UNIVERSIDAD DE GRANADA

Máster Universitario en Estadística Aplicada

TRABAJO FIN DE MÁSTER

Análisis de datos de proximidad para exploración y clasificación de textos

J. David Fernández Romero

Granada, Septiembre de 2021

Índice general

Resumen

El objetivo de este trabajo es mostrar la aplicación del análisis de datos de proximidad al análisis estadístico de textos.

El trabajo se encuentra dividido en 4 capítulos. En el primer capítulo se definen los conceptos mas importantes en el campo de la estadística textual, se realiza una breve revisión de su desarrollo histórico, se muestran sus principales aplicaciones y se describe una etapa propia de la estadística de variables textuales: el necesario tratamiento previo de estos datos para obtener un conjunto estructurado sobre el que poder aplicar las correspondientes técnicas estadísticas, introduciendo los diversos tratamientos aplicables para codificar el corpus lingüístico en tablas numéricas.

En el capítulo 2 se describen algunas técnicas estadísticas exploratorias para datos multivariantes:

- El análisis de componentes principales, técnica de aplicación muy generalizada por su potencia para abordar la reducción de la dimensionalidad y permitir la detección de variables latentes.
- El escalado multidimensional, técnica de aplicación específica sobre datos de proximidad que permite abordar el análisis desde el punto de vista de la similaridad.
- El análisis de correspondencias, técnica históricamente ligada al análisis estadístico de textos.

En la sección dedicada al escalado multidimensional se incluye la descripción del concepto de medida de proximidad para datos multivariantes, se definen los conceptos de distancia y similaridad, se introducen algunas de las métricas mas importantes y se describe la medida de similaridad mas utilizada para comparar textos: el coseno del ángulo entre dos vectores.

El tercer capítulo se dedica al problema de la clasificación mediante la metodología del análisis discriminante. Comienza con una breve descripción del problema de clasificación o discriminación y contiene una parte de desarrollo teórico centrada en el análisis discriminante clásico desarrollado por Fisher.

Por último, en el cuarto capítulo se muestra la aplicación de los conceptos y técnicas estadísticas anteriores sobre un corpus lingüístico. Para ello se realiza, en primer lugar, un análisis exploratorio de la variable textual objeto de estudio, analizando el vocabulario empleado en el corpus desde el punto de vista de la frecuencia relativa por documentos de las distintas palabras, de las palabras que mejor caracterizan los textos pertenecientes a cada uno de los autores y de la comparación en las frecuencias de uso de las palabras entre los mismos. Además, se aplica la técnica de *unfolding* para obtener una representación conjunta de los autores y las palabras mas representativas que permita interpretar su relación a través de las distancias en el espacio de representación conjunto. En una segunda etapa se aborda una metodología para resolver el problema de clasificación asociado a la identificación de los autores de cada uno de los textos mediante el análisis de las proximidades entre ellos. Esta tarea se realiza utilizando la distancia del coseno como medida de la similaridad entre dos textos y aplicando técnicas de escalado multidimensional con carácter previo a obtener un modelo de clasificación mediante análisis discriminante lineal.

Los resultados obtenidos muestran como la matriz de documentos-palabras asociada al corpus lingüístico no es apropiada para la aplicación del análisis discriminante, obteniendo un modelo que no mejora demasiado una clasificación por azar, mientras que la aplicación sobre esta matriz de la distancia del coseno entre documentos para posteriormente construir la matriz de disimilaridad asociada y aplicar MDS proporciona una configuración que, reduciendo considerablemente la dimensionalidad de la matriz, resulta, al aplicar análisis discriminante, en un buen modelo con un acierto en la identificación de autores cercano al 90 %.

Abstract

The aim of this paper is to show the application of proximity data analysis to the statistical analysis of texts.

The paper is divided into four chapters. In the first chapter, the most important concepts in the field of textual statistics are defined, a brief review of its historical development is made, its main applications are shown and a specific stage of the statistics of textual variables is described: the necessary prior treatment of these data to obtain a structured set on which the corresponding statistical techniques can be applied, introducing the different treatments applicable to codify the linguistic corpus in numerical tables.

Chapter 2 describes some exploratory statistical techniques for multivariate data:

- Principal component analysis, a technique of widespread application because of its power to address dimensionality reduction and to allow the detection of latent variables.
- Multidimensional scaling, a technique of specific application on proximity data that allows to approach the analysis from the point of view of similarity.
- Correspondence analysis, a technique historically linked to the statistical analysis of texts.

The section on multidimensional scaling includes a description of the concept of proximity measure for multivariate data, defines the concepts of distance and similarity, introduces some of the most important metrics and describes the most commonly used similarity measure for comparing texts: the cosine of the angle between two vectors.

The third chapter is devoted to the problem of classification using the methodology of discriminant analysis. It begins with a brief description of the problem of classification or discrimination and contains a part of theoretical development focused on the classical discriminant analysis developed by Fisher.

Finally, the fourth chapter shows the application of the previous concepts and statistical techniques on a real linguistic corpus. For this purpose, first of all, an exploratory analysis of the textual variable under study is carried out, analysing the vocabulary used in the corpus from the point of view of the relative frequency per document of the different words, of the words that best characterise the texts belonging to each of the authors and of the comparison in the frequencies of use of the words between them. In addition, the *unfolding* technique is applied to obtain a joint representation of the authors and the most representative words that allows us to interpret their relationship through the distances in the joint representation space. In a second stage, a methodology is addressed to solve the classification problem associated with the identification of the authors of each of the texts by analysing the proximities between them. This task is carried out using the cosine distance as a measure of the similarity between two texts and applying multidimensional scaling techniques prior to obtaining a classification model by means of linear discriminant analysis.

The results obtained show that the document-word matrix associated with the linguistic corpus is not appropriate for the application of discriminant analysis, obtaining a model that does not improve a classification by chance, while the application of the cosine distance between documents on this matrix to subsequently construct the associated dissimilarity matrix and apply MDS provides a configuration that, by considerably reducing the dimensionality of the matrix, results, when applying discriminant analysis, in a good model with a success rate in the identification of authors close to 90 %.

Capítulo 1

Análisis estadístico de textos.

Se suele definir la minería de datos como el proceso que, utilizando técnicas estadísticas y de las ciencias de la computación, pretende descubrir patrones no triviales de información desconocida en conjuntos de datos estructurados. En la actualidad este concepto se encuentra íntimamente ligado al de Big Data debido al desarrollo experimentado en las últimas décadas en la capacidad de almacenamiento y procesamiento de las computadoras que ha permitido la aplicación de determinadas técnicas estadísticas a volúmenes de datos que hasta hace poco resultaban imposibles de abordar.

Por su parte, se conoce como minería de textos al proceso de extracción, a partir de textos no estructurados, de patrones no triviales que proporcionan nueva información y conocimientos.

Podríamos entonces decir que la principal diferencia entre la minería de textos y la minería de datos es que la primera se aplica sobre información no estructurada mientras que la segunda lo hace sobre información estructurada. En realidad la minería de textos necesita una fase previa de procesamiento de los datos textuales (no estructurados) que los transforma en un conjunto de datos estructurados sobre los que se aplican las técnicas estadísticas que dan sustento a la minería de datos.

