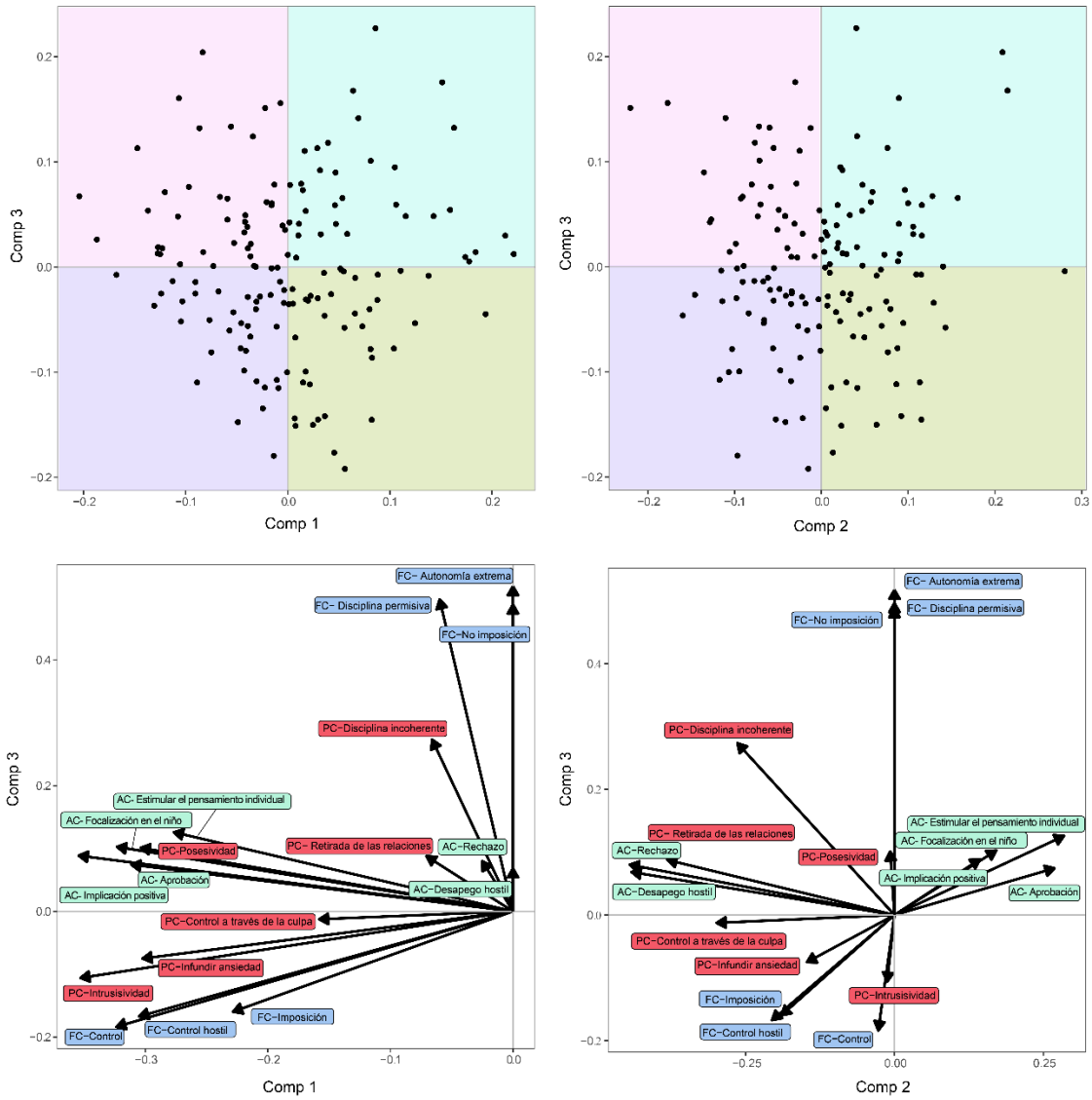


Figura 89. *Planos factoriales 1-3 y 2-3 para los modos A y B*

Biplot conjunto. Las interpretaciones de los elementos de la matriz Core pueden ser resumidos a su vez a partir de los resultados de las representaciones Biplot. En este análisis, debido al gran número de categorías de las subescalas del cuestionario y opiniones, el Biplot interactivo no es aconsejable. Además, este es útil en aquellos casos en que las condiciones presentan un orden temporal lógico (como podrían ser los años), pero en este caso esto no es así. Por ello, para representar gráficamente el comportamiento de todas las dimensiones se utiliza el Biplot conjunto. En el Biplot conjunto se representa en un mismo gráfico factorial la información de las componentes de dos modos, fijadas las componentes del tercero (el modo de referencia). En este caso, se representan

los Biplots conjuntos con $\boldsymbol{A}$ el modo de referencia (Figura 90, Figura 91, Figura 92, Figura 93). Esta es la primera cuestión a tener en cuenta: los modos a graficar y el modo de referencia. Siguiendo las recomendaciones de (P. Kroonenberg, 2008) y dado que los objetivos son examinar la diferencia de opiniones entre hijas y padres es lógico graficar los modos $\boldsymbol{B}$ y $\boldsymbol{C}$ y por ello el modo $\boldsymbol{A}$ es el seleccionado como modo de referencia. El lector debe tener en cuenta que podrían haberse fijado cualquiera de estos dos modos, generando nueva información para el estudio. Como se han retenido 4 componentes en el primer modo, se realizan 4 Biplots (uno para cada componente) que se interpretan de manera similar. Como cada una de las componentes sparse representa a un grupo de familias diferente, los cuatro Biplots sirven para explicar los patrones de comportamiento de estos cuatro grupos de niñas.

Además, dada la naturaleza del Biplot conjunto, recuérdese que cada gráfico representa a aquellas familias con peso alto en la componente correspondiente. El comportamiento de las familias con pesos nulos (hecho conseguido a partir de la aplicación de la descomposición sparse) no estarán evidenciadas en dichas componentes. En cuanto a la forma correcta de interpretación, téngase en cuenta que para aquellas familias con pesos positivos la interpretación de los resultados coincide con la interpretación de los métodos Biplot clásicos. Sin embargo, para aquellas observaciones con cargas negativas en la componente del tercer modo la interpretación se realizará a la inversa: observaciones (en este caso, opiniones) cercanas a una escala presentarán valores bajos en dicha escala.

Como ejemplo, se muestra la interpretación del Biplot conjunto asociado a la primera componente del modo A. En cuanto a las subescalas del cuestionario, pueden mencionarse cuatro grupos de subescalas relacionadas. Por un lado, la escala de la disciplina permisiva, con relaciones entre la aprobación, la estimación del pensamiento individual, la autonomía de las niñas y la no imposición. Además, estas variables están directamente correlacionadas con la implicación positiva y la focalización en el niño. Por otro lado aparece la escala relacionada con el control del comportamiento, con correlaciones altas y directas entre la posesividad, intrusismo, control y la provocación de ansiedad. Se observan además correlaciones altas entre el desapego hostil, retirada de las

relaciones, rechazo, imposición y control a través de la culpa. Las subescalas de la dimensión teórica AC positivas (aprobación, implicación positiva, estimulación del pensamiento individual) son prácticamente independientes de las subescalas rechazo y desapego. Algo similar ocurre entre los ítems de autonomía extrema, no imposición y disciplina permisiva con el control, la imposición y el control hostil. En el estudio de la diferencia de opiniones entre padres e hijas queda patente que:

1. La proyección de los juicios de los padres sobre si mismos plasman puntuaciones altas en aceptación y apoyo parental y disciplina permisiva, frente a puntuaciones más bajas en la opinión de las hijas con respecto a estas subescalas.
2. Las hijas perciben de sus padres un control a través de la culpa, rechazo y desapego, imposición de las cosas y retirada de las relaciones, frente a puntuaciones bajas de los padres con respecto a su opinión de su mismos en estas subescalas.

Esta interpretación es acorde a aquellas familias con pesos positivos en la primera componente del modo $\boldsymbol{A}$. Las familias con cargas negativas en dicha componente sufren los hechos a la inversa: las hijas perciben una autonomía extrema, disciplina permisiva, aceptación … mayor a la que los padres creen que aplican. Las hijas conciben un apoyo parental más favorable que el que sus padres pretenden dar.

Este desarrollo está actualmente en vías de escritura para ser posteriormente publicado en forma de artículo bajo el título “Sparse analysis in multi-way data context: Sparse Tucker-3-enet model”.

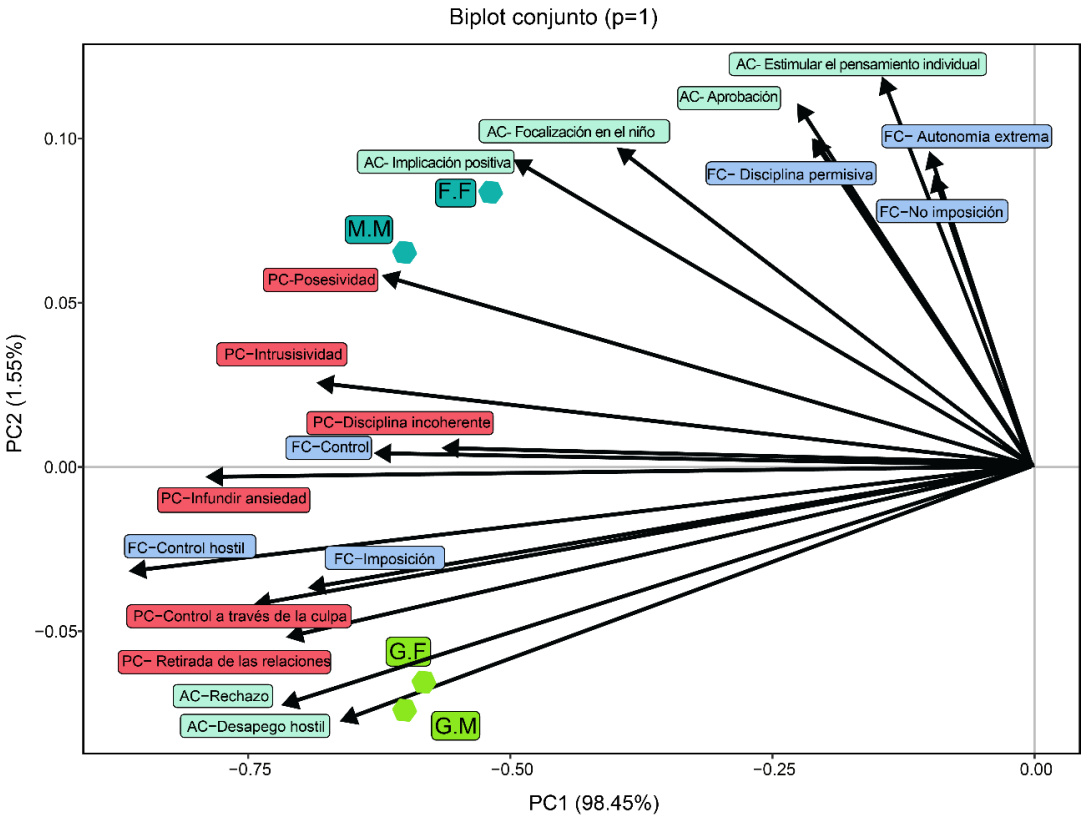


***Figura 90**. Biplot conjunto 1. Representación de juicios y escalas sobre la primera componente de familias*

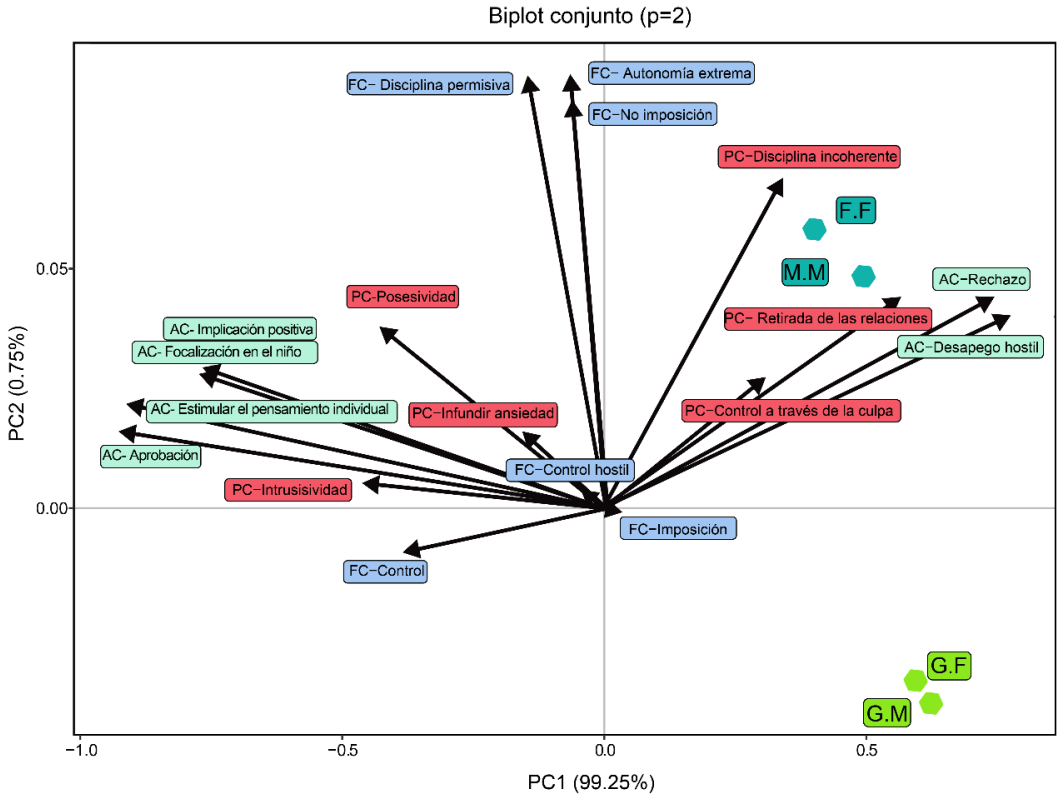


***Figura 91**. Biplot conjunto 2. Representación de juicios y escalas sobre la segunda componente de familias*

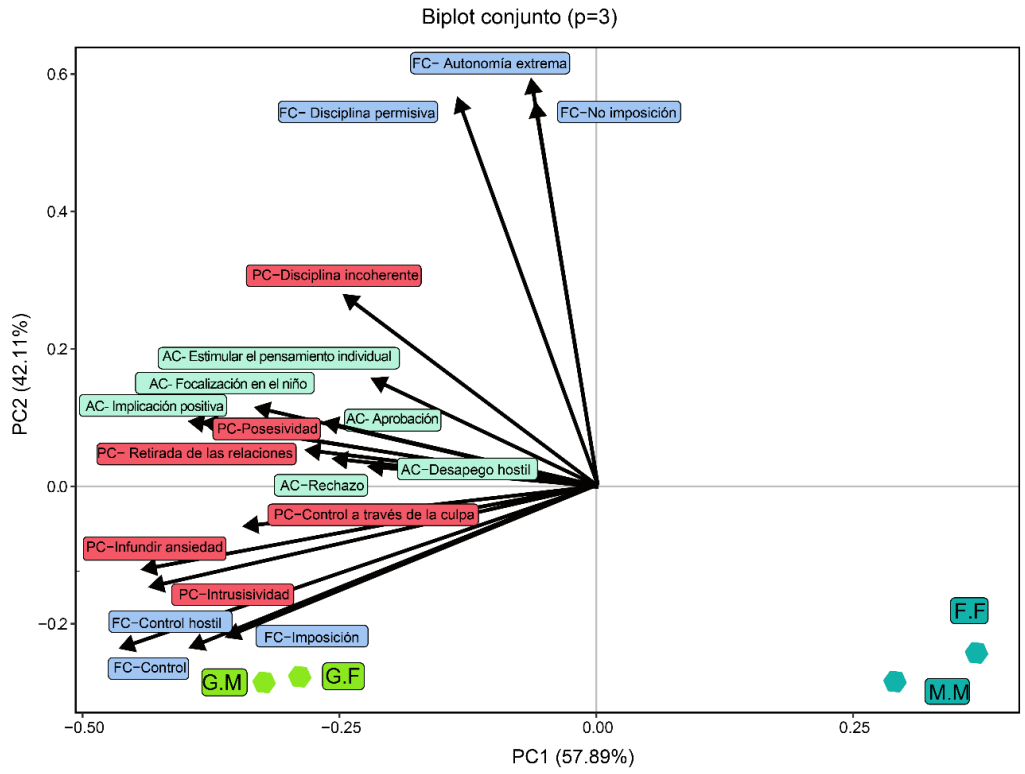


Figura 92. *Biplot conjunto 3. Representación simultánea de juicios y escalas sobre la tercera componente de familias*

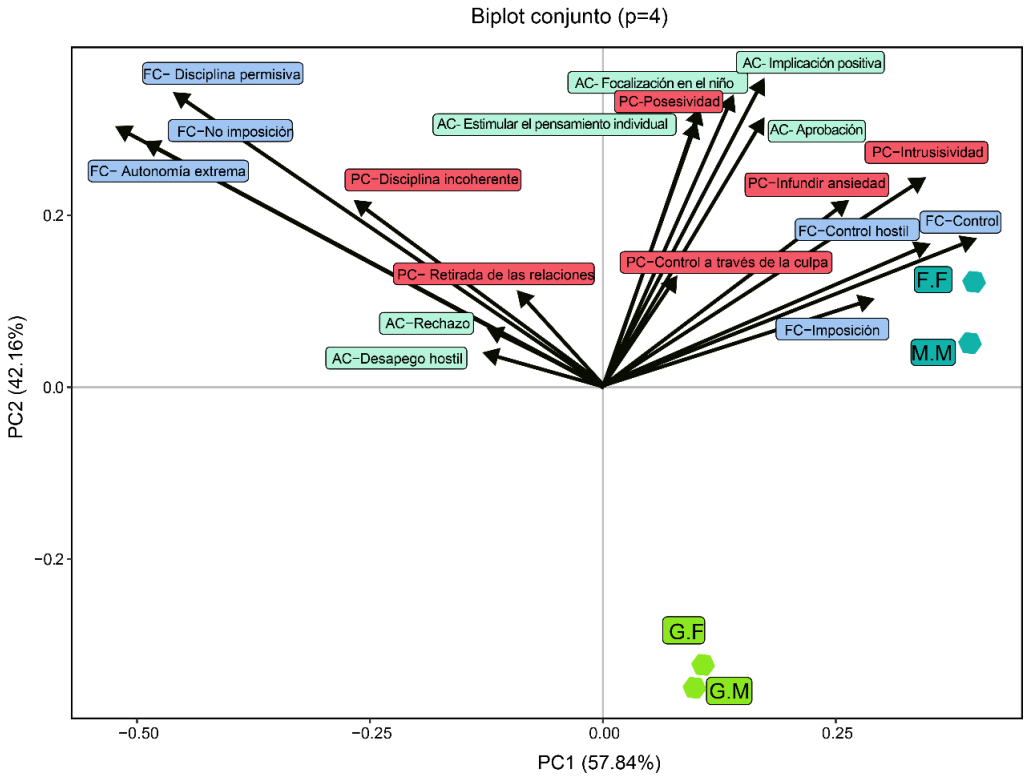


Figura 93*. Biplot conjunto 4. Representación simultánea de juicios y escalas sobre la cuarta componente de familias*

CONCLUSIONES

1. La descomposición en valores singulares, propuesta en el año 1936, se sitúa en el eje central de la estadística multivariante, como cimiento teórico de una gran cantidad de métodos de análisis de datos de dos, tres o n-vías, entre los que cabe destacar, el Análisis de Componentes Principales, los métodos Biplot, o los modelos de descomposición tensorial Tucker.

2. El Análisis de Componentes Principales es la técnica más utilizada para la reducción de la dimensionalidad. Cada componente principal se calcula, a partir de los vectores de la descomposición en valores singulares, como una combinación lineal de todas las variables de partida. La literatura ofrece soluciones clásicas como la umbralización, los métodos de rotación y las soluciones modernas sparse, para superar esta limitación.

3. De la exhaustiva revisión bibliográfica de los métodos sparse, en el ámbito de las componentes principales, se deduce que no existe unanimidad en la forma de generar coeficientes nulos, lo cual complica su extensión a las técnicas de análisis de datos de tres vías.

4. Siguiendo las ideas de Guillemot et al (2019), se ha desarrollado la extensión de la descomposición en valores singulares restringida, sparse y ortogonal, a la penalización Elastic net, basada en los métodos de proyección de un vector sobre la intersección de conjuntos convexos. Esta metodología la hemos denominado C_{enet}SVD.

5. Como extensión de dicha técnica se propone el análisis de componentes principales sparse restringido a Elastic net, denominado C_{enet}PCA y los métodos Biplot GH, JK y HJ sparse, denominados C_{enet}Biplots. Estas versiones solventan la deficiencia de algunos de los métodos existentes, que no generan ejes ortogonales y sparse simultáneamente.

6. La descomposición en valores y vectores singulares ortogonales y sparse restringidos a la bola Elastic net ha sido generalizada al caso de tensores multidimensionales a partir del ajuste de los modelos Tucker y la adaptación del modelo de Tuckals.

7. Para todas estas metodologías sparse se han implementado funciones específicas en R en forma de librería bajo el nombre “SparseCenet”.

8. Las contribuciones teóricas de esta investigación suponen un gran avance para el análisis de Big Data y el software desarrollado permitirá aplicar los aportes teóricos al análisis de datos en cualquier campo de la ciencia.

LÍNEAS FUTURAS

Investigaciones futuras pueden considerar la posibilidad de implementar otros métodos de regularización en el marco de la CSVD, incluso proponer algoritmos de proyección de un vector sobre conjuntos no convexos bajo la teoría matemática apropiada. Esto abriría un abanico de posibilidades en torno a métodos de penalización no convexa como SCAD, que tan buenos resultados ha demostrado en la literatura, o penalizaciones estructuradas como *Group* Lasso. Por otro lado, es obvio que el modelo C_{en}SVD puede ser extendido a diferentes metodologías, al igual que la SVD supone la base de muchas.

Otras posibles investigaciones futuras a tener en cuenta, relacionadas con ello, deben englobar:

1) Un posible uso de distintas técnicas de penalización en base al origen de los datos, así como alternativas en las funciones de penalización que mejoren la consistencia en la selección de variables proporcionada por Lasso. Para ello, aunque selecciona las variables importantes con alta probabilidad (Benner, Zucknick, Hielscher, Ittrich, & Mansmann, 2010), la metodología Bootstrap podría suponer un nuevo aporte de estabilización de las soluciones sparse. Esta metodología ha sido propuesta en técnicas de dos vías (Sill et al., 2015). Una de las líneas a seguir en esta investigación sería introducir este concepto en los métodos de descomposición de tensores.
2) Definir un algoritmo de solución más eficaz computacionalmente. Una de las principales cuestiones a tener en cuenta en el desarrollo de estos métodos es el orden computacional y la eficacia de los algoritmos, debido a la gran cantidad de datos que se manejan, pues pueden no ser eficientes ni en tiempo ni en memoria. Se propone en este sentido hacer uso de los métodos computacionales de paralelización para mejorar el tiempo de ejecución de las funciones propuestas.
3) Implementar de C_{enet}JIVE en R y comparación de los resultados obtenidos entre el JIVE clásico y esta metodología.
4) Combinar las técnicas Cenet y los métodos de clasificación.
5) Analizar la utilidad de estas técnicas en campos como las imágenes cerebrales en neurociencia.

REFERENCIAS

Abdi, H. (2007). Singular Value Decomposition (SVD) and Generalized Singular Value Decomposition (GSVD). In N. Salkind (Ed.), Encyclopedia of Measurement and STATIStic (pp. 907–912).

Abdi, H., & Valentin, D. (2006). *Mathématiques pour les sciences cognitives*. Presse Universitaire de France, pp.357, 2006.

Abdi, H., Valentin, D., Chollet, S., & Chrea, C. (2007). Analyzing assessors and products in sorting tasks: DISTATIS, theory and applications. *Food Quality Prefer, 18*, 627–640. https://doi.org/10.1016/j.foodqual.2006.09.003

Abdi, H., Williams, L. J., & Valentin, D. (2013). Multiple factor analysis: principal component analysis for multitable and multiblock data sets. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, *5*(2), 149–179. https://doi.org/10.1002/wics.1246

Abdi, H., Williams, L. J., Valentin, D., & Bennani-Dosse, M. (2012). STATIS and DISTATIS: optimum multi-table principal component analysis and three way metric multidimensional scaling. *Wiley Interdisciplinary Reviews: Computational STATIStics, 4*(2), 124-167. https://doi.org/10.1002/wics.198

Adarkwa, O., Schumacher, T., & Attoh-Okine, N. (2015). Multiway Analysis of bridge structural types in the National Bridge Inventory (NBI): A tensor decomposition approach. *Proceedings - 2014 IEEE International Conference on Big Data*, 1–6. https://doi.org/10.1109/BigData.2014.7004423

Ahmed, T., Raja, H., & Bajwa, W. U. (2019). Tensor Regression Using Low-rank and Sparse Tucker Decompositions.

Akaike, H. (1974). A new look at the STATIStical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Alarcón, R. (2006). Desarrollo de una Escala Factorial para Medir la Felicidad. *Interamerican Journal of Psychology, 40*(1), 99–106.

Alfeld, M., Wahabzada, M., Bauckhage, C., Kersting, K., Wellenreuther, G., & Falkenberg, G. (2014). Non-negative factor analysis supporting the interpretation of elemental distribution images acquired by XRF. *Journal of Physics: Conference Series, 499*(1), 012013. https://doi.org/10.1088/1742-6596/499/1/012013

Algamal, Z. Y., & Lee, M. H. (2015). Regularized logistic regression with adjusted adaptive elastic net for gene selection in high dimensional cancer classification. *Computers in Biology and Medicine, 67*, 136–145. https://doi.org/10.1016/j.compbiomed.2015.10.008

Allen, G. I. (2012). Sparse Higher-Order Principal Components Analysis. In *Artificial Intelligence and STATIStics* (pp. 27-36).

Allen, G. I., & Maletić-Savatic, M. (2011). Sparse non-negative generalized PCA with applications to metabolomics. *Bioinformatics, 27*(21), 3029–3035. https://doi.org/10.1093/bioinformatics/btr522

Alonso-Gutierrez, J., Kim, E. M., Batth, T. S., Cho, N., Hu, Q., Chan, L. J. G.,Petzold, C.J., Hillson, N.J., Adamsab, P.D., Keasling, J.D. et al. (2015). Principal component analysis of proteomics (PCAP) as a tool to direct metabolic engineering. *Metabolic Engineering, 28*, 123–133. https://doi.org/10.1016/j.ymben.2014.11.011

Amini, A. A., & Wainwright, M. J. (2008). High-dimensional analysis of semidefinite relaxations for sparse principal components. *2008 IEEE International Symposium on Information Theory*, 2454–2458. https://doi.org/10.1109/ISIT.2008.4595432

Amor-Esteban, V., García-Sánchez, I.-M., & Galindo-Villardón, M.-P. (2017). Analysing the Effect of Legal System on Corporate Social Responsibility (CSR) at the Country Level, from a Multivariate Perspective. *Social Indicators Research*, *140*(1), 435-452. https://doi.org/10.1007/s11205-017-1782-2

Anaya-Izquierdo, K., Critchley, F., & Vines, K. (2011). Orthogonal simple component analysis: A new, exploratory approach. *Annals of Applied STATIStics, 5*(1), 486–522. https://doi.org/10.1214/10-AOAS374

Andrés, A. R., Asongu, S. A., & Amavilah, V. (2015). The Impact of Formal Institutions on Knowledge Economy. *Journal of the Knowledge Economy, 6*(4), 1034–1062. https://doi.org/10.1007/s13132-013-0174-3

Arbuckle, J. L. (2014). Amos. Chicago: IBM SPSS.

Auzmendi, E. (1992). *Características y Medición*. Bilbao: Mensajero.

Ayala, G. (2018). *Bioinformática Estadística: Análisis estadístico de datos ómicos*. 581.

Ayyala, R., Ahmed, F., Ruzal-Shapiro, C., & Taylor, G. (2019). Prevalence of Burnout Among Pediatric Radiologists. *Journal of the American College of Radiology, 16*(4A), 518–522. https://doi.org/10.1016/j.jacr.2018.08.016

Bader, B. W., Berry, M. W., & Browne, M. (2008). Discussion tracking in enron email using PARAFAC. In *Survey of Text Mining II: Clustering, Classification, and Retrieval* (pp. 147–163). https://doi.org/10.1007/978-1-84800-046-9_8

Bandalos, D. L., & Finney, S. J. (2018). Factor Analysis. In *The Reviewer's Guide to Quantitative Methods in the Social Sciences* (pp. 98–122). https://doi.org/10.4324/9781315755649-8

Barahona, G., Barreiro, C., González-García, N., Hernández, S., Sánchez-Barba, M., & Galindo-Villardón, M. (2019). Dynamic CUR, an alternative to variable selection in CUR decomposition. *Investigación Operacional, 40*(3), 391–399.

Barahona, G. V., García, N. G., Sánchez-García, A. B., Barba, M. S., & Galindo-Villardón, M. P. (2018). Seven methods to determine the dimensionality of tests: Application to the general self-efficacy scale in twenty-six countries. *Psicothema, 30*(4), 442–448. https://doi.org/10.7334/psicothema2018.113

Bauschke, H. H., & Combettes, P. L. (2017). Convex Analysis and Monotone Operator Theory in Hilbert Spaces. In *CMS Books in Mathematics*. https://doi.org/10.1007/978-3-319-48311-5

Beaton, D., Chin Fatt, C. R., & Abdi, H. (2014). An ExPosition of multivariate analysis with the singular value decomposition in R. *Computational STATIStics and Data Analysis, 72*, 176–189. https://doi.org/10.1016/j.csda.2013.11.006

Benasseni, J., & Bennani-Dosse, M. (2012). Analyzing multiset data by the power STATIS-ACT method. *Advances in Data Analysis and Classification, 6*(1), 49–65.

Benigni, R., & Giuliani, A. (1994). Quantitative modeling and biology: The multivariate approach. American. *Journal of Physiology - Regulatory Integrative and Comparative Physiology*, *266*(5), 1697-1704. https://doi.org/10.1152/ajpregu.1994.266.5.r1697

Benzécri, J. (1973). *L'analyse des données*. Paris: Dunod.

Berg, E., Schmidt, M., Friedlander, M., & Murphy, K. (2008). Group Sparsity Via Linear-Time Projection. Development, 1–11.

Biggs, J., Kember, D., & Leung, D. Y. P. (2001). The Revised Two Factor Study Process Questionnaire : R-SPQ-2F The Revised Two Factor Study Process Questionnaire : R-SPQ-2F. *British Journal of Educational Psychology, 71*(1), 133–149. https://doi.org/10.1348/000709901158433

Birgin, E., Martínez, J., & Raydan, M. (2000). Nonmonotone spectral projected gradient methods on convex sets. *SIAM Journal on Optimization, 10*(4), 1196–1211. https://doi.org/10.1137/S1052623497330963

Björck, Å. (2015). *Numerical methods in matrix computations*. Cham: Springer

Bodor, A., Csabai, I., Mahoney, M. W., & Solymosi, N. (2012). rCUR: an R package for CUR matrix decomposition. *BMC Bioinformatics, 13*(1), 103. https://doi.org/10.1186/1471-2105-13-103

Borg, I., & Groenen, P. (2003). Modern Multidimensional Scaling: Theory and Applications. *Journal of Educational Measurement, 40*(3), 277–280. https://doi.org/10.1111/j.1745-3984.2003.tb01108.x

Boutsidis, C., & Gallopoulos, E. (2008). SVD based initialization: A head start for nonnegative matrix factorization. *Pattern Recognition, 41*(4), 1350–1362. https://doi.org/10.1016/j.patcog.2007.09.010

Boyd, S., & Dattorro, J. (2003). Alternating Projections. *EE392o, Stanford University*.