Los importantes avances en tecnología, tanto a nivel de hardware como de software, han propiciado un fuerte desarrollo de las técnicas de minería de datos, incidiendo especialmente en el caso de datos de origen textual.

Paralelamente al desarrollo tecnológico, la implantación de Internet ha resultado en la disponibilidad de una gran cantidad de contenido textual, de carácter muy diverso y fácilmente almacenable y procesable, lo que ha contribuido también al desarrollo de métodos y algoritmos para la detección de patrones no triviales de interés en los datos textuales.

Además, el hecho de que la forma mas común de almacenamiento de información sea en forma de texto ha provocado que la minería de textos se haya convertido en una de las áreas con mayor potencial de la minería de datos.

Sin embargo, la minería de textos es mucho más compleja que la minería de datos, fundamentalmente por el hecho de que esta última trata conjuntos de datos estructurados mientras que la primera lo hace con datos inicialmente desestructurados y en consecuencia mas confusos.

Una de las características más destacada de los datos textuales es su dispersión y alta dimensionalidad. Un corpus textual se puede representar como una matriz $n\ x\ d$, donde n es el número de documentos que componen el corpus y d es el número de términos distintos que aparecen en el conjunto de documentos. Así, el elemento (i, j) de la matriz representará la frecuencia normalizada del j-ésimo término en el i-ésimo documento. Esta representación numérica de los corpus textuales da lugar a matrices huecas y de muy elevada dimensionalidad, lo que condiciona en gran medida la aplicación de determinadas técnicas estadísticas.

1.1. Desarrollo histórico y aplicaciones destacadas.

Origen de la estadística textual.

El origen de la analítica de textos se remonta a la antigua Alejandría, cuando los gramáticos elaboraron el inventario de palabras de la Biblia. Este primitivo análisis textual se centraba fundamentalmente en el tipo y volumen del vocabulario empleado sin acometer el análisis del contenido de los documentos.

Durante las primeras décadas del siglo XX los lingüistas anglosajones abordaron el estudio de las concordancias de determinados vocablos en grandes autores literarios con la intención de inferir el contexto común alrededor de cada palabra.

A partir de la década de 1930, Zipf, Yule y Guiraud entre otros, realizaron importantes aportes teóricos al análisis estadístico de datos textuales, introduciendo leyes empíricas sobre la distribución de las palabras y resolviendo así algunos problemas planteados por los estilistas franceses.

Jean-Paul Benzécri, gran estadístico francés considerado el padre de la escuela francesa de análisis de datos, publicó en 1964 un curso de lingüística matemática que venía impartiendo en la Facultad de Ciencias de Rennes desde 1960. En él, Benzécri plantea un nuevo método de análisis descriptivo multivariante: la técnica del Análisis Factorial de Correspondencias. Al año siguiente y en esa misma facultad, Escofier defendía su tesis doctoral resaltando las principales propiedades de este método.

Con el Análisis de Correspondencias, Benzécri aporta un método estadístico para solucionar los problemas fundamentales de los lingüistas. El Análisis de Correspondencias puede ser aplicado en muchos campos, pero debemos tener en cuenta que, en comparación, el tratamiento de datos lingüísticos presenta particularidades propias debidas a la multidimensionalidad intrínseca de esta materia.

Las primeras aproximaciones a la minería de textos en el sentido ya de analizar la información contenida en los documentos tiene su origen en las tareas de catalogación de documentos que se extendieron rápidamente a la extracción de información gracias al desarrollo de técnicas de procesamiento del lenguaje natural, en el que tuvo y sigue teniendo gran importancia la necesidad de extraer información de los documentos de texto de manera automatizada. De esta forma, el análisis estadístico de textos y el procesamiento del lenguaje natural se han convertido en áreas complementarias y muy interrelacionadas, cuyas investigaciones se retroalimentan permanentemente.

En la segunda mitad del siglo XX el desarrollo de programas informáticos corre paralelo al de los métodos y aplicaciones a grandes conjuntos de datos. L. Lebart y A. Morineau desarrollan, en 1984, un módulo de tratamiento de textos en el sistema SPAD. Posteriormente, en 1988, M. Bécue Bertaut presenta su tesis doctoral titulada “Un sistema informático para el Análisis de Datos Textuales”, en la que desarrolla el programa SPAD.T. A partir de entonces se realizan grandes progresos en el análisis de respuestas libres a cuestiones abiertas y su relación con el resto de variables informadas en las encuestas.

En las últimas décadas del siglo XX la minería de textos buscaba automatizar, en cierta medida, el acceso a información concreta entre una gran cantidad de documentos de texto, por lo que las técnicas desarrolladas buscaban optimizar métodos para resumir los documentos manteniendo la información mas relevante de manera que se consiguiera disminuir el tamaño de los textos a tratar facilitando con ello las tareas de búsqueda, catalogación y clasificación.

En la actualidad, debido al desarrollo de la informática y las telecomunicaciones (especialmente Internet, la inteligencia artificial y el big data), son muchas las posibilidades de profundizar en el análisis de textos desde el punto de vista de su contenido y no únicamente del vocabulario. En este punto, es importante resaltar la diferencia entre el análisis textual como concepto general y el análisis estadístico de textos o simplemente análisis de datos textuales que consiste en la aplicación a la lingüística de técnicas propias de la estadística.

Principales aplicaciones.

Son muchas las aplicaciones de la estadística textual. Uno de sus objetivos principales es el descubrimiento y análisis de patrones de interés.

Entre las aplicaciones mas habituales podemos citar el análisis de sentimientos, el análisis de las respuestas a preguntas abiertas en encuestas o el filtrado de correos electrónicos.

El objetivo del análisis de sentimientos es, a partir de una opinión expresada en forma textual, deducir la actitud del individuo hacia el objeto de dicha opinión. Es una técnica de gran utilidad para personas o empresas cuyos productos o actividades tienen una dependencia importante de su proyección social por lo que ha experimentado un fuerte desarrollo e incrementado en gran medida su presencia en paralelo al auge de las redes sociales siendo uno de los campos de la minería de textos mas relacionados y que mas se beneficia de las técnicas de procesamiento del lenguaje natural.

La utilización de preguntas de respuesta libre es muy frecuente destacando su empleo habitual en encuestas de satisfacción de clientes, encuestas de opinión y estudios de preferencia en artículos de consumo. Estas preguntas permiten captar información que no es posible obtener mediante preguntas cerradas y cuentan con algunas ventajas sobre estas últimas como el hecho de un menor condicionamiento del encuestado y mayor fiabilidad acerca de sus opiniones reales, así como la eliminación de las limitaciones que conlleva

el diseño de preguntas cerradas en las que demasiadas veces se incluye una categoría "otros" que no aporta información válida sobre la verdadera respuesta del encuestado mas allá de no coincidir con ninguna de las opciones tenidas en cuenta en el diseño de las preguntas. Las técnicas de minería de textos permiten extraer la información contenida en las respuestas abiertas operando sobre las respuestas originales en los términos en que han sido expresadas por el encuestado al tiempo que permiten aumentar la calidad de la interpretación de las respuestas a preguntas cerradas en la misma encuesta.