Bria, M., Spânu, F., Băban, A., & Dumitraşcu, D. (2014). Maslach Burnout Inventory – General Survey: Factorial validity and invariance among Romanian healthcare professionals. *Burnout Research, 1*(3), 103–111. https://doi.org/10.1016/J.BURN.2014.09.001

Brink-Jensen, K. (2014). *Integrative Modeling and Inference in High Dimensional Genomic and Metabolic Data.* University of Copenhagen

Briz-Ponce, L., & García-Peñalvo, F. J. (2015). An Empirical Assessment of a Technology Acceptance Model for Apps in Medical Education. *Journal of Medical Systems, 39*(11), 176. https://doi.org/10.1007/s10916-015-0352-x

Bro, R., & Smilde, A. K. (2014). Principal component analysis. *Analytical Methods, 6*, 2812. https://doi.org/10.1039/c3ay41907j

Brunet, J. P., Tamayo, P., Golub, T. R., & Mesirov, J. P. (2004). Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences of the United States of America*,*101*(12), 4164–4169. https://doi.org/10.1073/pnas.0308531101

Cadima, J., & Jolliffe, I. T. (1995b). Loadings and correlations in the interpretation of principal components. *Journal of Applied STATIStics, 22*, 203–214. https://doi.org/10.1080/757584614

Cai, D., He, X., Wu, X., & Han, J. (2008). Non-negative Matrix Factorization on Manifold. *2008 Eighth IEEE International Conference on Data Mining*, 63–72. https://doi.org/10.1109/ICDM.2008.57

Candes, E. J., & Romberg, J. K. (2005). Signal recovery from random projections. *Computational Imaging III, 5674*, 76. https://doi.org/10.1117/12.600722

Carlier, A., & Kroonenberg, P. M. (1996). Decompositions and Biplots in three-way correspondence analysis. *Psychometrika, 61*, 355–373.

Carlin, M., & Garcés de los Fayos, E. (2010). El síndrome de burnout: Evolución histórica desde el contexto laboral al ámbito deportivo. *Anales de Psicología, 26*(1), 169–180.

Carmona-Saez, P., Pascual-Marqui, R. D., Tirado, F., Carazo, J. M., & Pascual-Montano, A. (2006). Biclustering of gene expression data by Non-smooth Non-negative Matrix Factorization. *BMC Bioinformatics, 7*, 78. https://doi.org/10.1186/1471-2105-7-78

Carrasco, G., Molina, J.L., Patino-Alonso, M.C., Castillo, M. D. C., Vicente-Galindo, M.P., & Galindo-Villardón, M.-P. (2019). Water quality evaluation through a multivariate STATIStical HJ-Biplot approach. *Journal of Hydrology, 577*, 123993. https://doi.org/10.1016/j.jhydrol.2019.123993

Carrol, J., De Soete, G., & Pruzansky, S. (1989). Fitting of the Latent Class model via iteratively reweighted least squares CANDECOMP with nonnegativity constraints. In *Multiway data analysis* (pp. 463–472). North-Holland Publishing Co.

Carroll, J., & Chang, J. (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of eckart-young decomposition. *Psychometrika, 35*(3), 283-319. https://doi.org/10.1007/BF02310791

Chao, S. F., Mccallion, P., & Nickle, T. (2011). Factorial validity and consistency of the Maslach Burnout Inventory among staff working with persons with intellectual disability and dementia. *Journal of Intellectual Disability Research, 55*(5), 529–536. https://doi.org/10.1111/j.1365-2788.2011.01413.x

Chen, Y., He, W., Yokoya, N., & Huang, T.-Z. (2019). Hyperspectral Image Restoration Using Weighted Group Sparsity-Regularized Low-Rank Tensor Decomposition. *IEEE Transactions on Cybernetics*, 1–15. https://doi.org/10.1109/tcyb.2019.2936042

Cheney, W., & Goldstein, A. A. (1959). Proximity Maps for Convex Sets. *Proceedings of the American Mathematical Society, 10*(3), 448. https://doi.org/10.2307/2032864

Choi, S. (2008). Algorithms for orthogonal nonnegative matrix factorization. In *2008 Proceedings of the International Joint Conference on Neural Networks* (pp. 1828-1832). https://doi.org/10.1109/IJCNN.2008.4634046

Cichocki, A. (2013). Tensor Decompositions: A New Concept in Brain Data Analysis? arXiv:1305.0395

Cichocki, A., Zdunek, R., Phan, A. H., & Amari, S. I. (2009). *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-Way Data Analysis and Blind Source Separation*. In Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-Way Data Analysis and Blind Source Separation (Vol. 1). Chichester: John Wiley & Sons, Ltd.

Clemmensen, L., Hastie, T., Witten, D., & Ersbøll, B. (2011). Sparse Discriminant Analysis. *Technometrics, 53*(4), 406–413. https://doi.org/10.1198/TECH.2011.08118

Colombani, C., Croiseau, P., Fritz, S., Guillaume, F., Legarra, A., Ducrocq, V., & Robert-Granié, C. (2012). A comparison of partial least squares (PLS) and sparse PLS regressions in genomic selection in French dairy cattle. *Journal of Dairy Science, 95*(4), 2120–2131. https://doi.org/10.3168/jds.2011-4647

Conesa, A., Prats-Montalbán, J. M., Tarazona, S., Nueda, M. J., & Ferrer, A. (2010). A multiway approach to data integration in systems biology based on Tucker3 and N-PLS. *Chemometrics and Intelligent Laboratory Systems, 104*(1), 101–111. https://doi.org/10.1016/j.chemolab.2010.06.004

Cong, F., Lin, Q. H., Kuang, L. D., Gong, X. F., Astikainen, P., & Ristaniemi, T. (2015). Tensor decomposition of EEG signals: A brief review. *Journal of Neuroscience Methods*, *248*, 59–69. https://doi.org/10.1016/j.jneumeth.2015.03.018

Cormen, T., Leiserson, C., Rivest, R., & Stein, C. (2009). *Introduction to algorithms*. Cambridge, MA: MIT press.

Crammer, K., & Singer, Y. (2002). On the learnability and design of output codes for multiclass problems. *Machine Learning, 47*(2–3), 201–233. https://doi.org/10.1023/A:1013637720281

Cross, H. J. (1969). College students' memories of their parents: A factor analysis of the CRPBI. *Journal of Consulting and Clinical Psychology, 33*(3), 275–278. https://doi.org/10.1037/h0027589

Croux, C., Filzmoser, P., Fritz, H., Sm--, F., Croux, C., Filzmoser, P., & Fritz, H. (2011). Robust sparse principal component analysis Robust Sparse Principal Component Analysis. *Technology*, *55*(2), 202-214. https://doi.org/10.1080/00401706.2012.727746

Cubilla-Montilla, M., Galindo-Villardón, P., Nieto-Librero, A., Vicente, M., & García-Sánchez, I. (2019). What companies do not disclose about their environmental policy and what institutional pressures may do to respect. *Corporate Social Responsibility and Environmental Management*, 1–17. https://doi.org/10.1002/csr.1874

Cubilla-Montilla, M., Torres, C., Nieto-Librero, A., & Villardon, P. (2019). SparseBiplots package.

d'Aspremont, A., Ghaoui, L. El, Jordan, M., & Lanckriet, G. (2007). A direct formulation for sparse PCA using semidefinite programming. *SIAM Review, 49*(3), 434-448. https://doi.org/10.1137/050645506

Daubechies, I., Fornasier, M., & Loris, I. (2008). Accelerated projected gradient method for linear inverse problems with sparsity constraints. *Journal of Fourier Analysis and Applications, 14*(5–6), 764–792. https://doi.org/10.1007/s00041-008-9039-8

De Lathauwer, L. (2006). A link between the canonical decomposition in multilinear algebra and simultaneous matrix diagonalization. *SIAM Journal on Matrix Analysis and Applications, 28*(3), 642–666. https://doi.org/10.1137/040608830

De Lathauwer, L., De Moor, B., & Vandewalle, J. (2000). A multilinear singular value decomposition. SIAM *Journal on Matrix Analysis and Applications, 21*(4), 1253–1278. https://doi.org/10.1137/S0895479896305696

Deth, J. Van, Montero, J., & Westholm, A. (2007). *Citizenship and involvement in European democracies: A comparative analysis*. Abingdon: Routledge.

Díaz-Faes, A. A., González-Albo, B., Galindo, M. P., & Bordons, M. (2013). HJ-Biplot as tool of matrix inspection for bibliometrical data. *Revista Española de Documentación Científica, 36*, 1–16.

Ding, C., He, X., & Simon, H. D. (2005). On the Equivalence of Nonnegative Matrix Factorization and Spectral Clustering. *Proceedings of the Fifth SIAM International Conference on Data Mining (SDM), 4*, 606–610. https://doi.org/10.1137/1.9781611972757.70

Douglas Carroll, J., Pruzansky, S., & Kruskal, J. B. (1980). Candelinc: A general approach to multidimensional analysis of many-way arrays with linear constraints on parameters. *Psychometrika, 45*(1), 3–24. https://doi.org/10.1007/BF02293596

Drineas, P., Kannan, R., & Mahoney, M. W. (2006). Fast Monte-Carlo algorithms for matrices III: Computing a compressed approximate matrix decomposition. *SIAM Journal on Computing, 36*, 184–206. https://doi.org/10.1137/S0097539704442702

Drineas, P., Mahoney, M. W., & Muthukrishnan, S. (2007). Relative-Error CUR Matrix Decompositions. *SIAM Journal on Matrix Analysis and Applications, 30*(2), 40. https://doi.org/10.1137/07070471X

Duchi, J., Shalev-Shwartz, S., Singer, Y., & Chandra, T. (2008). Efficient projections onto the ℓ1-ball for learning in high dimensions. Proceedings of the 25th International Conference on Machine Learning, 272–279.

Eckart, C., & Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika, 1*(3), 211–218. https://doi.org/10.1007/BF02288367

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least Angle Regression. *Annals of STATIStics, 32*(2), 407–499.

Egido, J. (2017). dynBiplotGUI: Full Interactive GUI for Dynamic Biplot in R version 1.1.5 from CRAN.

Elder, J., Miner, G., Nisbet, B., Fast, A., Hill, T., & Delen, D. (2012). *Practical text mining and STATIStical analysis for non-structured text data applications*. Oxford: Academic Press.

Elias, M., & Sánchez-Gelabert, A. (2014). Connection between attitudes towards studies and learning actions among university students. *Revista de Estudios e Investigación en Psicología y Educación,1*, 3-14. https://doi.org/10.17979/reipe.2014.1.1.17

Engelen, S., & Hubert, M. (2011). Detecting outlying samples in a parallel factor analysis model. *Analytica Chimica Acta, 705*(1–2), 155–165. https://doi.org/10.1016/j.aca.2011.04.043

Engelhardt, B. E., & Stephens, M. (2010). Analysis of population structure: A unifying framework and novel methods based on sparse factor analysis. *PLoS Genetics, 6*(9). https://doi.org/10.1371/journal.pgen.1001117

Escofier, B., & Pagès, J. (1983). Méthode pour l'analyse de plusieurs groupes de variables. Application à la caractérisation de vins rouges du Val de Loire. *Revue de STATIStique Appliquée, 31*(2), 43–59.

Escofier, B., & Pagès, J. (1994). Multiple factor analysis (AFMULT package). *Computational STATIStics & Data Analysis, 18*(1), 121–140. https://doi.org/10.1016/0167-9473(94)90135-X

Escoufier, B., & Pagès, J. (1984). *Analyse factorielle multiple*. Paris: Cahiers du BURO.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American STATIStical Association, 96*(456), 1348–1360. https://doi.org/10.1198/016214501753382273

Faye-Dumanget, C., Carré, J., Le Borgne, M., & Boudoukha, P. A. H. (2017). French validation of the Maslach Burnout Inventory-Student Survey (MBI-SS). *Journal of Evaluation in Clinical Practice, 23*(6), 1247–1251. https://doi.org/10.1111/jep.12771

Feldt, T., Rantanen, J., Hyvönen, K., Mäkikangas, A., Huhtala, M., Pihlajasaari, P., & Kinnunen, U. (2014). The 9-item Bergen Burnout Inventory: factorial validity across organizations and measurements of longitudinal data. *Industrial Health, 52*, 102–112. https://doi.org/10.2486/indhealth.2013-0059

Févotte, C., Bertin, N., & Durrieu, J. L. (2009). Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis. *Neural Computation, 21*, 793–830. https://doi.org/10.1162/neco.2008.04-08-771

Filzmoser, P., Gschwandtner, M., & Todorov, V. (2012). Review of sparse methods in regression and classification with application to chemometrics. *Journal of Chemometrics, 26*(3–4), 42–51. https://doi.org/10.1002/cem.1418

Franceschini, A., Lin, J., von Mering, C., & Jensen, L. J. (2016). SVD-phy: improved prediction of protein functional associations through singular value decomposition of phylogenetic profiles. *Bioinformatics, 32*(7), 1085–1087. https://doi.org/10.1093/bioinformatics/btv696

Freudenberger, H. (1974). *The free clinic handbook*. Society for the Psychological Study of Social Issues.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *NIH Public Access, 33*(1), 1–20.

Frieze, A., Kannan, R., & Vempala, S. (2013). Fast Monte-Carlo algorithms for finding low-rank approximations. *Proceedings 39th Annual Symposium on Foundations of Computer Science, 53*, 370–378. https://doi.org/10.1109/SFCS.1998.743487

Frutos, E. (2015). *Análisis de datos acoplados: modelo T3-PCA*. Universidad de Salamanca.

Frutos, E., Galindo, M. P., & Leiva, V. (2014). An interactive Biplot implementation in R for modeling genotype-by-environment interaction. *Stochastic Environmental Research and Risk Assessment, 28*(7), 1629–1641. https://doi.org/10.1007/s00477-013-0821-z

Frutos, E., & Galindo, P. (2014). GGEBiplotGUI: interactive GGE Biplots in R.

Gabriel, K. R. (1971). The Biplot graphic display of matrices with application to principal component analysis. *Biometrika, 58*, 453–467.

Galindo, M. P. (1986). An alternative for simultaneous representation: HJ-Biplot. *Questíio: Quaderns d'Estadística, Sistemes, Informatica i Investigació Operativa, 10*, 12–23.

Gallo, M. (2015). Tucker3 Model for Compositional Data. *Communications in STATIStics - Theory and Methods, 44*(21), 4441–4453. https://doi.org/10.1080/03610926.2013.798664

Galloa, M., Todorovb, V., & Palmaa, M. Di. (2017). R Visual Tools for Three-way Data Analysis.

Gao, C., Ma, Z., & Zhou, H. H. (2017). Sparse CCA: Adaptive estimation and computational barriers. *Annals of STATIStics, 45*(5), 2074–2101. https://doi.org/10.1214/16-AOS1519

García, J. M., Herrero, S., & León, J. L. (2007). Validez factorial del Maslach Burnout Inventory (MBI) en una muestra de trabajadores del Hospital Psiquiátrico Penitenciario de Sevilla. *Apuntes de Psicología, 25*(2), 157–174.

García, M. I., Duarte, A. F., Rivera, O. I., Villalba, G. E., & Capacho, N. S. (2017). Learning approaches, academic performance and related factors; in students that curse last year of the programs of the faculty of health sciences. *Educación Médica*.

Gaujoux, R., & Seoighe, C. (2010). A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics, 11*, 367. https://doi.org/10.1186/1471-2105-11-367

Geraldo, J., del Rincón, B., & del Rincón, D. (2011). Estructura latente y consistencia interna del R-SPQ-2F: reinterpretando los enfoques de aprendizaje en el EEES. *Revista de Investigación Educativa, 29*(2), 277–293.

Gil-Monte, P. (2002). The factorial validity of the Maslach Burnout Inventory-General Survey. *Salud Publica de Mexico, 44*(1), 33–40.

Gillis, N. (2014). The Why and How of Nonnegative Matrix Factorization. 1–25.

Giordani, P., & Kiers, H. A. L. (2018). A review of tensor-based methods and their application to hospital care data. *STATIStics in Medicine, 37*(1), 137–156. https://doi.org/10.1002/sim.7514

Giordani, P., Kiers, H. A. L., & Del Ferraro, M. A. (2014). Three-way component analysis using the R package ThreeWay. *Journal of STATIStical Software, 57*(7), 1–23. https://doi.org/10.18637/jss.v057.i07

Giordani, P., & Rocci, R. (2016). Remedies for degeneracy in Candecomp/Parafac. Springer *Proceedings in Mathematics and STATIStics, 167*, 213–227. https://doi.org/10.1007/978-3-319-38759-8_16

Gligorijević, V., Malod-Dognin, N., & Pržulj, N. (2016). Integrative methods for analysing big data in precision medicine. *Proteomics, 16*, 741–758. https://doi.org/10.1002/pmic.201500396

Gloaguen, A., Guillemot, V., Tenenhaus, A., Gloaguen, A., Guillemot, V., & Tenenhaus, A. (2017). An efficient algorithm to satisfy l1 and l2 constraints. *49èmes Journées de STATIStique*. France

Gracia, R., Ferrer, J., Ayora, A., Alonso, M., Amutio, A., & Ferrer, R. (2019). Aplicación de un programa de mindfulness en profesionales de un servicio de medicina intensiva. Efecto sobre el burnout, la empatía y la autocompasión. *Medicina Intensiva, 43*(4), 207–216. https://doi.org/10.1016/J.MEDIN.2018.02.005

Greenacre, M. (2017). *Correspondence analysis in practice*. Boca Raton, Chapman and Hall/CRC.