La potencia de las técnicas estadísticas aplicadas al análisis de textos ha posibilitado, por ejemplo, el desarrollo de herramientas informáticas que permiten automatizar la organización del correo electrónico en carpetas y específicamente la clasificación o no como spam de los correos recibidos por los usuarios. Para ello, se han construido clasificadores que categorizan en spam o no spam el correo entrante mediante la extracción de cada correo entrante de los términos más relevantes (en el sentido de contar con una menor probabilidad de constituir spam).

1.2. Construcción de la matriz de *documentos x palabras.*

La variable textual, expresada en tablas de recuentos, implica una mayor complejidad en su forma que las variables cuantitativas o cualitativas "puras". Sin embargo, esta complejidad junto a la dificultad de tratamiento de este tipo de variables aporta un mayor apego a la realidad de los resultados obtenidos.

1.2.1. Preprocesamiento del corpus y segmentación en unidades léxicas.

Llamamos *corpus* al conjunto de textos a analizar. Puede tratarse de artículos de opinión, relatos de diversos autores o épocas, comentarios en una red social, respuestas libres a una pregunta abierta en una encuesta, etc.

Es fundamental realizar un cuidadoso procesamiento previo del corpus inicial para una correcta identificación de las unidades léxicas cuyas ocurrencias se van a contar. Este preprocesamiento deberá estar formado por una serie de reglas bien definidas que aporten

estabilidad, facilidad de comprensión y reproducibilidad. Es habitual seguir la norma lexicométrica desarrollada por Muller en 1977 y posteriormente completada por Labbé en 1990, que contempla:

- Utilizar un corrector ortográfico automático potente que tenga en cuenta las reglas gramaticales.
- Realizar una limpieza de notaciones normativas (eliminando por ejemplo, las mayúsculas al inicio de frases o las abreviaciones ambiguas). Se trata, en definitiva, de dotar de un estatus único a cada carácter del texto (por ejemplo, el punto que indica el fin de una frase es diferente del punto que separa ciertas abreviaciones como D.N.I. o N.I.E.).
- Definir los signos considerados de puntuación, de manera que todos los demás signos son tratados como parte del conjunto de las letras.
- Si se considera necesario, *lematizar* el corpus, es decir, transformar cada palabra en la *entrada del diccionario* a la que se asocia.
- Definir lo que se conoce como *stoplist* (o conjunto de *stopwords*), que no es mas que la lista de palabras que se eliminan del estudio por considerar que no aportan información (resulta habitual eliminar las preposiciones, artículos y conjunciones).
- En estudios *comparativos*, tales como en el análisis de emociones o el análisis de encuestas con preguntas abiertas, es común establecer un *umbral de frecuencia*, ya que se considera que la comparación sólo tendrá sentido entre palabras que se utilicen con al menos una determinada frecuencia.

Lematización.

La principal dificultad al acometer el preproceso de todo corpus es la definición de la unidad léxica, o *unidad de segmentación* del corpus, puesto que será la base del análisis estadístico que vamos a realizar.

Podríamos adoptar la *palabra*, en su concepción lexicográfica como secuencia de letras delimitada a izquierda y derecha por espacio en blanco o signo de puntuación, (también denominada forma gráfica), pero así no obtendríamos una unidad léxica claramente determinada, por lo que se suele optar por el *lema*[1].

[1] *lema*: entrada del diccionario asociada a una palabra.

En el caso del idioma *castellano* la lematización de un texto requiere convertir:

- las flexiones verbales al infinitivo
- los sustantivos a singular
- los adjetivos al masculino singular

Aun cuando se utilice un analizador morfo-sintáctico de alta calidad para automatizar el proceso de lematización, pueden subsistir ambigüedades únicamente solventables mediante una operación manual.

La lematización del corpus previa a su análisis estadístico proporciona las siguientes ventajas:

- Se reduce la variabilidad entre respuestas: las frases "*me gusta ver series*" y "*por la noche veo una serie*" tienen en común los lemas "*ver*" y "*serie*", pero en su forma original no comparten ninguna forma gráfica
- Se limita la pérdida de unidades textuales, puesto que la frecuencia de cada lema considerado será la suma de las frecuencias de las formas gráficas reagrupadas en el lema: en un texto en el que aparezca 18 veces la forma gráfica *permanencia*, 1 vez *permanecen*, 3 veces *permaneciendo* y 5 veces *permaneceremos*, el lema *permanecer* tendrá una frecuencia de 27 ocurrencias, cuando ninguna de las formas gráficas alcanza este umbral
- Se asocia a cada lema (o forma gráfica) su categoría sintáctica, aspecto fundamental en diversos momentos del tratamiento, puesto que hace posible seleccionar palabras por su categoría o utilizar la categoría para *apoyar* la interpretación de los resultados

Es aconsejable tomar en consideración ambos tipos de unidad léxica (palabra o forma gráfica, y lema) repitiendo incluso las distintas fases del tratamiento en cada caso puesto que los resultados obtenidos se enriquecerán mutuamente.

Stoplist.

En determinados casos se separan las unidades léxicas consideradas en dos grupos: las consideradas *llenas* (que aportan significado por sí solas, como los sustantivos y los verbos, y a veces pero no siempre, los adjetivos y los adverbios) y las consideradas *herramientas* (o gramaticales, no consideradas informativas, tales como los artículos, las preposiciones y

las conjunciones). En algunas aplicaciones las unidades léxicas gramaticales se consideran no útiles.

El conjunto de palabras consideradas no útiles se denomina *stoplist*, pero debe tenerse en cuenta que en determinados casos, como el caso de las respuestas libres de encuestas, los adverbios, las negaciones o los adjetivos, no deben eliminarse del tratamiento puesto que pueden ser extremadamente importantes.

***Stematización* y/o reagrupación de sinónimos.**

Puede resultar de utilidad realizar una *stematización*[2], consistente en el reagrupamiento de varios lemas provenientes de una misma raíz.

También, en determinados estudios, puede ser apropiada la unificación de sinónimos.

Debemos tener en cuenta que estos reagrupamientos deben realizarse con posterioridad a un primer tratamiento para garantizar que no se distorsionen los resultados.

Umbral de frecuencia.

El último concepto importante relacionado con el preproceso del corpus es el que hace referencia a la determinación de una frecuencia mínima de presencia en el corpus para que la unidad léxica sea considerada en el estudio. A veces se sustituye por el establecimiento *a priori* del número de palabras a incluir.

1.2.2. Codificación del texto: tabla léxica y tabla léxica agregada.

Una vez realizada la fase de preproceso mediante la segmentación del corpus en unidades léxicas y, en su caso, operadas la eliminación de las incluidas en la stoplist y las reagrupaciones necesarias, obtendremos los *glosarios*, tanto de unidades léxicas como de segmentos repetidos.