Groves, M. (2005). Problem-based learning and learning approach: Is there a relationship? *Advances in Health Sciences Education, 10*(4), 315–326. https://doi.org/10.1007/s10459-005-8556-3

Guillemot, V., Beaton, D., Gloaguen, A., Löfstedt, T., Levine, B., Raymond, N., Tenenhaus, A., & Abdi, H. (2019). A constrained singular value decomposition method that integrates sparsity and orthogonality. *PLOS ONE, 14*(3), e0211463. https://doi.org/10.1371/journal.pone.0211463

Guo, J., James, G., Levina, E., Michailidis, G., & Zhu, J. (2010). Principal component analysis with sparse fused loadings. *Journal of Computational and Graphical STATIStics, 19*(4), 930-946. https://doi.org/10.1198/jcgs.2010.08127

Zou, H., Hastie, T. & Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical STATIStics, 15*(2), 265–286. https://doi.org/10.1198/106186006X113430

Haferlach, T., Kohlmann, A., Wieczorek, L., Basso, G., Te Kronnie, G., Béné, M. C., De Vos, J., Hernández, J. M., Hofmann, W. K., Mills, K. I., et al. (2010). Clinical utility of microarray-based gene expression profiling in the diagnosis and subclassification of leukemia: Report from the international microarray innovations in leukemia study group. *Journal of Clinical Oncology, 28*(15), 2529–2537. https://doi.org/10.1200/JCO.2009.23.4732

Halbesleben, J. R. B., & Demerouti, E. (2005). The construct validity of an alternative measure of burnout: Investigating the English translation of the Oldenburg Burnout Inventory. *Work & Stress, 19*(3), 208–220. https://doi.org/10.1080/02678370500340728

Hanafi, M., & Kiers, H. A. L. (2006). Analysis of K sets of data, with differential emphasis on agreement between and within sets. *Computational STATIStics and Data Analysis, 51*(3), 1491–1508. https://doi.org/10.1016/j.csda.2006.04.020

Harshman, R. A. (1978). Models for analysis of asymmetrical relationships among n objects or stimuli.

Harshman, R. A., & Lundy, M. E. (1984a). *Data preprocessing and the extended PARAFAC model*. In Research Methods for Multi-mode Data Analysis (pp. 216–284). New York: Praeger.

Harshman, R. A., & Lundy, M. E. (1984b). *The PARAFAC model for three-way factor analysis and multidimensional scaling*. In Research methods for multimode data analysis (pp. 122–215). Praeger, New York

Harshman, R. A (1970). Foundations of the PARAFAC procedure: Models and conditions for an “explanatory” multimodal factor analysis. *UCLA Working Papers in Phonetics, 16*(10), 1–84.

Hartigan, J. A. (1975). *Clustering Algorithms*. Wiley.

Hartingan, J. A. (1972). Direct Clustering of a Data Matrix. *Journal of the American STATIStical Society, 67*, 123–129. https://doi.org/10.1080/01621459.1972.10481214

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of STATIStical learning*. New York: Springer.

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *STATIStical learning with sparsity: the lasso and generalizations*. Boca Raton: Chapman and Hall/CRC.

Hausman, R. (1982). Constrained multivariate analysis. Optimization in STATIStics.

Hazan, E. (2006). Approximate convex optimization by online game playing. *arXiv cs/0610119*.

Heiler, M., & Schnörr, C. (2006). Learning Sparse Representations by Non-Negative Matrix Factorization and Sequential Cone Programming. *Journal of Machine Learning Research, 7*, 1385-1407.

Hernández-Pina, F., García-Sanz, M. P., & Maquilón-Sánchez, J. (2004). Análisis del cuestionario de procesos de estudio-2 factores de Biggs en estudiantes universitarios españoles. *Revista de La Facultad de Ciencias de La Educación, 6*, 117–138.

Hernández-Pina, F., Rodríguez, M., Ruiz, E., & Esquivel, J. (2010). Enfoques de aprendizaje en alumnos universitarios de la titulación de Ciencias de la Actividad Física y del Deporte de España y México. *Revista Iberoamericana de Educación, 53*(7), 1-11.

Hesterberg, T., Choi, N., Meier, L., & Fraley, C. (2008). Least angle and ℓ1 penalized regression: A review. *STATIStics Surveys, 2*, 61-93.

Hijazi, H., & Chan, C. (2013). A classification framework applied to cancer gene expression profiles. *Journal of Healthcare Engineering, 4*(2), 255–283. https://doi.org/10.1260/2040-2295.4.2.255

Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Application to nonorthogonal problems. *Technometrics, 12*(1), 69–82. https://doi.org/10.1080/00401706.1970.10488634

Hoerl, A., & Kennard, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55-67. https://doi.org/10.1080/00401706.1970.10488634#.XdpYyOhKhPY

Hore, V., Viñuela, A., Buil, A., Knight, J., McCarthy, M. I., Small, K., & Marchini, J. (2016). Tensor decomposition for multiple-tissue gene expression experiments. *Nature Genetics, 48*(9), 1094–1100. https://doi.org/10.1038/ng.3624

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

Hotelling, H. (1936). Relations between two sets of variables. *Biometrika, 28*(3–4), 321–377. https://doi.org/10.1093/biomet/28.3-4.321

Hoyer, P. (2002). Non-negative sparse coding. *Neural Networks for Signal Processing - Proceedings of the IEEE Workshop,* 557–565. https://doi.org/10.1109/NNSP.2002.1030067

Hoyer, P. (2004). Non-negative Matrix Factorization with Sparseness Constraints. *The Journal of Machine Learning Research, 5*, 1457–1469. https://doi.org/10.1109/ICMLC.2011.6016966

Hsieh, C. (2012). Burnout Among Public Service Workers: The Role of Emotional Labor Requirements and Job Resources. *Review of Public Personnel Administration, 34*(4), 379–402. https://doi.org/10.1177/0734371X12460554

Huang, H. H., Liu, X. Y., & Liang, Y. (2016). Feature selection and cancer classification via sparse logistic regression with the hybrid L1/2 +2 regularization. *PLoS ONE, 11*(5). https://doi.org/10.1371/journal.pone.0149675

Hussein, A. (2015). Arabic document similarity analysis using n-grams and singular value decomposition. *IEEE 9th International Conference on Research Challenges in Information Science (RCIS)*, 445–455.

Hyung, Z., Lee, K., & Lee, K. (2014). Music recommendation using text analysis on song requests to radio stations. *Expert Systems with Applications, 41*(5), 2608–2618. https://doi.org/10.1016/J.ESWA.2013.10.035

Ikemoto, H., & Adachi, K. (2016). Sparse Tucker2 analysis of three-way data subject to a constrained number of zero elements in a core array. *Computational STATIStics and Data Analysis, 98*, 1–18. https://doi.org/10.1016/j.csda.2015.12.007

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2014). *An Introduction to STATIStical Learning: with Applications in R.* New York: Springer.

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.

Jere, S., Dauwels, J., Asif, M. T., Vie, N. M., Cichocki, A., & Jaillet, P. (2014). Extracting commuting patterns in railway networks through matrix decompositions. *2014 13th International Conference on Control Automation Robotics and Vision, ICARCV 2014*, 541–546. https://doi.org/10.1109/ICARCV.2014.7064362

Jiang, T., & Sidiropoulos, N. D. (2004). Kruskal's Permutation Lemma and the identification of CANDECOMP/PARAFAC and bilinear models with constant modulus constraints. *IEEE Transactions on Signal Processing, 52*(9), 2625–2636. https://doi.org/10.1109/TSP.2004.832022

Jianqing, Fan; Runze, L. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American STATIStical Association, 96*(456), 1348–1360. https://doi.org/10.1198/016214501753382273#.XdpZ3-hKhPZ

Johnstone, I. M., & Lu, A. Y. (2009a). On Consistency and Sparsity for Principal Components Analysis in High Dimensions. *Journal of the American STATIStical Association, 104*(486), 682–693. https://doi.org/10.1198/jasa.2009.0121

Johnstone, I. M., & Lu, A. Y. (2009b). Sparse Principal Components Analysis. *Computers & Geosciences, 19*(3), 1–29. https://doi.org/10.1198/jasa.2009.0121

Jolliffe, I. (1995). Rotation of principal components: choice of normalization constraints. *Journal of Applied STATIStics, 22*(1), 29-35. https://doi.org/10.1080/757584395

Jolliffe, I. (2002). Principal component analysis. Aberdeen: Springer.

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

Jolliffe, I. T., Trendafilov, N. T., & Uddin, M. (2003). A modified principal component technique based on the LASSO. *Journal of Computational and Graphical STATIStics, 12*(3), 531–547. https://doi.org/10.1198/1061860032148

Journée, M., & Nesterov, Y. (2010). Generalized power method for sparse principal component analysis. *The Journal of Machine Learning Research, 11*, 517–553.

Juneja, A., Rana, B., & Agrawal, R. K. (2016). A combination of singular value decomposition and multivariate feature selection method for diagnosis of schizophrenia using fMRI. *Biomedical Signal Processing and Control, 27*, 122–133. https://doi.org/10.1016/j.bspc.2016.02.009

Kaiser, H. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika, 23*(3), 187-200. https://doi.org/10.1007/BF02289233

Kalliath, T., O'Driscoll, M., Gillespie, D., & Bluedorn, A. (2000). A test of the Maslach Burnout Inventory in three samples of healthcare professionals. *Work & Stress, 14*(1), 35–50. https://doi.org/10.1080/026783700417212

Kass, R. E., & Raftery, A. E. (2012). Bayes Factors. *Journal of the American STATIStical Association, 90*(430), 773-795. https://doi.org/ 10.1080/01621459.1995.10476572#.XdpbYuhKhPY

Kawash, G. F., & Clewes, J. L. (1988). A factor analysis of a short form of the crpbi: Are children's perceptions of control and discipline multidimensional? *Journal of Psychology: Interdisciplinary and Applied, 122*(1), 57–67. https://doi.org/10.1080/00223980.1988.10542943

Khamisa, N., Oldenburg, B., Peltzer, K., & Ilic, D. (2015). Work Related Stress, Burnout, Job Satisfaction and General Health of Nurses. *International Journal of Environmental Research and Public Health, 12*(1), 652–666. https://doi.org/10.3390/ijerph120100652

Khatavkar, R. (2007). *Sparse and orthogonal singular value decomposition*. Kansas State University

Kiers, H. A. L. (2000). Towards a standardized notation and terminology in multiway analysis. *Journal of Chemometrics, 14*(3), 105–122. https://doi.org/ 10.1002/1099-128X%28200005/06%2914%3A3<105%3A%3AAID-CEM582>3.0.CO%3B2-I

Kiers, H. A. L., & Berge, J. M. F. (1994). Hierarchical relations between methods for simultaneous component analysis and a technique for rotation to a simple simultaneous structure. *British Journal of Mathematical and STATIStical Psychology, 47*(1), 109–126. https://doi.org/10.1111/j.2044-8317.1994.tb01027.x

Kiers, H. A. L., & Smilde, A. K. (1998). Constrained three-mode factor analysis as a tool for parameter estimation with second-order instrumental data. *Journal of Chemometrics, 12*(2), 125–147. https://doi.org/10.1002/(SICI)1099-128X(199803/04)12:2<125::AID-CEM504>3.0.CO;2-D

Kim, H. J., Ollila, E., & Koivunen, V. (2013). Sparse regularization of tensor decompositions. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 3836–3840. https://doi.org/10.1109/ICASSP.2013.6638376

Kim, H. J., Ollila, E., Koivunen, V., & Croux, C. (2013). Robust and sparse estimation of tensor decompositions. *2013 IEEE Global Conference on Signal and Information Processing, GlobalSIP 2013 - Proceedings*, 965–968. https://doi.org/10.1109/GlobalSIP.2013.6737053

Kim, H., & Park, H. (2008). Nonnegative Matrix Factorization Based on Alternating Nonnegativity Constrained Least Squares and Active Set Method. *SIAM Journal on Matrix Analysis and Applications, 30*(2), 713–730. https://doi.org/10.1137/07069239X

Kim, H., Park, H., & Eldén, L. (2007). Non-negative tensor factorization based on alternating large-scale non-negativity-constrained least squares. *Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering, BIBE*, 1147–1151. https://doi.org/10.1109/BIBE.2007.4375705

Kim, J., & Park, H. (2008). *Sparse nonnegative matrix factorization for clustering*. Georgia Institute of Technology.

Kim, Y., Bismeijer, T., Zwart, W., Wessels, L., & Vis, D. (2019). Genomic data integration by WON-PARAFAC identifies interpretable factors for predicting drug-sensitivity in vivo. *Nature Communications, 10*(1), 5034. https://doi.org/10.1038/s41467-019-13027-2

Kim, Y. D., & Choi, S. (2007). Nonnegative tucker decomposition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/CVPR.2007.383405

Kojima, H. (1975). Inter-battery factor analysis of parents' and children's reports of parental behavior. *Japanese Psychological Research, 17*(1), 33–48. https://doi.org/10.4992/psycholres1954.17.33

Kolda, T., & Bader, B. (2009). Tensor Decompositions and Applications. *SIAM Review*, *51*(3), 455–500.

Krijnen, W. P., Dijkstra, T. K., & Stegeman, A. (2008). On the non-existence of optimal solutions and the occurrence of "degeneracy" in the CANDECOMP/PARAFAC model. *Psychometrika, 73*(3), 431–439. https://doi.org/10.1007/s11336-008-9056-1

Kristensen, T. S., Borritz, M., Villadsen, E., & Christensen, K. B. (2005). The Copenhagen Burnout Inventory: A new tool for the assessment of burnout. *Work & Stress, 19*(3), 192–207. https://doi.org/10.1080/02678370500297720

Kroonenberg, P. (2008). *Applied Multiway Data Analysis*. New Jersey: John Wiley & Sons

Kroonenberg, P., & Leeuw, J. De. (1980). Principal component analysis of three-mode data by means of alternating least squares algorithms. *Psychometrika, 45*(1), 69-97. https://doi.org/10.1007/BF02293599

Kroonenberg, P. M., Harshman, R. A., & Murakami, T. (2009). Analysing three-way profile data using the Parafac and Tucker3 models illustrated with views on parenting. *Applied Multivariate Research, 13*(2), 5. https://doi.org/10.22329/amr.v13i1.2833

Kroonenberg, P. M., & Leeuw, D. J. (1980). Principal Component Analysis of 3-Mode Data by Means of Alternating Least-Squares Algorithms. *Psychometrika, 45*(1), 69–97. https://doi.org/10.1007/Bf02293599

Kruskal, J. B. (1977). Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and STATIStics. *Linear Algebra and Its Applications, 18*(2), 95–138. https://doi.org/10.1016/0024-3795(77)90069-6

Kruskal, J. B. (1989). Rank*, decomposition, and uniqueness for 3-way and n-way arrays*. In Multiway data analysis (pp. 7–18).

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. New York, USA: Springer.

L'Hermier des Plantes, H. (1976). *Structuration Des Tableaux A Trois Indices De La STATIStique: theorie et application d'une méthode d'analyse conjointe*. Université Des Sciences et Techniques Du Languedoc, Montpellier.