Las tablas generadas permiten estudiar la distribución de las palabras entre individuos[3] (*tabla léxica*: cruzando individuos y palabras) o entre categorías de individuos (*tabla léxica agregada*: cruzando categorías de individuos y palabras). En adelante, para una

[2] *stematización*: neologismo formado a partir de *stem*, que en inglés significa raíz.

[3] Entiéndase individuo en sentido amplio en el contexto adecuado, por ejemplo, en el estudio de encuestas con respuesta abierta será cada una de las respuestas individuales, en el caso de un corpus formado por varios documentos, cada uno de los documentos individuales, etc

mejor adecuación de la terminología utilizada al caso de aplicación, se utilizará el término *documento* como equivalente a *individuo*.

Tabla léxica *documentos x palabras.*

Para acometer el estudio de la distribución del vocabulario entre los documentos se codifica la variable textual en una tabla de frecuencias *Documentos* x *Palabras* que contiene el número de veces que en cada documento aparece cada una de las palabras.

Tabla léxica agregada *categorías x palabras.*

En determinadas ocasiones puede resultar de interés la comparación de la distribución de palabras entre categorías de documentos, en función de una variable categórica. En este caso, para cada categoría se cuenta el número de veces que cada palabra aparece en los documentos de la categoría, es decir, se agregan los documentos por categorías, obteniendo una tabla léxica agregada *Categorías* x *Palabras*.

Palabras características.

En los análisis estadísticos de *corpus* de texto que aborden un estudio *comparativo* entre distintos *corpus* o subconjuntos de un mismo *corpus*, o bien pretendan *clasificar* los individuos (o documentos) en varios grupos, resultará de interés obtener las palabras características de cada uno de los *corpus*, *subcorpus* o grupos, es decir, las palabras con frecuencia anormalmente alta en un determinado grupo respecto de su frecuencia global.

Uno de los métodos que se utiliza para identificar las palabras características de un determinado *subcorpus* es mediante la medida conocida como *tfIdf* que consiste en multiplicar la frecuencia de cada término por un factor corrector denominado *idf: inverse document frequency* que se obtiene mediante el logaritmo del cociente entre el número total de documentos y el número de documentos que contienen el término.

$$idf(pal) = log\left(\frac{n}{n_{pal}}\right)$$

Este indicador permite detectar el vocabulario que caracteriza un subconjunto en comparación con el conjunto total y, por lo tanto, no debe emplearse para comparar subconjuntos entre sí estableciendo parecidos entre los mismos. El valor de este indicador radica en su utilidad, junto con las adecuadas técnicas multivariantes, para la interpretación de los resultados.

1.3. Técnicas estadísticas multivariantes.

En la mayor parte de los casos el tratamiento previo del corpus textual da lugar a matrices de términos-documentos de una elevada dimensionalidad, lo que tradicionalmente ha constituido un serio inconveniente para el tratamiento de estos datos. Aunque la propia etapa de preprocesado incorpora técnicas para abordar este problema, ha sido principalmente el avance tecnológico de las últimas décadas lo que en mayor medida ha facilitado este tratamiento. En cualquier caso, el problema de la dimensionalidad sigue constituyendo un reto y su reducción es un objetivo principal en la mejora de los algoritmos existentes y en el desarrollo de nuevas metodologías.

Las técnicas de análisis estadístico multivariante son herramientas de uso común en muchas disciplinas: desde la psicología hasta la inteligencia artificial pasando por la sociología, la economía, la ingeniería, la medicina o las ciencias ambientales. En la actualidad, muchas de estas técnicas soportan los procesos de extracción de conocimiento que se encuentran detrás de los métodos conocidos como minería de datos y constituyen la base de las técnicas de inteligencia artificial.

El análisis de datos multivariantes comprende el estudio estadístico de varias variables medidas sobre elementos de una población y puede plantearse a dos niveles:

- Extraer la información contenida en los datos, para lo que se utilizan métodos exploratorios que extienden al caso multivariante las técnicas estadísticas descriptivas habituales para resumir los valores de las variables y describir su estructura de dependencia, así como realizar representaciones gráficas y elegir transformaciones de las variables que simplifiquen su descripción.
- Obtener conclusiones sobre la población que ha generado los datos, para lo que es preciso construir un modelo que explique dicha generación y permita realizar predicciones sobre datos futuros. Se utilizan para ello métodos inferenciales con los que se pretende generar conocimiento sobre el problema que subyace en los datos disponibles.

El objetivo último de las técnicas multivariantes exploratorias puede consistir en uno o varios de los siguientes:

- Resumir los datos en un conjunto de nuevas variables que resulten de aplicar determinadas transformaciones a las variables originales
- Identificar, si existen, grupos homogéneos de individuos o variables
- Clasificar nuevas observaciones en grupos predefinidos
- Relacionar conjuntos de variables entre sí

Transformar las variables originales posibilita simplificar la descripción de los datos reduciendo su dimensionalidad. En el caso mas extremo, reducir la dimensionalidad de un conjunto de datos multivariante a dos únicas variables indicadoras permite la representación bidimensional de los individuos y con ello la visualización e interpretación de las relaciones entre ellos.

Entre los métodos exploratorios multivariantes orientados a la obtención de nuevas variables indicadoras que sinteticen la información de las variables originales destaca, cuando los datos son continuos, el análisis de componentes principales. Esta técnica permite determinar las dimensiones necesarias para representar adecuadamente los datos. Cuando se cuenta con información sobre similitudes o semejanzas entre los individuos e interesa encontrar las dimensiones de dichas similitudes el concepto de componentes principales se generaliza en las técnicas de escalado multidimensional, mientras que en el caso de los datos textuales se ha utilizado tradicionalmente otra generalización de dicho concepto, en este caso dirigida al análisis de datos cualitativos, denominada análisis de correspondencias.

Por otra parte, el análisis de las similitudes entre los individuos permite encontrar grupos homogéneos cuando el sentido o significado de la similaridad es, a priori, desconocido, aunque en muchos casos las motivaciones o causas de esta similitud puedan permanecer ocultas a la interpretación del analista.

Para estudiar si los datos forman o no un grupo homogéneo y, si existen varios grupos, identificar los elementos que pertenecen a cada uno de ellos, se utilizan los métodos del análisis de conglomerados o análisis cluster, conocidos también como métodos de clasificación automática o no supervisada que permiten, a través del agrupamiento de los datos, abordar distintos objetivos:

- Cuando se conoce o sospecha que los datos son heterogéneos así como el número de grupos en que se podrían dividir: particionar los datos en un número de grupos preestablecido de forma que cada elemento pertenezca a uno y sólo uno de los grupos, todos los elementos pertenezcan a algún grupo y cada grupo sea internamente homogéneo.
- Cuando no existe información alguna acerca del grado de heterogeneidad de los datos: obtener una estructura de los elementos ordenada en niveles de forma jerárquica en función de su grado de similitud, de manera que los niveles superiores contienen a los inferiores, lo que permite al analista particionar los datos a posteriori en el número de grupos que considere mas apropiado.
- Cuando se abordan problemas con un elevado número de variables y resulta de interés realizar un estudio exploratorio previo para dividir las variables en grupos: clasificar las variables, lo que a su vez puede facilitar el planteamiento posterior de modelos de reducción de dimensionalidad.

La obtención de modelos para generar conocimiento sobre la población de la que provienen los datos se puede abordar también mediante la reducción del número de variables con la metodología del análisis factorial que, como en el caso del análisis de correspondencias y el escalado multidimensional, se puede interpretar como una generalización de las componentes principales. Se trata de reemplazar un determinado conjunto de variables por un número menor de factores o variables latentes, no observables, que permitan predecir los valores de las variables originales.

El análisis de la heterogeneidad de la población desde el punto de vista inferencial se realiza mediante lo que se conoce como técnicas de clasificación supervisada. El calificativo de supervisada hace referencia a que se parte de una muestra de elementos cuya clasificación es conocida a partir de los que se obtendrá el modelo para la clasificación de futuras observaciones. Para ello se pueden utilizar distintos métodos estadísticos multivariantes cuya elección dependerá de las características concretas del problema a resolver. Así, cuando todas las variables son continuas, es frecuente aplicar el análisis discriminante clásico de Fisher, que resulta óptimo bajo el supuesto de normalidad multivariante (aunque los datos originales no sean normales es posible aplicar una transformación para obtener normalidad). Si no todas las variables son continuas el problema de clasificación se abordará con otros métodos de clasificación, como los basadas en modelos de respuesta cualitativa, árboles de clasificación, máquinas de los vectores soporte, redes neuronales, etc.

Capítulo 2

Técnicas exploratorias para datos multivariantes.

2.1. Análisis de componentes principales.

Dadas n observaciones de p variables, el análisis de componentes principales persigue representar adecuadamente esta información, con la mínima pérdida de información, mediante un número menor de variables construidas como combinaciones lineales de las originales. La aplicación de esta técnica permite transformar las variables originales, en general correladas, en nuevas variables incorreladas, facilitando la interpretación de los datos. Al mismo tiempo, la representación de las observaciones en un espacio de menor dimensión propicia la identificación de las posibles variables latentes o no observadas que generan los datos.

En otras palabras podemos decir que el análisis de componentes principales consiste en encontrar un subespacio de dimensión menor que p de forma que la proyección sobre él de los n puntos conserve su estructura con la menor distorsión posible. Esta condición de distorsión mínima equivale a exigir que las distancias entre los puntos originales y sus proyecciones en el subespacio obtenido sean lo más pequeñas posible.

Así, dado un elemento x_i y una dirección definida por un vector de norma unidad, $a_1 = (a_{11}, \ldots, a_{1p})'$, la proyección de x_i sobre esta dirección será:

$$z_i = a_{11}x_{i1} + \cdots + a_{1p}x_{ip} = a_1' x_i$$

y el vector que representa esta dirección será $z_i a_1$. Si r_i es la distancia entre x_i y su proyección sobre a_1, entonces

$$minimizar(\sum_{i=1}^{n} r_i^2) = \sum_{i=1}^{n} |x_i - z_i a_1^2|$$

siendo $|u|$ la norma euclídea o módulo de u.

Ahora bien,

$$x_i' x_i = z_i^2 + r_i^2 \Rightarrow \sum_{i=1}^{n} x_i' x_i = \sum_{i=1}^{n} z_i^2 + \sum_{i=1}^{n} r_i^2 \Rightarrow minimizar \sum_{i=1}^{n} r_i^2 = maximizar \sum_{i=1}^{n} z_i^2$$

y, puesto que las proyecciones z_i son variables de media cero, maximizar la suma de sus cuadrados equivale a maximizar su varianza y, en definitiva, el criterio de minimizar la distorsión equivale a encontrar la dirección de proyección que maximice la varianza de los datos proyectados.

2.1.1. Obtención de las componentes principales.

Supongamos ahora que la matriz X de dimensiones n x p contiene los valores de p variables observadas sobre n elementos, y supongamos que los valores de cada variable se encuentran centrados respecto de su media, de forma que las variables de la matriz X tienen media cero y la matriz de varianzas y covarianzas de X es $S = \frac{1}{n} X'X$.

La primera componente principal se define como la combinación lineal de las variables originales que tiene varianza máxima y sus valores para los n elementos se representarán por el vector $z_1 = Xa_1$, que tendrá media cero puesto que las variables originales tienen media cero, y cuya varianza será:

$$Var(z_1) = \frac{1}{n} z_1' z_1 = \frac{1}{n} a_1' X'X a_1 = a_1' S a_1$$

Para que la maximización de la expresión anterior tenga solución se impone la restricción $a_1' a_1 = 1$ que se introduce mediante el multiplicador de Lagrange:

$$M = a_1' S a_1 - \lambda(a_1' a_1 - 1)$$

Esta expresión se maximiza de la forma habitual derivando respecto de los componentes de a_1 e igualando a cero obteniéndose que

$$Sa_1 = \lambda a_1$$

lo que implica que a_1 es un vector propio de S y λ su correspondiente valor propio.

Entonces,

$$Sa_1 = \lambda a_1 \Rightarrow a_1' S a_1 = \lambda a_1' a1 = \lambda \Rightarrow \lambda = Var(z_1)$$

y, puesto que buscamos maximizar $Var(z_1)$, λ será el mayor valor propio de la matriz S y su vector asociado, a_1, definirá los coeficientes de las variables originales en la primera componente principal.

La segunda componente principal proporciona el mejor plano de proyección de X y se calcula estableciendo como función objetivo que la suma de las varianzas de las dos primeras componentes principales sea máxima, es decir, siendo a_1 y a_2 los vectores que definen el plano, la función objetivo será:

$$\phi = a_1' S a_1 + a_2' S a_2 - \lambda_1(a_1' a_1 - 1) - \lambda_2(a_2' a_2 - 1)$$

con las restricciones de que las direcciones tengan módulo unidad, $a_i' a_i = 1,\ i = 1, 2$.

De nuevo, derivando e igualando a cero obtenemos la solución:

$$Sa_1 = \lambda_1 a_1$$

$$Sa_2 = \lambda_2 a_2$$

que indica que a_1 y a_2 son vectores propios de S. Tomando los vectores propios de norma uno y sustituyendo en la función objetivo se obtiene que, en el máximo, su valor es

$$\phi = \lambda_1 + \lambda_2$$

y por tanto, λ_1 y λ_2 deben ser los dos autovalores mayores de S, y a_1 y a_2 sus autovectores.

Nótese que z_1 y z_2 están incorreladas, puesto que:

$$a_1' a2 = 0 \Rightarrow Cov(z_1, z_2) = a_1' S a_2 = 0$$

Análogamente puede demostrarse que el espacio de dimensión r que mejor representa a los n puntos p-dimensionales viene definido por los vectores propios asociados a los r mayores valores propios de S. Estas direcciones se denominan *direcciones principales* de los datos y las nuevas variables que definen, componentes *principales*.