Lai, Z., Xu, Y., Yang, J., Tang, J., & Zhang, D. (2013). Sparse tensor discriminant analysis. *IEEE Transactions on Image Processing, 22*(10), 3904–3915. https://doi.org/10.1109/TIP.2013.2264678

Lê Cao, K. A., Rossouw, D., Robert-Granié, C., & Besse, P. (2008). A sparse PLS for variable selection when integrating omics data. *STATIStical Applications in Genetics and Molecular Biology, 7*(1). https://doi.org/10.2202/1544-6115.1390

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature, 401*(6755), 788–791. https://doi.org/10.1038/44565

Lee, D. D., & Seung, H. S. (2001). *Algorithms for Non-negative Matrix Factorization.* In Advances in neural information processing systems (pp. 556-562). https://doi.org/10.1109/IJCNN.2008.4634046

Lee, M., Shen, H., Huang, J. Z., & Marron, J. S. (2010). Biclustering via Sparse Singular Value Decomposition*. Biometrics, 66*(4), 1087–1095. https://doi.org/10.1111/j.1541-0420.2010.01392.x

Lenhardt, L., Zeković, I., Dramićanin, T., Milićević, B., Burojević, J., & Dramićanin, M. D. (2017). Characterization of cereal flours by fluorescence spectroscopy coupled with PARAFAC. *Food Chemistry, 229*, 165–171. https://doi.org/10.1016/j.foodchem.2017.02.070

Li, B., Tian, B.-B., & Liu, J. (2016). A Simple Review of Sparse Principal Component Analysis. *Intelligent Computing Theories and Technology, 7996*, 443–449. https://doi.org/10.1007/978-3-642-39482-9

Li, G., & Xue, R. (2018). A New Privacy-Preserving Data Mining Method Using Non-negative Matrix Factorization and Singular Value Decomposition. *Wireless Personal Communications, 102*(2), 1799–1808. https://doi.org/10.1007/s11277-017-5237-5

Li, S. Z., Hou, X. W., Zhang, H. J., & Cheng, Q. S. (2001). Learning spatially localized, parts-based representation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1*. https://doi.org/10.1109/cvpr.2001.990477

Li, T. (2005). A general model for clustering binary data. *Proceeding of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining - KDD '05,* 188. https://doi.org/10.1145/1081870.1081894

Li, X. G., Lv, X. L., & Zhang, Y. (2013). Application of PARAFAC method in petroleum organic measurement and analysis. *Advanced Materials Research, 753*, 2269–2272. https://doi.org/10.4028/www.scientific.net/AMR.753-755.2269

Li, X., Xu, D., Zhou, H., & Li, L. (2018). Tucker Tensor Regression and Neuroimaging Analysis. *STATIStics in Biosciences, 10*(3), 520–545. https://doi.org/10.1007/s12561-018-9215-6

Li, Y., & Ngom, A. (2012). A new Kernel non-negative matrix factorization and its application in microarray data analysis. *2012 IEEE Symposium on Computational Intelligence and Computational Biology, CIBCB 2012,* 371–378. https://doi.org/10.1109/CIBCB.2012.6217254

Lisowska, K. M., Olbryt, M., Student, S., Kujawa, K. A., Cortez, A. J., Simek, K., … Kupryjańczyk, J. (2016). Unsupervised analysis reveals two molecular subgroups of serous ovarian cancer with distinct gene expression profiles and survival. *Journal of Cancer Research and Clinical Oncology, 142*(6), 1239–1252. https://doi.org/10.1007/s00432-016-2147-y

Liu, C., Harley, J., Bergés, M., Greve, D., & Oppenheim, I. (2015). Robust ultrasonic damage detection under complex environmental conditions using singular value decomposition. *Ultrasonics, 58*, 75–86.

Liu, J., Liu, J., Wonka, P., & Ye, J. (2012). Sparse non-negative tensor factorization using columnwise coordinate descent. *Pattern Recognition, 45*(1), 649–656. https://doi.org/10.1016/j.patcog.2011.05.015

Liu, X., Wang, S., Zhang, H., Zhang, H., Yang, Z.-Y., & Liang, Y. (2019). Novel regularization method for biomarker selection and cancer classification. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. https://doi.org/10.1109/tcbb.2019.2897301

Lock, E. (2012). *Vertical integration of multiple high-dimensional datasets*. University of North Carolina.

Lock, E. F., Hoadley, K. A., Marron, J. S., & Nobel, A. B. (2013). Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *Annals of Applied STATIStics, 7*(1), 523–542. https://doi.org/10.1214/12-AOAS597

Lock, E., & Li, G. (2018). Supervised multiway factorization. *Electronic Journal of STATIStics, 12*(1), 1150. https://doi.org/10.1214/18-EJS1421

Loera, B., Converso, D., & Viotti, S. (2014). Evaluating the Psychometric Properties of the Maslach Burnout Inventory-Human Services Survey (MBI-HSS) among Italian Nurses: How Many Factors Must a Researcher Consider? *PLoS ONE, 9*(12), e114987. https://doi.org/10.1371/journal.pone.0114987

López, H., Pedrosa, I., Vicente-Galindo, M., Suárez-Álvarez, J., Galindo-Villardón, M., & García-Cueto, E. (2014). Multivariate analysis of burnout syndrome in latin-american priests. *Psicothema , 26*(2), 227–234. https://doi.org/10.7334/psicothema2013.178.

Louis, D. N., Perry, A., Reifenberger, G., von Deimling, A., Figarella-Branger, D., Cavenee, W. K., … Ellison, D. W. (2016). The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. *Acta Neuropathologica, 131*(6), 803–820. https://doi.org/10.1007/s00401-016-1545-1

Lu, H., Plataniotis, K. N., Venetsanopoulos, A., & More, & 0. (2013). *Multilinear Subspace Learning: Dimensionality Reduction of Multidimensional Data*. Boca Raton: Chapman & Hall/CRC

Lundy, M., Harshamn, R., & Kruskal, J. (1989). *A two-stage procedure incorporating good features of both trilinear and quadrilinear models*. In Elsevier (Ed.), Multiway Data Analysis (pp. 123–130).

Lykou, A., & Whittaker, J. (2010). Sparse CCA using a lasso with positivity constraints. *Computational STATIStics and Data Analysis, 54*(12), 3144–3157. https://doi.org/10.1016/j.csda.2009.08.002

Maculan, N., & de Paula, G. (1989). A linear-time median-finding algorithms for projecting a vector on the simplex of Rn. *Operations Research Letters, 8*(1989), 219–222.

Mahoney, M. W., & Drineas, P. (2009). CUR matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences of the United States of America, 106*(3), 697–702. https://doi.org/10.1073/pnas.0803205106

Mairal, J., Bach, F., Ponce, J., & Sapiro, G. (2009). Online dictionary learning for sparse coding. *Proceedings of the 26th International Conference on Machine Learning*, 1–8. https://doi.org/10.1145/1553374.1553463

Mairal, J., Bach, F., Ponce, J., & Sapiro, G. (2010). Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research, 11*, 19–60. https://doi.org/10.1145/1756006.1756008

Margolies, P. J., & Weintraub, S. (1977). The revised 56-item CRPBI as a research instrument: Reliability and factor structure*. Journal of Clinical Psychology, 33*(2), 472–476. https://doi.org/10.1002/1097-4679(197704)33:2<472::AID-JCLP2270330230>3.0.CO;2-S

Martínez-Montes, E., Sánchez-Bornot, J., & Valdés-Sosa, P. (2008). Penalized PARAFAC analysis of spontaneous EEG recordings. *STATIStica Sinica*, *18*(4), 1449–1464.

Marton, F., & Säljö, R. (1976). On qualitative differences in learning : i-outcome and process. *British Journal of Educational Psychology, 46*, 4–11. https://doi.org/10.1111/j.2044-8279.1976.tb02980.x

Maslach, C., & Jackson, S. (1981). *MBI: Maslach burnout inventory*. Palo Alto: Consulting Psychologists Press.

Maslach, C., Jackson, S., Leiter, M., & Schaufeli, W. (1986). *Maslach burnout inventory*. Palo Alto: Consulting Psychologists Press

McCabe, G. (1984). Principal variables. *Technometrics, 26*(2), 137-144. https://doi.org/10.1080/00401706.1984.10487939

McDonald, R. (1999). *Test theory: A unified treatment*. Taylor & Francis.

Meier, L. (2008). The group lasso for logistic regression. *Journal of the Royal STATIStical Society, 70*(1), 53-71. https://doi.org/10.1111/j.1467-9868.2007.00627.x/full

Mellers, B., Stone, E., Atanasov, P., Rohrbaugh, N., Emlen, S., Ungar, L., Bishop, M. M., Horowitz, M., & Tetlock, P. (2015). The psychology of intelligence analysis: Drivers of prediction accuracy in world politics. *Journal of Experimental Psychology: Applied, 21*(1), 1–14. https://doi.org/10.1037/xap0000040

Mendes, S., Fernández-Gómez, M. J., Galindo, M. P., Morgado, F., Maranhão, P., Azeiteiro, U., & Bacelar-Nicolau, P. (2009). The study of bacterioplankton dynamics in the Berlengas Archipelago (West coast of Portugal) by applying the HJ-Biplot method. Arquipelago Life and Marine Sciences, 26, 25–35.

Mészáros, V., Ádám, S., Szabó, M., Szigeti, R., & Urbán, R. (2014). The Bifactor Model of the Maslach Burnout Inventory-Human Services Survey (MBI-HSS)-An Alternative Measurement Model of Burnout. *Stress and Health, 30*(1), 82–88. https://doi.org/10.1002/smi.2481

Meuwissen, T. H. E., Indahl, U. G., & Ødegård, J. (2017). Variable selection models for genomic selection using whole-genome sequence data and singular value decomposition. *Genetics Selection Evolution, 49*(1), 94. https://doi.org/10.1186/s12711-017-0369-3

Miyoshi, T., Tanioka, K., Yamamoto, S., Yadohisa, H., Hiroyasu, T., & Hiwa, S. (2019). Short-term effects on brain functional network caused by focused-attention meditation revealed by Tucker3 clustering on graph theoretical metrics. *BioRxiv*, 765693. https://doi.org/10.1101/765693

Moghaddam, B., Weiss, Y., & Avidan, S. (2007). Spectral method for sparse linear discriminant analysis.

Mondéjar-Jiménez, J., & Vargas-Vargas, M. (2010). Determinant factors of attitude towards quantitative subjects: differences between sexes. *Teaching and Teacher Education, 26*(3), 688–693. https://doi.org/10.1016/j.tate.2009.10.004

Mondéjar Jimenez, J., Vargas, M., & Bayot Mestre, A. (2008). Medición de la actitud hacia la estadística. Influencia de los procesos de estudio. *Electronic Journal of Research in Educational Phychology, 6*(3), 729–748. https://doi.org/10.25115/ejrep.v6i16.1303

Mørup, M., Hansen, L. K., & Arnfred, S. M. (2008). Algorithms for sparse nonnegative tucker decompositions. *Neural Computation, 20*(8), 2112–2131. https://doi.org/10.1162/neco.2008.11-06-407

Muñoz, J., & Mato, M. D. (2008). Análisis de las actitudes respecto a las matemáticas en alumnos de ESO. *Revista de Investigación Educativa, 26*(1), 209–226.

Muris, P. (2010). *Normal and abnormal fear and anxiety in children and adolescents*. Elsevier.

Nerici, I. G. (1969). *Hacia una didáctica general dinámica*. Buenos Aires: Kapelusz.

Nieto-Librero, A. B., Sierra, C., Vicente-Galindo, M. P., Ruíz-Barzola, O., & Galindo-Villardón, M. P. (2017). Clustering Disjoint HJ-Biplot: A new tool for identifying pollution patterns in geochemical studies. *Chemosphere, 176*, 389–396. https://doi.org/10.1016/j.chemosphere.2017.02.125

Nieto-Librero, A. B., Galindo-Villardón, M., Leiva, V., & Vicente-Galindo, M. (2014). A Methodology for Biplots based on bootstrapping with R. *Revista Colombiana de Estadística, 37*(2), 367-397. https://doi.org/10.15446/rce.v37n2spe.47944

Nieto-Librero, A. B. (2015). *Versión inferencial de los métodos Biplot basada en remuestreo bootstrap y su aplicación a tablas de tres vías*. Universidad de Salamanca.

Ning-min, S., & Jing, L. (2015). A Literature Survey on High-Dimensional Sparse Principal Component Analysis. *International Journal of Database Theory and Application, 8*(6), 57–74. https://doi.org/10.14257/ijdta.2015.8.6.06

O'Connell, M. J., & Lock, E. F. (2016). R.JIVE for exploration of multi-source molecular data. *Bioinformatics, 32*(18):2877-9. https://doi.org/10.1093/bioinformatics/btw324

Ortas, E., Álvarez, I., Jaussaud, J., & Garayar, A. (2015). The impact of institutional and social context on corporate environmental, social and governance performance of companies committed to voluntary corporate social responsibility initiatives. *Journal of Cleaner Production, 108*, 673–684. https://doi.org/10.1016/j.jclepro.2015.06.089

Paatero, P., & Tapper, U. (1994). Positive Matrix Factorization - A Nonnegative Factor Model with Optimal Utilization of Error Estimates of Data Values. *Environmetrics, 5*(2), 111–126. https://doi.org/10.1002/env.3170050203

Papalexakis, E. E., Sidiropoulos, N. D., & Bro, R. (2013). From K-means to higher-way co-clustering: Multilinear decomposition with sparse latent factors. *IEEE Transactions on Signal Processing, 61*(2), 593-506. https://doi.org/10.1109/TSP.2012.2225052

Paul, L. C., & Al Sumam, A. (2012). Face recognition using principal component analysis method. *International Journal of Advanced Research in Computer Engineering & Technology, 1*(9), 135–139.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine, 2*(6), 559–572. https://doi.org/10.1080/14786440109462720

Peharz, R., & Pernkopf, F. (2012). Sparse nonnegative matrix factorization with ℓ0-constraints. *Neurocomputing, 80*, 38–46. https://doi.org/10.1016/j.neucom.2011.09.024

Potluru, V. K., Plis, S. M., Roux, J. Le, Pearlmutter, B. A., Calhoun, V. D., & Hayes, T. P. (2013). Block Coordinate Descent for Sparse NMF. *arXiv: 1301.3527.*

Prosser, M., & Trigwell, K. (2014). Qualitative variation in approaches to university teaching and learning in large first-year classes. *Higher Education, 67*(6), 783–795. https://doi.org/10.1007/s10734-013-9690-0

Puntanen, S. (2011). Projection Matrices, Generalized Inverse Matrices, and Singular Value Decomposition by Haruo Yanai, Kei Takeuchi, Yoshio Takane. *International STATIStical Review, 79*(3), 503–504. https://doi.org/10.1111/j.1751-5823.2011.00159_24.x

Qi, X., Luo, R., & Zhao, H. (2013). Sparse principal component analysis by choice of norm. *Journal of Multivariate Analysis, 114*, 127-160. https://doi.org/10.1016/j.jmva.2012.07.004

Qian, Y., Jia, S., Zhou, J., & Robles-Kelly, A. (2011). Hyperspectral unmixing via L1/2 sparsity-constrained nonnegative matrix factorization. *IEEE Transactions on Geoscience and Remote Sensing, 49*(11), 4287–4297. https://doi.org/10.1109/TGRS.2011.2144605

Ramsay, J. O., & Silverman, B. W. (2005). *Functional data analysis*. Springer.

Ricci, G., De Gemmis, M., & Semeraro, G. (2012). Matrix and Tensor Factorization Techniques applied to Recommender Systems: a Survey. *International Journal of Computer and Information Technology, 1(*1), 94-98.

Rocci, R., & Giordani, P. (2010). A weak degeneracy revealing decomposition for the CANDECOMP/PARAFAC model. *Journal of Chemometrics*, *24*(2), 57-66. https://doi.org/10.1002/cem.1272

Rodríguez-Rosa, M., Gallego-Álvarez, I., & Galindo-Villardón, M. P. (2019). Spatio-temporal analysis of economic, social, and environmental issues in the framework of sustainable development in worldwide countries. *Sustainable Development, 27*(3), 429–447. https://doi.org/10.1002/sd.1916

Roux, J. Le, Weninger, F., & Hershey, J. R. (2015). Sparse NMF – half-baked or well done? *Mitsubishi Electric Research Labs (MERL)*, no. TR2015-023.

Sádecká, J., Uríčková, V., Hroboňová, K., & Májek, P. (2015). Classification of Juniper-Flavoured Spirit Drinks by Multivariate Analysis of Spectroscopic and Chromatographic Data. *Food Analytical Methods, 8*(1), 58–69. https://doi.org/10.1007/s12161-014-9869-8

Salakhutdinov, R., Roweis, S., & Ghahramani, Z. (2002). On the convergence of bound optimization algorithms. In *Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence* (pp. 509-516). Morgan Kaufmann Publishers Inc.