2.1.2. Relación entre las componentes principales y las variables originales.

Si X (y por tanto también S) tiene rango p, existirán tantas componentes principales como variables y se obtendrán calculando los valores propios de S mediante:

$$|S - \lambda I| = 0$$

y sus vectores propios asociados:

$$(S - \lambda_i I) a_i = 0$$

En este caso S será simétrica y definida positiva, y en consecuencia los términos λ_i serán reales y positivos., siendo ortogonales cualesquiera a_i y a_j vectores asociados a las raíces λ_i y λ_j, $\lambda_i \neq \lambda_j$.

Si S es semidefinida positiva de rango $r < p$, es decir $p - r$ variables son combinación lineal del resto, habrá únicamente r valores propios positivos en S y el resto serán nulos.

Si Z es la matriz cuyas columnas contienen los valores obtenidos para las p componentes en los n elementos, las nuevas variables se relacionan con las originales mediante $Z = XA$, con $A'A = I$. Obsérvese que calcular los componentes principales no es mas que calcular una transformación ortogonal A de las variables X para obtener unas nuevas variables Z incorreladas entre sí, lo que puede interpretarse como elegir unos nuevos ejes coordenados que coincidan con los *ejes naturales* de los datos.

2.1.3. Componentes principales estandarizadas.

Es preciso hacer notar que cuando las escalas de medida de las variables son muy distintas, la maximización de su variabilidad dependerá decisivamente de estas escalas de medida y las variables con valores mas grandes tendrán mas peso en el análisis. Así, si una variable tiene una varianza mucho mayor que las demás, la primera componente principal coincidirá muy aproximadamente con esa variable, situación no deseable si es debida a una diferencia en la escala de medida de esta variable respecto de las demás. Asimismo, aun cuando la escala de medida de todas las variables sea la misma, si las variabilidades son muy distintas, al calcular la primera componente principal tendrán mucha mas influencia las variables con varianzas mas elevadas.

Para evitar estas situaciones es conveniente estandarizar las variables con carácter previo al cálculo de las componentes principales, transformando la ecuación a maximizar en:

$$M' = 1 + \sum_{i=1}^{p} \sum_{j=i+1}^{p} a_i a_j r_{ij}$$

en la que r_{ij} es el coeficiente de correlación lineal entre las variables i y j, de forma que ahora la solución depende de las correlaciones.

Las componentes principales se obtendrán entonces calculando los vectores y valores propios de la matriz de coeficientes de correlación R y se denominarán *componentes principales estandarizadas*.

Por lo tanto, cuando las variables originales están medidas en unidades distintas es conveniente calcular las componentes principales estandarizadas, mientras que si están medidas en la misma unidad se podrá aplicar indistintamente cualquiera de las dos opciones pero, si las diferencias entre las varianzas de las variables son informativas y el analista pretende preservar esta información deberá tenerlas en cuenta y no proceder a la estandarización previa de las variables originales.

2.1.4. Interpretación.

Si existe una alta correlación entre todas las variables, la primera componente principal tendrá todas sus coordenadas del mismo signo y puede interpretarse como un promedio ponderado de todas las variables, constituyendo un factor global de *forma*. Las demás

componentes principales se interpretan como factores de forma y habitualmente incluirán coordenadas positivas y negativas contraponiendo grupos de variables.

2.1.5. Selección del número de componentes.

No existe una regla única para decidir el número de componentes principales idóneo, dependiendo de las características del problema concreto ante las que el analista deberá seguir uno u otro criterio. Las reglas mas utilizadas son las siguientes:

- Buscar un "codo" en el gráfico de λ_i frente a i, es decir localizar en dicho gráfico el punto a partir del que los autovalores λ_i son aproximadamente iguales y elegir un número de componentes que excluya los asociados a λ_i pequeños y aproximadamente de la misma magnitud.
- Seleccionar componentes hasta alcanzar una proporción de varianza determinada (usualmente $\geq 90\,\%$). Debe aplicarse con cierta precaución ante la posibilidad de que exista un componente de "tamaño" que recoja el porcentaje de variabilidad total escogido y se desestimen otros componentes que serían adecuados para interpretar la "forma" de las variables.
- Establecer una cota desechando las componentes asociadas a autovalores inferiores (suele utilizarse la varianza media: $\sum_{i=1}^{n} \frac{\lambda_i}{p}$). En el caso estandarizado, el valor medio de las componentes es 1 y esta regla equivale a seleccionar los valores propios mayores que 1. De nuevo, debe aplicarse con precaución.

2.2. Escalado multidimensional.

El escalado multidimensional (MDS) es una técnica que representa las medidas de similaridad (o disimilaridad) entre pares de objetos como distancias entre puntos en un espacio de dimensionalidad menor. De este modo la representación MDS muestra los objetos como puntos en un plano n-dimensional en los que la distancia entre puntos es menor cuanto mas similares son los objetos, lo que facilita al analista identificar comportamientos regulares que podrían permanecer ocultos al estudiar las matrices numéricas en su dimensión original.

2.2.1. Medidas de proximidad para datos multivariantes.

A la hora de abordar el estudio de un conjunto de datos multivariantes, típicamente formado por n individuos sobre los que se han obtenido valores de p variables, es de interés poder establecer el grado de *similitud* entre diferentes individuos de cara, por ejemplo, a realizar agrupaciones o facilitar su representación gráfica y la interpretación de sus relaciones.

En estadística multivariante es habitual distinguir entre medidas de asociación para individuos y para variables aunque técnicamente estas medidas son válidas tanto para unos como para otras.

Existen muchas medidas de asociación multivariante. Entre las mas conocidas podemos citar el coseno del ángulo entre vectores, el coeficiente de correlación, la distancia euclídea, la distancia de Mahalanobis o distancias basadas en el estadístico χ^2.

Cada medida diferente refleja la asociación entre dos variables (o individuos) en un sentido particular por lo que es de gran importancia elegir, en cada situación, la medida apropiada para el problema concreto de que se trate.

En el caso del análisis de variables textuales es generalizado cuantificar el grado de similaridad entre dos textos mediante el coseno del ángulo que forman los vectores que los representan.

Distancias y similaridades.

Se denomina *distancia* o *métrica* entre dos puntos, x_i y x_j, pertenecientes a $\mathbb{R}^p$ a la función $d : \mathbb{R}^p$ x $\mathbb{R}^p \to \mathbb{R}^+$ que verifica las siguientes propiedades:

- $d(x_i, x_j) \geq 0$
- $d(x_i, x_i) = 0, \quad \forall i$
- $d(x_i, x_j) = d(x_j, x_i)$
- $d(x_i, x_j) \leq d(x_i, x_p) + d(x_p, x_j), \ \forall \ x_j \in \ \mathbb{R}^p$

De forma similar, se puede definir la *similaridad* entre dos puntos, x_i y x_j, pertenecientes a $\mathbb{R}^p$ como la función $s : \mathbb{R}^p$ x $\mathbb{R}^p \to \mathbb{R}^+$ que verifica las siguientes propiedades:

- $s(x_i, x_j) \leq s_0$
- $s(x_i, x_i) = s_0 \forall i$
- $s(x_i, x_j) = s_0 \Rightarrow x_i = x_j$
- $s(x_i, x_j) = s(x_j, x_i)$
- $|s(x_i, x_p) + s(x_p, x_j)|s(x_i, x_j) \geq s(x_i, x_p)s(x_p, x_j),\ \forall\ x_p \in\ \mathbb{R}^p$

Las métricas de Minkowski y la distancia euclídea.