Samaranayake, D., & Seneviratne, S. (2012). Validity of the Maslach Burnout Inventory – Human Services Survey among Sri Lankan Nursing Officers. *Psychological Studies, 57*(1), 101–111. https://doi.org/10.1007/s12646-011-0135-5

Săndica, A. M., Dudian, M., & Ştefănescu, A. (2018). Air pollution and human development in Europe: A new index using principal component analysis. *Sustainability*, *10*(2). https://doi.org/10.3390/su10020312

Sanguansat, P. (2012). *Principal Component Analysis: Engineering Applications*. InTech

Sauzay L, Hanafi M, Qannari EM, Schlich P. Analysede K+1 tableaux à l'aide de la méthode STATIS:application en évaluation sensorielle, *9ième Journées Européennes Agro-industrie et Méthodes STATIStiques*. Montpellier (France). 1 – 23

Sch, J. (2005). A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics, *STATIStical applications in genetics and molecular biology, 4*(1). https://doi.org/10.2202/1544-6115.1175

Schaefer, E. S. (1965). Children's Reports of Parental Behavior: An Inventory. *Child Development, 36*(2), 413. https://doi.org/10.2307/1126465

Schaufeli, W., Martinez, I., Pinto, A., Salanova, M., & Bakker, A. (2002). Burnout and engagement in university students: A cross-national study. *Journal of Cross-Cultural Psychology, 33*(5), 464–481. https://doi.org/10.1177/0022022102033005003

Schaufeli WB, Leiter MP, Maslach C, & Jackson SE. (1996). The Maslach Burnout Inventory: General Survey (MBI-GS).

Schludermann, E., & Schludermann, S. (1970). Replicability of factors in children's report of parent behavior (CRPBI). *Journal of Psychology: Interdisciplinary and Applied, 76*(2), 239–249. https://doi.org/10.1080/00223980.1970.9916845

Schwarz, N., & Bohner, G. (2001). The construction of attitudes. *Blackwell handbook of social psychology: Intraindividual processes, 1*, 436-457.

Shao, J., Wang, Y., Deng, X., & Wang, S. (2011). Sparse linear discriminant analysis by thresholding for high dimensional data. *The Annals of STATIStics, 39*(2), 1241–1265. https://doi.org/10.1214/10-aos870

Shen, H., & Huang, J. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis, 99*(6), 1015-1034. https://doi.org/10.1016/j.jmva.2007.06.007

Shirom, A., & Melamed, S. (2006). A comparison of the construct validity of two burnout measures in two groups of professionals. *International Journal of Stress Management, 13*(2), 176–200. https://doi.org/10.1037/1072-5245.13.2.176

Sidiropoulos, N. D., & Bro, R. (2000). On the uniqueness of multilinear decomposition of N-way arrays. *Journal of Chemometrics, 14*(3), 229–239. https://doi.org/10.1002/1099-128X(200005/06)14:3<229::AID-CEM587>3.0.CO;2-N

Silva, A. D., Taveira, M. do C., Marques, C., & Gouveia, V. V. (2015). Satisfaction with Life Scale Among Adolescents and Young Adults in Portugal: Extending Evidence of Construct Validity. *Social Indicators Research, 120*(1), 309–318. https://doi.org/10.1007/s11205-014-0587-9

Simon, N., Friedman, J., & Hastie, T. (2014). A Blockwise Descent Algorithm for Group-penalized Multiresponse and Multinomial Regression. *Journal of STATIStical Software, 20*(2), 1-15.

Singh, K. P., Malik, A., Singh, V. K., Basant, N., & Sinha, S. (2006). Multi-way modeling of hydro-chemical data of an alluvial river system-A case study. *Analytica Chimica Acta, 571*(2), 248–259. https://doi.org/10.1016/j.aca.2006.04.080

Skillicorn, D., & Leuprecht, C. (2015). Deception in speeches of candidates for public office. *Journal of Data Mining and Digital Humanities, 43*.

Smilde, A, Bro, R., & Geladi, P. (2004). *Multi-way Analysis with Applications in the Chemical Sciences*. Chichester: John Wiley & Sons

Smilde, A., Måge, I., Næs, T., Hankemeier, T., Lips, M., Kiers, H., Acar, E. & Bro, R. (2017). Common and distinct components in data fusion. *Journal of Chemometrics, 31*(7). https://doi.org/10.1002/cem.2900

Spearman, C. (1904). “General Intelligence”, Objectively Determined and Measured. T*he American Journal of Psychology, 15*(2), 201–292.

Stegeman, A. (2007). Degeneracy in candecomp/parafac and indscal explained for several three-sliced arrays with a two-valued typical rank. *Psychometrika, 72*(4), 601–619. https://doi.org/10.1007/s11336-007-9022-3

Stegeman, A., Berge, J. M. F. Ten, & Lathauwer, L. De. (2006). Sufficient conditions for uniqueness in Candecomp/Parafac and Indscal with random component matrices. *Psychometrika, 71*(2), 219–229. https://doi.org/10.1007/s11336-006-1278-2

Stegeman, A., & Ten Berge, J. M. F. (2006). Kruskal’s condition for uniqueness in Candecomp/Parafac when ranks and k-ranks coincide. *Computational STATIStics and Data Analysis, 50*, 210–220. https://doi.org/10.1016/j.csda.2004.07.015

Strazar, M., Zitnik, M., Zupan, B., Ule, J., & Curk, T. (2016). Orthogonal matrix factorization enables integrative analysis of multiple RNA binding proteins. *Bioinformatics, 32*(10), 1527–1535. https://doi.org/10.1093/bioinformatics/btw003

Sun, W. W., Lu, J., Liu, H., & Cheng, G. (2017). Provable sparse tensor decomposition. *Journal of the Royal STATIStical Society: Series B (STATIStical Methodology), 79*(3), 899–916. https://doi.org/10.1111/rssb.12190

Team, R. C. (2019). R: A language and environment for STATIStical computing. https://www.r-project.org/

Ten Berge, J. M. F., & Sidiropoulos, N. D. (2002). On uniqueness in CANDECOMP/PARAFAC. *Psychometrika, 67*(3), 399–409. https://doi.org/10.1007/BF02294992

Thara, S., & Sidharth, S. (2017). Aspect based sentiment classication: Svd features. *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2370–2374. https://doi.org/10.1109/ICACCI.2017.8126201

Theis, F. J., Stadlthanner, K., & Tanaka, T. (2005, September). First results on uniqueness of sparse non-negative matrix factorization. In *2005 13th European Signal Processing Conference* (pp. 1-4).

Thurstone, L. (1935). *The vectors of mind.* The University of Chicago science series

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal STATIStical Society, Series B, 58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tomasi, G., & Bro, R. (2006). A comparison of algorithms for fitting the PARAFAC model. *Computational STATIStics and Data Analysis, 50*(7), 1700–1734. https://doi.org/10.1016/j.csda.2004.11.013

Torres-Salinas, D., Robinson-García, N., Jiménez-Contreras, E., Herrera, F., & López-Cózar, E. D. (2013). On the use of Biplot analysis for multivariate bibliometric and scientific indicators. *Journal of the American Society for Information Science and Technology, 64*(7), 1468–1479. https://doi.org/10.1002/asi.22837

Trendafilov, N. T. (2014). From simple structure to sparse components: A review. *Computational STATIStics, 29,* 431–454. https://doi.org/10.1007/s00180-013-0434-5

Trendafilov, N. T., Unkel, S., & Krzanowski, W. (2013). Exploratory factor and principal component analyses: Some new aspects. *STATIStics and Computing, 23*(2), 209–220. https://doi.org/10.1007/s11222-011-9303-7

Trygg, J. (2002). O2-PLS for qualitative and quantitative analysis in multivariate calibration. *Journal of Chemometrics, 16*(6), 283–293. https://doi.org/10.1002/cem.724

Tucker, L. (1972). Relations between multidimensional scaling and three-mode factor analysis. *Psychometrika, 37*(1), 3-27. https://doi.org/10.1007/BF02291410

Tucker, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika, 31*(3), 279–311. https://doi.org/10.1007/BF02289464

Ullah, R. (2016). Learning environment, approaches to learning and learning preferences: medical students versus general education students. *Journal of Pakistan Medical Association, 16*(66), 541–544.

Valiente, R. M., Magaz, A., Chorot, P., & Sandín, B. (2016). Estructura factorial del cuestionario de percepción de estilos de crianza CRPBIAbreviado. 3, 69–78.

Vallejo-Arboleda, A., Vicente-Villardón, J. L., & Galindo-Villardón, M. P. (2007). Canonical STATIS: Biplot analysis of multi-table group structured data based on STATIS-ACT methodology. *Computational STATIStics and Data Analysis, 51*(9), 4193–4205. https://doi.org/10.1016/j.csda.2006.04.032

Van Benthem, M. H., Lane, T. W., Davis, R. W., Lane, P. D., & Keenan, M. R. (2011). PARAFAC modeling of three-way hyperspectral images: Endogenous fluorophores as health biomarkers in aquatic species. *Chemometrics and Intelligent Laboratory Systems, 106*(1), 115–124. https://doi.org/10.1016/j.chemolab.2010.09.003

van den Berg, E., & Friedlander, M. P. (2009). Probing the Pareto Frontier for Basis Pursuit Solutions. *SIAM Journal on Scientific Computing, 31*(2), 890–912. https://doi.org/10.1137/080714488

van der Kloet, F. M., Sebastián-León, P., Conesa, A., Smilde, A. K., & Westerhuis, J. A. (2016). Separating common from distinctive variation. *BMC Bioinformatics, 17*(S5), S195. https://doi.org/10.1186/s12859-016-1037-2

Van Deun, K., Smilde, A. K., Thorrez, L., Kiers, H. A. L., & Van Mechelen, I. (2013). Identifying common and distinctive processes underlying multiset data. *Chemometrics and Intelligent Laboratory Systems, 129*, 40–51. https://doi.org/10.1016/j.chemolab.2013.07.005

van Deun, K., van Mechelen, I., Thorrez, L., Schouteden, M., de Moor, B., van der Werf, M. J., … Kiers, H. A. L. (2012). DISCO-SCA and properly applied GSVD as swinging methods to find common and distinctive processes. *PLoS ONE, 7*(5). https://doi.org/10.1371/journal.pone.0037840

Vázquez, J., Vicente-Galindo, M., & Galindo, M. (2011). Variables que inciden en la seguridad de las escuelas públicas de los Estados Unidos. *Revista de Pedagogía, 44*(1), 141–165.

Vega-Hernández, M. C., Patino-Alonso, M. C., & Galindo-Villardón, M. P. (2018). Multivariate characterization of university students using the ICT for learning. *Computers and Education, 121*, 124–130. https://doi.org/10.1016/j.compedu.2018.03.004

Vellone, E., Barbaranelli, C., Lee, C., & Riegel, B. (2015). Measures of self-care in heart failure: Issues with factorial structure and reliability. *Heart & Lung, 44*(1), 82–83. https://doi.org/10.1016/j.hrtlng.2014.08.008

Vesty, G., Sridharan, V. G., Northcott, D., & Dellaportas, S. (2018). Burnout among university accounting educators in Australia and New Zealand: determinants and implications. *Accounting and Finance, 58,* 255–277. https://doi.org/10.1111/acfi.12203

Vichi, M., & Kiers, H. A. L. (2001). Factorial k-means analysis for two-way data. *Computational STATIStics and Data Analysis, 37*, 49–64. https://doi.org/10.1016/S0167-9473(00)00064-5

Vichi, M., & Saporta, G. (2009). Clustering and disjoint principal component analysis. *Computational STATIStics & Data Analysis, 53*(8), 3194-3208. https://doi.org/10.1016/j.csda.2008.05.028

Villegas Barahona, G. (2018). Modelo estadístico pedagógico para la toma de decisiones administrativas y académicas con impacto en el mejoramiento continuo del rendimiento de los estudiantes universitarios, basado en los métodos de selección CUR. Universidad de Salamanca.

Vines, S. (2000). Simple principal components. *Journal of the Royal STATIStical Society: Series C, 49*(4), 441-451. https://doi.org/10.1111/1467-9876.00204/abstract

Wang, G., Kossenkov, A. V., & Ochs, M. F. (2006). LS-NMF: A modified non-negative matrix factorization algorithm utilizing uncertainty estimates. *BMC Bioinformatics, 7*, 1–10. https://doi.org/10.1186/1471-2105-7-175

Wang, H., Li, R., & Tsai, C. L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika, 94*(3), 553–568. https://doi.org/10.1093/biomet/asm053

Wang, S. H., Zhan, T. M., Chen, Y., Zhang, Y., Yang, M., Lu, H. M., Wang, H-N, Liu, B. & Phillips, P. (2016). Multiple Sclerosis Detection Based on Biorthogonal Wavelet Transform, RBF Kernel Principal Component Analysis, and Logistic Regression. *IEEE Access, 4*, 7567–7576. https://doi.org/10.1109/ACCESS.2016.2620996

Wang, Z., Yuan, W., & Montana, G. (2015). Sparse multi-view matrix factorization: A multivariate approach to multiple tissue comparisons. *Bioinformatics, 31*(19), 3163–3171. https://doi.org/10.1093/bioinformatics/btv344

Wesseling, P., & Capper, D. (2018). WHO 2016 Classification of gliomas. Neuropathology and Applied Neurobiology, 44(2), 139–150. https://doi.org/10.1111/nan.12432

West, C. P., Dyrbye, L. N., & Shanafelt, T. D. (2018). Physician burnout: contributors, consequences and solutions. *Journal of Internal Medicine, 283*(6), 516–529. https://doi.org/10.1111/joim.12752

Wijaya, M. E., Billah, M. S., & Ahn, H. (2018). Political attitude estimation through Facebook like: a South Korean case study. Asian Journal of Political Science, 26(1), 87–102. https://doi.org/10.1080/02185377.2017.1402357

Wild, K., Scholz, M., Ropohl, A., Bräuer, L., Paulsen, F., & Burger, P. (2014). Strategies against Burnout and Anxiety in Medical Education – Implementation and Evaluation of a New Course on Relaxation Techniques (Relacs) for Medical Students. *PLoS ONE, 9*(12), e114967. https://doi.org/10.1371/journal.pone.0114967

Wise, S. (1985). The development and validation of a scale measuring attitudes toward STATIStics. *Educational and psychological measurement, 45*(2), 401-405. https://doi.org/10.1177/001316448504500226

Witriw, A., Molina, S., & Ferrari, M. (2014). Enfoques de aprendizaje utilizados por estudiantes universitarios en las Áreas Básica y Gestión-Alimentos de la carrera de Nutrición de la UBA. *Revista Argentina de Educación Superior, 6*(9), 195–207.

Witten, D. M., & Tibshirani, R. (2010). A framework for feature selection in clustering. *American STATIStician, 105*(490), 713–726. https://doi.org/10.1198/jasa.2010.tm09415.A

Witten, D. M., Tibshirani, R., & Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *BioSTATIStics, 10*, 515–534. https://doi.org/10.1093/bioSTATIStics/kxp008

Wong, K. K., Rostomily, R., & Wong, S. T. C. (2019). Prognostic gene discovery in glioblastoma patients using deep learning. *Cancers, 11*(1), 1–15. https://doi.org/10.3390/cancers11010053

Xu, W., Liu, X., & Gong, Y. (2003). Document clustering based on non-negative matrix factorization. *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval - SIGIR '03, 267*. https://doi.org/10.1145/860435.860485

Yang, Z., & Michailidis, G. (2015). A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. *Bioinformatics, 32*(1), 1–8. https://doi.org/10.1093/bioinformatics/btv544

Yavuz, G., & Doğan, N. (2014). Maslach Burnout Inventory-Student Survey (MBI-SS): A Validity Study. *Procedia - Social and Behavioral Sciences, 116*(2014), 2453–2457. https://doi.org/10.1016/j.sbspro.2014.01.590

Ye, J., & Jin, Z. (2013). Nonnegative matrix factorization on orthogonal subspace with smoothed L0 norm constrained. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 7751* LNCS, 1–7. https://doi.org/10.1007/978-3-642-36669-7_1

Yoo, J., & Choi, S. (2010). Orthogonal nonnegative matrix tri-factorization for co-clustering: Multiplicative updates on Stiefel manifolds. *Information Processing and Management, 46*(5), 559–570. https://doi.org/10.1016/j.ipm.2009.12.007

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal STATIStical Society: Series B (STATIStical Methodology), 68*(1), 49–67. https://doi.org/10.1111/j.1467-9868.2005.00532.x

Zahedi, J., & Rounaghi, M. M. (2015). Application of artificial neural network models and principal component analysis method in predicting stock prices on Tehran Stock Exchange. *Physica A: STATIStical Mechanics and Its Applications, 438,* 178–187. https://doi.org/10.1016/j.physa.2015.06.033

Zander, E., & Matthies, H. G. (2007). Tensor product methods for stochastic problems. *Proceedings in Applied Mathematics and Mechanics,* *7*(1), 2040067–2040068. https://doi.org/10.1002/pamm.200700773

Zhan, L., Liu, Y., Wang, Y., Zhou, J., Jahanshad, N., Ye, J., & Thompson, P. M. (2015). Boosting brain connectome classification accuracy in Alzheimer's disease using higher-order singular value decomposition. *Frontiers in Neuroscience, 9*, 257. https://doi.org/10.3389/fnins.2015.00257

Zhang, C. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of STATIStics, 38*(2), 894–942. https://doi.org/10.1214/09-AOS729

Zhang, S., Liu, C. C., Li, W., Shen, H., Laird, P. W., & Zhou, X. J. (2012). Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Research, 40*(19), 9379–9391. https://doi.org/10.1093/nar/gks725

Zhang, Y., Li, R., & Tsai, C. L. (2010). Regularization parameter selections via generalized information criterion. *Journal of the American STATIStical Association, 105*(489), 312–323. https://doi.org/10.1198/jasa.2009.tm08013

Zhang, Z., Member, S., Xu, Y., & Member, S. (2015). A survey of sparse representation : algorithms and applications. *IEEE access, 3*, 490-530. https://doi.org/10.1109/ACCESS.2015.2430359

Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American STATIStical Association, 101*(476), 1418–1429. https://doi.org/10.1198/016214506000000735

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal STATIStical Society. Series B: STATIStical Methodology, 67*(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse Principal Component Analysis. *Journal of Computational and Graphical STATIStics, 15*(2), 265–286. https://doi.org/10.1198/106186006X113430

Zou, H., & Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. The *Annals of STATIStics, 37*(4), 1733–1751. https://doi.org/10.1214/08-AOS625

Zubair, S., & Wang, W. (2013). Tensor dictionary learning with sparse tucker decomposition. *2013 18th International Conference on Digital Signal Processing*, DSP 2013. https://doi.org/10.1109/ICDSP.2013.6622725

ANEXOS

ANEXO 1- Cuestionarios y material suplementario

Cuestionario de Medición de la Actitud hacia la Didáctica

Act1	Creo que la asignatura de Didáctica se me va a dar bastante mal
Act2	Creo que la Didáctica será útil para mi futuro profesional
Act3	Es mejor dejar la Didáctica para los expertos y no incluirla en mi plan de estudios
Act4	Saber los contenidos de la asignatura de Didáctica incrementará mis posibilidades de trabajo
Act5	Un buen educador ha de haber estudiado Didáctica
Act6	La formación Didáctica me ayudará a entender mejor que es lo que se hace ante los procesos de enseñanza-aprendizaje
Act7	Me siento tranquilo ante el aprendizaje de los contenidos teóricos y prácticos de la Didáctica
Act8	Para el desarrollo profesional de mi carrera considero que hay otras asignaturas más importantes que la Didáctica
Act9	Trabajar con la Didáctica hace que me sienta muy nervioso
Act10	La Didáctica puede ser útil para los expertos en Pedagogía; pero no para otros profesionales
Act11	La formación en Didáctica mejora la experiencia profesional
Act12	Cuando me enfrento a un texto de Didáctica me siento incapaz de pensar con claridad
Act13	Estoy entusiasmado ante la posibilidad de utilizar la Didáctica en mi trabajo
Act14	Si tuviera la posibilidad me inscribiría en más cursos de Didáctica
Act15	Estudiar Didáctica es una diversión para mi
Act16	Estudiar Didáctica es una pérdida de tiempo
Act17	Me gustaría continuar mi formación Didáctica siguiendo cursos avanzados de esta materia
Act18	La mayor parte de los alumnos se benefician siguiendo un curso de Didáctica
Act19	La materia que se imparte en un curso de Didáctica es muy poco interesante
Act20	La Didáctica es un aspecto inseparable de los procesos educativos
Act21	Pensar que tengo que hacer un curso de Didáctica me pone nervioso
Act22	La Didáctica es una de las asignaturas que más temo
Act23	Tengo confianza en mí mismo-a cuando me enfrento al estudio de la Didáctica
Act24	La Didáctica es agradable y estimulante para mí
Act25	La Didáctica me parece lo suficientemente abstracta como para ser útil en mi profesión
Act26	La formación en Didáctica es importante para mi desarrollo en mi campo de estudios
Act27	Creo que sería importante que se impartiera pronto la Didáctica en la preparación para los profesionales de la Educación

Estructura factorial propuesta por Mondéjar et al. (2008)

Cuestionario R-SPQ-2F

Apr1	Me doy cuenta de que estudiar me proporciona un sentimiento de profunda satisfacción personal
Apr2	Al elaborar o estudiar un tema, no me encuentro satisfecho hasta que me he formado mis propias conclusiones sobre él
Apr3	Mi objetivo es aprobar el curso haciendo el mínimo trabajo posible
Apr4	Sólo estudio seriamente lo que se da en las clases o lo que está en los programas detallados de las asignaturas
Apr5	Me parece que cualquier tema puede llegar a ser altamente interesante una vez que te metes en él
Apr6	Encuentro interesantes la mayoría de los nuevos temas y empleo tiempo extra intentando obtener mayor información sobre ellos
Apr7	Dado que no encuentro el curso muy interesante voy en mi trabajo a lo mínimo
Apr8	Aprendo las cosas repitiéndolas hasta que me las sé de memoria incluso aunque no las comprenda
Apr9	Estudiar temas académicos puede ser a veces tan apasionante como leer una buena novela o ver una buena película
Apr10	Me hago preguntas a mí mismo sobre los temas importantes hasta que los comprendo totalmente
Apr11	Creo que puedo aprobar la mayoría de las evaluaciones memorizando los aspectos clave en lugar de intentar comprenderlos
Apr12	Generalmente limito mi estudio a lo que está específicamente ordenado, porque creo que es innecesario hacer cosas extra
Apr13	Trabajo duro en mis estudios porque encuentro los temas interesantes
Apr14	Empleo bastante de mi tiempo libre en buscar más información sobre temas interesantes que se han discutido en las diferentes clases
Apr15	Me parece que no ayuda estudiar los temas en profundidad. Confunde y hace perder el tiempo cuando todo lo que se necesita es un conocimiento por encima de los temas
Apr16	Creo que los profesores no deberían esperar que los alumnos dedicaran mucho tiempo a estudiar cosas que no van a caer en el examen
Apr17	Voy a la mayoría de las clases con preguntas a las que desearía encontrar respuesta
Apr18	Es muy importante para mí echar un vistazo a la mayoría de las lecturas recomendadas que tienen que ver con las clases
Apr19	No le encuentro sentido a aprender contenidos que probablemente no caerán en el examen
Apr20	Me parece que la mejor manera de pasar los exámenes es recordar las respuestas de las posibles preguntas

Estructura bifactorial propuesta por Biggs et al. (2001) y Hernández-Pina et al. (2005)

Maslach Burnout Inventory – Human Services Survey (MBI-HSS)

MBI-1	Me siento emocionalmente agotado/a por mi trabajo
MBI-2	Me siento cansado/a al final de la jornada de trabajo
MBI-3	Me siento fatigado/a cuando me levanto por la mañana y tengo que enfrentarme con otro día de trabajo
MBI-4	Fácilmente comprendo cómo se sienten los clientes
MBI-5	Creo que trato a los clientes como si fuesen objetos impersonales
MBI-6	Trabajar todo el día con mucha gente es un esfuerzo
MBI-7	Trato eficazmente los problemas de los clientes
MBI-8	Siento que mi trabajo me está desgastando
MBI-9	Creo que estoy influyendo positivamente, con mi trabajo, en la vida de los demás
MBI-10	Me he vuelto más insensible con la gente desde que ejerzo esta profesión
MBI-11	Me preocupa el hecho de que este trabajo me esté endureciendo emocionalmente
MBI-12	Me siento muy activo/a
MBI-13	Me siento frustrado/a con mi trabajo
MBI-14	Creo que estoy trabajando demasiado
MBI-15	No me preocupa lo que les ocurra a los clientes
MBI-16	Trabajar directamente con personas me produce estrés
MBI-17	Fácilmente puedo crear una atmósfera relajada con los pacientes
MBI-18	Me siento estimulado/a después de trabajar en contacto con los clientes
MBI-19	He conseguido muchas cosas útiles en mi profesión
MBI-20	Me siento como si estuviera al límite de mis posibilidades
MBI-21	En mi trabajo trato los problemas que se me presentan con mucha calma
MBI-22	Creo que los clientes me culpan de algunos de sus problemas

Tabla S1. Contribución relativa del factor al elemento

	Pre		Post	
	Eje 1	Eje 2	Eje 1	Eje 2
Interés				
I13	836,95	163,05	994,38	5,62
I14	996,64	3,36	999,53	0,47
I15	999,2	0,8	977,54	22,46
I17	934,97	65,03	988,81	11,19
I18	999,77	0,23	996,65	3,35
I24	996,6	3,4	999,03	0,97
Ansiedad				
A1	900,2	99,8	994,71	5,29
A7	874,31	125,69	986,65	13,35
A9	509,2	490,8	918,16	81,84
A12	807,57	192,43	998,47	1,53
A21	723,71	276,29	983,28	16,72
A22	932,22	67,78	997,95	2,05
A23	748,4	251,6	993,34	6,66
Utilidad presente				
UPre3	878,92	121,08	840,16	159,84
UPre10	950	50	816,32	183,68
UPre16	725,82	274,18	924,57	75,43
UPre25	269,22	730,78	761,57	238,43
Utilidad profesional				
UPro2	918,7	81,3	922,21	77,79
UPro4	208,24	791,76	664,16	335,84
UPro5	619,03	380,97	548,05	451,95
UPro6	689,49	310,51	781,52	218,48
UPro11	824,59	175,41	954,9	45,1
UPro19	812,9	187,1	890,49	109,51
UPro20	512,32	487,68	916,72	83,28
UPro26	793,46	206,54	846,21	153,79
UPro27	832,09	167,91	887,74	112,26
Act8	917,27	82,73	927,79	72,21
Enfoque profundo				
P1	992,44	7,56	446,7	553,3
P2	499,83	500,17	403,28	596,72
P5	296,08	703,92	149,31	850,69
P6	278,13	721,87	236,93	763,07
P9	745,88	254,12	248,31	751,69
P10	404	596	6,15	993,85
P13	483,65	516,35	63,86	936,14
P14	521,5	478,5	74,22	925,78

	Pre		Post	
	Eje 1	Eje 2	Eje 1	Eje 2
P17	92,64	907,36	117,56	882,44
P18	399,64	600,36	12,98	987,02
Enfoque superficial				
S3	705,37	294,63	56,61	943,39
S4	672,32	327,68	89,94	910,06
S7	847,31	152,69	394,96	605,04
S8	543,64	456,36	159,4	840,6
S11	442,26	557,74	142,16	857,84
S12	606,25	393,75	184,47	815,53
S15	991,82	8,18	176,05	823,95
S16	318,22	681,78	243,22	756,78
S19	520,51	479,49	90,51	909,49
S20	897,46	102,54	8,62	991,38

Tabla S2. Cargas factoriales del SPCA y del PCA con rotación Varimax del cuestionario Actitud hacia la Estadística

	Sparse PCA				PCA con Varimax			
Ítems	**SPC1**	**SPC2**	**SPC3**	**SPC4**	**PC1**	**PC2**	**PC3**	**PC4**
ACT2	-0,447				-0,281	-0,051	-0,007	-0,037
ACT4	-0,296				-0,296	0,086	-0,111	0,017
ACT5	-0,487				-0,322	0,038	-0,009	-0,078
ACT6	-0,436				-0,395	0,003	0,015	0,064
ACT11	-0,217				-0,209	0,076	-0,006	-0,222
ACT19				0,046	0,078	0,109	0,084	0,127
ACT20	-0,039				-0,222	0,107	-0,024	-0,115
ACT26	-0,426				-0,254	-0,013	-0,067	0,009
ACT27	-0,199				-0,449	-0,009	0,068	0,043
ACT1		0,373			-0,004	0,299	0,093	-0,068
ACT7		-0,021	0,116		-0,013	-0,163	-0,030	-0,224
ACT9		0,446			-0,017	0,438	-0,120	0,078
ACT12		0,058			-0,149	0,203	0,098	0,098
ACT21		0,587			0,067	0,409	-0,034	-0,010
ACT22		0,531			0,046	0,528	-0,090	-0,108
ACT23		-0,141			0,040	-0,285	-0,049	-0,045
ACT3	0,127				0,251	0,125	-0,043	0,048
ACT10			-0,761		0,021	-0,018	-0,231	0,632
ACT16			-0,580		0,101	0,025	-0,066	0,350
ACT25			0,265		0,143	0,114	-0,104	-0,456
ACT13				-0,293	-0,148	-0,036	-0,312	0,055
ACT14				-0,425	-0,109	0,023	-0,475	0,063
ACT15				-0,572	0,176	-0,075	-0,444	-0,232
ACT17				-0,582	-0,007	0,033	-0,462	-0,052
ACT24		-0,107		-0,258	0,014	-0,201	-0,249	-0,094
ACT18					-0,133	-0,057	-0,075	-0,112
ACT8					-0,010	-0,029	0,232	-0,074

CRBPI-30 (Niños)

MY MOTHER IS A PERSON WHO ...		
1. ... makes me feel better after talking over my worries with her.		
Not like	Somewhat like	A lot like
2. ... tells me of all the things she has done for me.		
Not like	Somewhat like	A lot like
3. ... believes in having a lot of rules and sticking with them.		
Not like	Somewhat like	A lot like
4. ... smiles at me often.		
Not like	Somewhat like	A lot like
5. ... says, if I really cared for her, I would not do things that cause her to worry.		
Not like	Somewhat like	A lot like
6. ... insists that I must do exactly as I am told.		
Not like	Somewhat like	A lot like
7. ... is able to make me feel better when I am upset.		
Not like	Somewhat like	A lot like
8. ... is always telling me how I should behave.		
Not like	Somewhat like	A lot like
9. ... is very strict with me.		
Not like	Somewhat like	A lot like
10. ... enjoys doing things with me.		
Not like	Somewhat like	A lot like
11. ... would like to be able to tell me what to do all the time.		
Not like	Somewhat like	A lot like
12. ... gives hard punishment.		
Not like	Somewhat like	A lot like
13. ... cheers me up when I am sad.		
Not like	Somewhat like	A lot like
14. ... wants to control whatever I do.		
Not like	Somewhat like	A lot like
15. ... is easy with me.		
Not like	Somewhat like	A lot like
16. ... gives me a lot of care and attention.		
Not like	Somewhat like	A lot like
17. ... is always trying to change me.		
Not like	Somewhat like	A lot like
18. ... let’s me off easy when I do something wrong.		
Not like	Somewhat like	A lot like
19. ... makes me feel like the most important person in her life.		

MY MOTHER IS A PERSON WHO ...		
Not like	Somewhat like	A lot like
20. ... only keeps rules when it suits her.		
Not like	Somewhat like	A lot like
21. ... gives me as much freedom as I want.		
Not like	Somewhat like	A lot like
22. ... believes in showing her love for me.		
Not like	Somewhat like	A lot like
23. ... is less friendly with me, if I do not see things her way.		
Not like	Somewhat like	A lot like
24. ... let's me go anyplace I please without asking.		
Not like	Somewhat like	A lot like
25. ... often praises me.		
Not like	Somewhat like	A lot like
26. ... will avoid looking at me when I have disappointed her.		
Not like	Somewhat like	A lot like
27. ... let's me go out any evening I want.		
Not like	Somewhat like	A lot like
28. ... is easy to talk to.		
Not like	Somewhat like	A lot like
29. ... if I have hurt her feelings, stops talking to me until I please her again.		
Not like	Somewhat like	A lot like
30. ... let's me do anything I like to do.		
Not like	Somewhat like	A lot like

MY FATHER IS A PERSON WHO ...		
1. ... makes me feel better after talking over my worries with him.		
Not like	Somewhat like	A lot like
2. ... tells me of all the things he has done for me.		
Not like	Somewhat like	A lot like
3. ... believes in having a lot of rules and sticking with them.		
Not like	Somewhat like	A lot like
4. ... smiles at me often.		
Not like	Somewhat like	A lot like
5. ... says, if I really cared for him, I would not do things that cause him to worry.		
Not like	Somewhat like	A lot like
6. ... insists that I must do exactly as I am told.		
Not like	Somewhat like	A lot like
7. ... is able to make me feel better when I am upset.		