Unas distancias empleadas muy frecuentemente en análisis multivariante son las denominadas distancias o métricas de *Minkowski*, definidas en función de un parámetro r de la siguiente forma:

$$d_{ij}^{(r)} = \left[\sum_{s=1}^{p}(x_{is} - x_{js})^r\right]^{1/r}$$

Es inmediato que la distancia de *Minkowski* de parámetro $r = 2$ es la distancia euclídea, que es la mas utilizada pero tiene el inconveniente de que depende de las unidades de medida de las variables, por lo que en su lugar se suelen utilizar las denominadas *métricas euclídeas ponderadas* en las que se divide cada variable por un término que elimine el efecto de la escala. Es decir

$$d_{ij} = [(x_i - x_j)' M^{-1}(x_i - x_j)]^{1/2}$$

siendo M una matriz diagonal empleada para estandarizar las variables y conseguir que la medida sea invariante ante cambios de escala.

La matriz M puede no ser diagonal pero siempre deberá ser singular y definida positiva de forma que se verifique la condición $d_{ij} \geq 0$.

De nuevo, en el caso $M = I$ tendremos la distancia euclídea.

La distancia de Mahalanobis.

Esta distancia se obtiene al tomar $M = S$ y definir la distancia entre un punto y su vector de medias:

$$d_i = [(x_i - \bar{x})' S^{-1} (x_i - \bar{x})]^{1/2}$$

Es frecuente referirse al valor de esta distancia al cuadrado, d_i^2, con la misma denominación: distancia de Mahalanobis y así se hace a lo largo de este trabajo en el apartado dedicado al *Análisis discriminante*.

El coseno del ángulo entre vectores.

Si $x_i = (x_{i1}, \ldots, x_{ip})'$ y $x_j = (x_{j1}, \ldots, x_{jp})'$ son dos vectores p-dimensionales, su producto escalar, o suma de sus productos cruzados, es $x_i' x_j = \sum_{s=1}^{p} x_{is} x_{js}$, y $||x_i||^2 = \sum_{s=1}^{p} x_{is}^2 = \sum_{s=1}^{p} x_{is} x_{is}$ es su norma al cuadrado o suma de cuadrados. Entonces, si β es el ángulo formado por los vectores x_i y x_j, su producto escalar se puede expresar como:

$$x_i' x_j = ||x_i|| ||x_j|| cos(\beta)$$

de donde

$$cos(\beta) = \frac{x_i' x_j}{||x_i|| ||x_j||} = \frac{\sum_{s=1}^{p} x_{is} x_{js}}{\sqrt{\sum_{s=1}^{p} x_{is}^2 \sum_{s=1}^{p} x_{js}^2}}$$

Es fácil comprobar que cuando los vectores se encuentran centrados respecto de su media esta medida coincide con el coeficiente de correlación de Pearson.

El coseno del ángulo formado por dos vectores es por tanto una medida de similaridad entre los mismos que toma valores entre -1 y 1 siendo, además, la mejor medida para establecer el paralelismo entre dos vectores, al ser en este caso 1 en valor absoluto. Además, el coseno es invariante ante homotecias excepto un eventual cambio de signo y, por tanto, independiente de la longitud de los vectores considerados. Todas estas características hacen que sea una medida muy utilizada en el análisis textual a la hora de cuantificar el grado de similitud entre dos textos.

2.2.2. Escalado métrico.

MDS trabaja sobre una matriz de distancias (o proximidades) $D = [d_{ij}]_{n_x n}$. En el caso general estas distancias serán euclídeas, es decir, supuestas dos observaciones, x_i y x_j en un espacio p-dimensional: $d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + \cdots + (x_{ip} - x_{jp})^2}$.

Estas distancias no se ven alteradas si se centran las variables respecto de su media, es decir:

$$d_{ij}^2 = \sum_{s=1}^{p}(x_{is} - x_{js})^2 = \sum_{s=1}^{p}[(x_{is} - \bar{x}_s) - (x_{js} - \bar{x}_s)]^2$$

y por lo tanto no hay pérdida de generalidad en suponer que las variables tienen media cero.

Se trata de encontrar, a partir de la matriz de distancias D, una matriz $\tilde{X}$ n x p con variables de media cero. Su matriz de covarianzas tendrá la forma:

$$S = \frac{\tilde{X}'\tilde{X}}{n}$$

siendo $Q = \tilde{X}\tilde{X}'$ la correspondiente matriz de productos cruzados que puede a su vez interpretarse como una matriz de similitudes entre los n elementos. Sus términos serán de la forma:

$$q_{ij} = \tilde{x}_i'\tilde{x}_j = \sum_{s=1}^{p} x_{is}x_{js}$$

con lo que las distancias pueden ser deducidas inmediatamente a partir de Q:

$$d_{ij}^2 = \sum_{s=1}^{p}(x_{is} - x_{js})^2 = \sum_{s=1}^{p} x_{is}^2 + \sum_{s=1}^{p} x_{js}^2 - 2\sum_{s=1}^{p} x_{is}x_{js} = q_{ii} + q_{jj} - 2q_{ij}$$

Dado que $\tilde{X}'1 = 0$, también $Q1 = 0$, es decir, $\sum_{i=1}^{n} q_{ij} = 0$ (y por tanto, al ser Q simétrica, $\sum_{j=1}^{n} q_{ij} = 0$). Imponiendo estas restricciones resulta:

$$\sum_{i=1}^{n} d_{ij}^2 = \sum_{i=1}^{n} q_{ii} + nq_{jj} = t + nq_{jj}$$

$$\sum_{j=1}^{n} d_{ij}^2 = \sum_{j=1}^{n} q_{jj} + nq_{ii} = t + nq_{ii}$$

con $t = traza(Q) = \sum_{i=1}^{n} q_{ii}$, y entonces

$$\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^2 = 2nt \;\Rightarrow\; d_{ij}^2 = \frac{1}{n}\sum_{i=1}^{n} d_{ij}^2 - \frac{t}{n} + \frac{1}{n}\sum_{j=1}^{n} d_{ij}^2 - \frac{t}{n} - 2q_{ij} = d_{i.}^2 + d_{.j}^2 - d_{..}^2 - 2q_{ij}$$

Por lo tanto

$$q_{ij} = -\frac{1}{2}(d_{ij}^2 - d_{i.}^2 - d_{.j}^2 + d_{..}^2)$$

donde $d_{i.}^2$, $d_{.j}^2$ y $d_{..}^2$ son, respectivamente, la media por filas, por columnas y total de los elementos de D, es decir:

$$d_{i.}^2 = \frac{1}{n}\sum_{j=1}^{n} d_{ij}^2$$

$$d_{.j}^2 = \frac{1}{n}\sum_{i=1}^{n} d_{ij}^2$$

$$d_{..}^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^2$$

Hemos obtenido entonces la matriz Q de similitud a partir de la matriz de distancias D.

Si suponemos que Q es definida positiva de rango p podemos expresarla como $Q = V\Lambda V'$, donde V contiene los vectores propios que corresponden a los valores propios no nulos de Q y Λ es diagonal y contiene los valores propios de Q.

Entonces, podemos escribir:

$$Q = (V\Lambda^{1/2})(\Lambda^{1/2}V')$$

obteniendo una matriz $V\Lambda^{1/2} = Y$, n x p, con p variables incorreladas que reproducen la métrica inicial.

Se debe observar que la matriz obtenida no estará formada por las variables originales sino por sus componentes principales pues existe una indeterminación en el problema al partir únicamente de la matriz de distancias D que es función de la matriz de similitud, Q, y esta es invariante ante rotaciones de las variables originales.

Es frecuente que la matriz de distancias D no sea compatible con una métrica euclídea, pero también lo es que la matriz de similitud Q obtenida a partir de ella tenga p valores

propios positivos y mayores que el resto y, si estos restantes $n - p$ valores propios no nulos son mucho menores, podemos obtener una representación aproximada de los puntos originales mediante los p vectores propios asociados a los primeros p valores propios positivos de Q, en cuyo caso las representaciones conservarán la distancia entre los puntos de manera aproximada.

El procedimiento para obtener las *coordenadas principales* de los puntos originales a partir de una matriz de distancias D es el siguiente:

1. Se construye $Q = -\frac{1}{2}PDP$, con $P = I - \frac{1}{n}11'$ (se puede comprobar que en estas condiciones Q es semidefinida positiva y D compatible con una métrica euclídea)
2. Se obtienen los valores propios de Q y se toman los r mayores de forma que los restantes $n - r$ valores sean próximos a 0.
3. Se considera $Q \approx (V_r\Lambda_r^{1/2})(\Lambda_r^{1/2}V_r')$ lo que implica tomar como coordenadas de los puntos originales $Y_r = V_r\Lambda_r^{1/2}$ y por tanto $y_i = v_i\sqrt{\lambda_i}$ donde λ_i es un valor propio de Q y v_i su vector propio asociado.

Además, es posible calcular la precisión obtenida mediante la aproximación realizada a partir de los p valores propios positivos de Q. Por ejemplo, mediante el coeficiente propuesto por *Mardia*:

$$m_{1,p} = 100\frac{\sum_{i=1}^{p}\lambda_i}{\sum_{i=1}^{p}|\lambda_i|}$$

2.2.2.1. Relación entre coordenadas y componentes principales.

El escalado multidimensional está muy relacionado con el análisis de componentes principales. Ambos persiguen la reducción de la dimensionalidad de los datos originales. En el caso del análisis de componentes principales se obtienen los valores propios de la matriz $X'X$ y se proyectan las variables sobre las direcciones obtenidas para obtener los valores de las componentes principales, mientras que en el caso del escalado multidimensional las coordenadas principales se obtienen directamente como vectores propios de la matriz XX', pero si la matriz de similaridades Q proviene de una métrica euclídea los dos métodos conducen al mismo resultado.

No obstante es necesario precisar que el escalado multidimensional tiene aplicación en una gama de problemas mayor, puesto que siempre es posible obtener las coordenadas principales, aun cuando la matriz de distancias D no provenga exactamente de un conjunto de variables originales, como en el caso del escalado multidimensional *no métrico,* en los que la matriz de partida esta compuesta por las diferencias o *disimilitudes* entre objetos, obtenidas en general por las opiniones de un conjunto de *jueces* o por procedimientos de *ordenación*.

2.2.3. Escalado no métrico.

En estos casos se parte de la premisa de que la matriz de disimilaridades se relaciona con una matriz de distancias de una manera compleja, es decir, las variables explicativas de las similitudes entre los elementos comparados determinan unas distancia euclídeas, d_{ij}, entre los mismos relacionadas con las similitudes, δ_{ij}, mediante una función desconocida:

$$\delta_{ij} = f(d_{ij})$$

imponiendo la única condición de que f sea monótona, es decir, $\delta_{ij} > \delta_{ih} \Leftrightarrow d_{ij} > d_{ih}$, con lo que se pretende encontrar unas coordenadas capaces de reproducir las distancias originales a partir únicamente de la condición de monotonía. Para ello será preciso definir un criterio de bondad del ajuste que sea invariante ante transformaciones monótonas de los datos y un algoritmo para obtener las coordenadas optimizando el criterio anterior.

Aunque para este tipo de problemas no existe solución única y se han propuesto diversos procedimientos, uno de los mas utilizados consiste en minimizar las diferencias entre las distancias derivadas de las coordenadas principales, $\hat{d}_{ij}$, y las similitudes de partida, δ_{ij}, para todos los términos de la matriz, es decir: minimizar $\sum_{i<j}(\delta_{ij} - \hat{d}_{ij})^2$. Al estandarizar esta cantidad se obtiene un criterio de ajuste denominado STRESS, S^2, dado por la expresión:

$$S^2 = \frac{\sum_{i<j}(\delta_{ij} - \hat{d}_{ij})^2}{\sum_{i<j}\delta_{ij}^2}$$

Otro criterio habitual es el conocido como S-STRESS basado en minimizar las distancias al cuadrado, $\hat{d}_{ij}$, que se determinarán obteniendo p coordenadas principales que se emplean

como variables implícitas, y_{ij}, $i = 1, \ldots, n, \quad j = 1, \ldots, p$, de forma que las distancias euclídeas entre ellas se expresan en la forma:

$$\hat{d}_{ij}^2 = \sum_{s=1}^{p} (y_{is} - y_{js})^2$$

Se suele considerar como valor inicial de estas variables la solución proporcionada por las *coordenadas principales* y se itera para mejorar la solución minimizando el criterio S^2. El número de dimensiones para obtener una buena representación, p, se suele estimar probando distintos valores y estudiando la evolución de S^2. Una vez fijado p se plantea el problema de minimizar S^2 para las distancias entre las variables y_{ij}. Para ello, se deriva S^2 respecto a cada término y_{ip} y se iguala a cero, obteniendo:

$$\frac{\partial S^2}{\partial y_{ip}} = 2 \sum_{j=1}^{n} (\delta_{ij} - \hat{d_{ij}}) \frac{\partial \hat{d_{ij}}}{\partial y_{ip}} = 0$$

Dado que $\frac{\partial \hat{d_{ij}}}{\partial y_{ip}} = \frac{(y_{ip} - y_{jp})}{\hat{d_{ij}}}$, sustituyendo en las ecuaciones anteriores llegamos a las ecuaciones:

$$y_{ip} \sum_{j=1}^{n} \frac{(\delta