MY FATHER IS A PERSON WHO ...

Not like Somewhat like A lot like

8. ... is always telling me how I should behave.

Not like Somewhat like A lot like

9. ... is very strict with me.

Not like Somewhat like A lot like

10. ... enjoys doing things with me.

Not like Somewhat like A lot like

11. ... would like to be able to tell me what to do all the time.

Not like Somewhat like A lot like

12. ... gives hard punishment.

Not like Somewhat like A lot like

13. ... cheers me up when I am sad.

Not like Somewhat like A lot like

14. ... wants to control whatever I do.

Not like Somewhat like A lot like

15. ... is easy with me.

Not like Somewhat like A lot like

16. ... gives me a lot of care and attention.

Not like Somewhat like A lot like

17. ... is always trying to change me.

Not like Somewhat like A lot like

18. ... let’s me off easy when I do something wrong.

Not like Somewhat like A lot like

19. ... makes me feel like the most important person in him life.

Not like Somewhat like A lot like

20. ... only keeps rules when it suits him.

Not like Somewhat like A lot like

21. ... gives me as much freedom as I want.

Not like Somewhat like A lot like

22. ... believes in showing his love for me.

Not like Somewhat like A lot like

23. ... is less friendly with me, if I do not see things his way.

Not like Somewhat like A lot like

24. ... let’s me go anyplace I please without asking.

Not like Somewhat like A lot like

25. ... often praises me.

Not like Somewhat like A lot like

26. ... will avoid looking at me when I have disappointed him.

Not like Somewhat like A lot like

MY FATHER IS A PERSON WHO ...		
27. ... let’s me go out any evening I want.		
Not like	Somewhat like	A lot like
28. ... is easy to talk to.		
Not like	Somewhat like	A lot like
29. ... if I have hurt his feelings, stops talking to me until I please him again.		
Not like	Somewhat like	A lot like
30. ... let’s me do anything I like to do.		
Not like	Somewhat like	A lot like

Tabla S3. Cargas factoriales de la matriz de componentes para el modo 1, reteniendo 4 componentes sparse mediante C_{enet}Tucker3

Girls	**C_{enet}C1**	**C_{enet}C2**	**C_{enet}C3**	**C_{enet}C4**
G1	-0,06	-0,13	0,05	
G2	-0,20	0,13	0,07	-0,03
G3	-0,01	-0,12	-0,11	0,07
G4	-0,03	0,07	-0,03	0,08
G5	-0,02	0,06	0,06	-0,05
G6	-0,09	0,03	-0,11	-0,05
G7	-0,02	0,01	-0,11	0,01
G8	-0,05	0,09	-0,08	-0,13
G9	0,21	0,12	0,03	-0,03
G10	0,12	0,09	-0,05	-0,07
G11	0,04	0,09	-0,14	0,06
G12	-0,11	-0,08	-0,01	-0,07
G13	0,06	0,21	0,17	0,08
G14	0,01	0,05	-0,07	0,08
G15		-0,11	-0,10	-0,03
G16	-0,07	-0,09	0,00	0,05
G17	0,05	0,09	0,04	-0,04
G18	-0,06	-0,07	0,13	0,08
G19	-0,04	-0,10	0,02	0,01
G20	0,01	-0,03	0,01	0,12
G21		-0,13	0,04	-0,06
G22	-0,11	-0,07	0,05	-0,09
G23	-0,12	0,06	0,07	
G24	-0,02	0,12	0,06	-0,02
G25	0,12	-0,04	0,05	0,08
G26	0,07	-0,03	-0,06	0,00
G27	0,10	0,02	0,09	0,08
G28	0,17	-0,03	0,01	-0,14
G29	-0,02	-0,22	0,15	-0,06
G30	-0,09	0,02	-0,03	-0,11
G31		-0,06	-0,02	-0,02
G32	0,01	0,02	-0,15	0,16
G33	0,02	-0,05	-0,03	-0,15
G34	0,08	-0,10	-0,08	0,05
G35	0,01	-0,02	-0,03	0,01
G36	-0,13	0,01	-0,04	-0,10
G37	-0,01			0,02
G38		-0,05	-0,02	0,21
G39	0,06	-0,02	-0,19	0,06
G40	0,03	-0,10	-0,03	-0,09
G41	0,15	-0,03	0,18	-0,01
G42	-0,14		0,05	-0,04
G43	0,11	-0,12		-0,03
G44	0,06	-0,02	0,03	-0,10

Girls	$C_{enet}C1$	$C_{enet}C2$	$C_{enet}C3$	$C_{enet}C4$
G45	-0,01	-0,08	0,08	0,03
G46	-0,02	-0,05		-0,03
G47	0,01	0,11	-0,11	0,06
G48	-0,04	-0,05	-0,10	0,07
G49	0,02	0,05	0,05	0,13
G50	0,04	0,03	-0,03	0,07
G51	0,05	-0,14	0,09	0,09
G52	0,11	-0,07	0,06	
G53	-0,04	0,07	-0,06	0,01
G54	-0,03	0,08	-0,04	-0,03
G55	0,03	-0,05	-0,15	-0,04
G56	-0,05	0,02	-0,04	0,05
G57	0,05	0,28	0,00	0,08
G58	-0,01	-0,10	-0,18	-0,19
G59	0,08	-0,02	-0,09	0,01
G60	-0,01	-0,18	0,16	0,02
G61	0,16	-0,05	0,05	0,19
G62	0,18	-0,11	0,01	-0,02
G63	-0,03	0,01	-0,03	0,08
G64	-0,13	0,03	0,01	-0,06
G65	-0,13	0,04	0,02	
G66	0,14	-0,04	0,05	-0,09
G67	-0,05	-0,07	-0,05	-0,01
G68	0,01	-0,03	0,04	-0,13
G69	0,10	-0,06	-0,08	-0,06
G70	0,05	0,07		0,07
G71	0,05	0,02	0,06	0,13
G72	-0,13	0,02	0,01	-0,01
G73	-0,06	-0,02	-0,06	0,17
G74		0,13	-0,03	0,02
G75	-0,04	-0,01	0,01	0,07
G76	-0,09	-0,09	-0,01	0,10
G77	-0,12	0,02	0,02	-0,03
G78	-0,03	0,14	0,00	-0,18
G79	0,01	-0,03	0,08	0,05
G80	0,02	0,06	-0,15	0,08
G81	0,04	-0,16	-0,05	-0,02
G82	0,08	-0,07	0,10	-0,02
G83	-0,07	-0,03	-0,02	0,13
G84	0,02	-0,03	0,11	0,06
G85	0,16	-0,06	0,13	0,04
G86	-0,07	0,08	-0,08	-0,07
G87	-0,03	0,05		0,08
G88	-0,06	-0,09	0,06	-0,02
G89	0,03	0,02	0,09	0,04
G90	0,08	0,12	-0,15	0,03
G91	-0,11	0,09	0,16	-0,08

Girls	$C_{enet}C1$	$C_{enet}C2$	$C_{enet}C3$	$C_{enet}C4$
G92	0,02		-0,03	-0,11
G93	-0,17	0,12	-0,01	-0,01
G94		-0,03	-0,04	
G95		-0,05	0,04	0,01
G96	-0,04		0,03	-0,08
G97	-0,04	0,11	0,04	-0,05
G98	0,01	0,01	0,03	-0,03
G99	-0,02	-0,15	-0,03	-0,02
G100	0,03	-0,06	0,11	-0,02
G101	-0,04		0,04	0,05
G102		0,07	0,01	0,16
G103	-0,07	-0,09	0,07	0,05
G104	0,03	0,11	0,03	
G105	-0,08		0,01	0,11
G106	-0,03	-0,04	-0,11	-0,03
G107	0,02	-0,09	-0,10	-0,03
G108	-0,08	-0,07	-0,05	0,01
G109	-0,03	0,04	0,12	-0,08
G110	-0,02	0,01	-0,13	0,10
G111	-0,04	-0,02	-0,03	0,05
G112	0,02	0,09	-0,11	-0,10
G113	0,14	0,06	-0,01	-0,19
G114	-0,09	-0,01	0,13	0,12
G115	0,19	0,04	-0,04	-0,04
G116	0,05	-0,10		-0,03
G117	-0,05	-0,04	-0,15	0,04
G118	0,00	0,05	0,08	-0,03
G119	-0,15	0,08	0,11	
G120	-0,04		-0,08	-0,05
G121	0,07	-0,08	-0,04	0,07
G122	-0,01	0,04	-0,12	0,02
G123	0,08	0,06	-0,04	
G124	-0,12	-0,03	-0,03	-0,08
G125	0,04	-0,08	0,12	0,01
G126	-0,02	0,10	0,06	-0,05
G127	0,07	-0,06	-0,01	-0,14
G128	0,06	0,14	-0,06	0,03
G129	0,07	-0,11	0,14	-0,08
G130	0,09	0,04	0,23	0,10
G131	0,01	-0,02	-0,14	0,08
G132	0,04	0,01	-0,01	0,03
G133	0,09	0,11	-0,01	0,02
G134	-0,01	0,02	0,04	-0,04
G135	-0,04	0,03	0,05	-0,12
G136	-0,10	-0,11	-0,03	-0,09
G137	-0,10	-0,06	0,08	-0,02
G138	-0,08	0,21	0,20	0,07

Girls	$C_{enet}C1$	$C_{enet}C2$	$C_{enet}C3$	$C_{enet}C4$
G139	-0,01	-0,07	-0,01	-0,02
G140	-0,05	0,02	0,02	0,05
G141	-0,04	-0,04	0,02	-0,04
G142	-0,10	0,02	-0,05	0,03
G143	-0,01		-0,06	0,23
G144	0,01	0,10	0,07	-0,11
G145	0,09	0,03	-0,03	-0,11
G146	0,02	-0,04	-0,03	0,11
G147	0,22	0,09	0,01	
G148	-0,19		0,03	0,04
G149	-0,04	0,04	-0,07	-0,23
G150	0,05	0,16	0,07	-0,11
G151	0,18	0,09	0,01	-0,10
G152	-0,11	0,01		
G153	0,05	0,01	-0,18	0,01

ANEXO 2 - Librerías de R y Bioconductor empleadas en esta investigación y funciones de propia elaboración

Se recogen a continuación algunos de los paquetes de R empleados en el análisis de los datos de esta memoria de Tesis Doctoral.

Repositorio CRAN

BiplotbootGUI (2019): Análisis Biplot.
Elasticnet: Sparse PCA
ggplot2: Funciones gráficas.
gplots: Funciones gráficas.
glmnet: Incorporación de métodos de regularización Lasso y Elastic net en regresión lineal, regresión logística y multinomial, regresión de Poisson y modelo de Cox.
MultBiplotR (2019): Análisis Biplot.
NMF: Factorización no negativa
rCUR: CUR decomposition package.
r.jive: Descomposición JIVE de un conjunto de matrices que comparten una dimensión.
rrcov3way: Métodos de análisis de datos multivía (PARAFAC y Tucker) y su extensión a versiones robustas y composicionales.
sparseLDA: Análisis lineal discriminante sparse.
ThreeWay: Candecomp/Parafac y modelos Tucker para el análisis de array de tres vías.

Repositorio Bioconductor

biobase: Funciones base de Bioconductor.

affy: Métodos para arrays de Affymetrix.

limma: Modelos lineales para datos de microarrays.

annotate: Anotación de microarrays.

simpleaffy: Very simple high-level analysis of Affymetrix data.

sva: Eliminación de efecto batch y otro tipo de variabilidad en experimentos de alto rendimiento.

Funciones de elaboración propia para la implementación de las metodologías propuestas, próximamente publicadas en el paquete de R *SparseCenetSVD*

Proyección de vectores sobre un espacio restringido	
lasso.proj()	Proyección de un vector sobre la bola $\mathfrak{B}^{\ell_1}$ (E Berg et al., 2008)
enet.proj()	Proyección de un vector sobre la bola $\mathfrak{B}^{\ell_1+\ell_2}$. Pseudocódigo propuesto por (Mairal et al., 2010)
	Proyección de un vector sobre la bola $\mathfrak{B}^{(\ell_1+\ell_2)\cap\ell_2}$.
Descomposición en valores singulares restringida: extensión a Elastic net	
csvd.enet (X, Q=2, tau.u = rep(1.4, Q), tau.v = rep(1.4, Q),alpha.u=1e-16, alpha.v=1e-16, itermax.pi=1000, itermax.pocs=1000, eps.pi=1e-16, eps.pocs=1e-16, init.svd="svd")	C_{enet}SVD
Análisis de componentes principales sparse restringido: extensión a Elastic net	
pca.enet(X, Q=2, tau.u = 1.4, tau.v = 1.4,alpha.u=1e-16, alpha.v=1e-16, itermax.pi=1000, itermax.pocs=1000, eps.pi=1e-16, eps.pocs=1e-16, init.svd="svd", init.transf=1, obs.names=FALSE, plot.axis=c(1,2))	C_{enet}PCA
Métodos Biplot clásicos (JK, GH, HJ) sparse restringidos: extensión a Elastic net	
Biplot.enet (X, Q=2, tau.u = 1.4, tau.v = 1.4, alpha.u=1e-16, alpha.v=1e-16, Biplot.type=2, itermax.pi=1000, itermax.pocs=1000, eps.pi=1e-16, eps.pocs=1e-16, init.svd="svd", init.transf=1, plot.axis=c(1,2), names.obs=FALSE, select.cur=FALSE, variables.cur=1, weighted.cur=FALSE, method.cur="top.scores")	C_{enet}Biplot
Descomposición tensorial Tucker sparse restringida: extensión a Elastic net	
tucker.enet(X,tucker.type=3,I,J,K,p=2,q=2,r=2, tau.uA = 1.4, tau.uB = 1.4,tau.uC=1.4, alpha.uA=1e-16, alpha.uB=1e-16,alpha.uC=1e-16, tau.vA = sqrt(I), tau.vB = sqrt(J),tau.vC=sqrt(K), alpha.vA=1e-16, alpha.vB=1e-16,alpha.vC=1e-16, itermax.pi=100,itermax.pocs=100,itermax.tucker=100, eps.pi=1e-16, eps.pocs=1e-16, eps.tucker=1e-16,	C_{enet}Tucker

init.tucker="svd", init.svd="svd", center=F, scale=F, center.mode="A", scale.mode="B")	
factorplots.enet(A, B, C, Qa.plot=c(1,2), Qb.plot=c(1,2), Qc.plot=c(1,2))	Gráficos factoriales
intBiplot.enet(A,B,C,G,Q.plot=c(1,2))	Biplot interactivo
jointBiplot.enet(A,B,C,G, Q.plot=c(1,2), mode.fix="C")	Biplots conjuntos
Funciones auxiliares	
init.transformation(X,t)	Preprocesamiento de matrices de dos vías (t=1 datos brutos; t=2 eliminar la media global; t=3 centrado por columnas; t=4 estandarización por columnas; t=5 centrado por filas; t=6 estandarización por filas; t=7 doble centrado)
cv.alpha_enet<-function(data, Q, nfolds=10, set.alpha.v=c(),parallel=FALSE, type.measure="MSE", tau.u=sqrt(dim(data)[2]), tau.v=1.4,alpha.u=0.000001,ntau=5, itermax.pi=500, itermax.pocs=500,eps.pi=1e-16, eps.pocs=1e-16, init.svd="svd")	CV para la selección de α
bic.tau_enet<-function(data, Q, ntau=3, tau.u=sqrt(dim(data)[1]), alpha.u=0.000001, alpha.v=0.5, itermax.pi=500, itermax.pocs=500, eps.pi=1e-16, eps.pocs=1e-16, init.svd="svd")	BIC para la selección del parémtro de regularización

Aprovecho la ocasión para agradecer al grupo de trabajo (Guillemot et al., 2019) el uso de sus códigos de la metodología CSVD en R, públicos en el repositorio github (https://github.com/vguillemot/csvd), los cuales se han utilizado para su adaptación a C_{enet}SVD bajo el siguiente copyright: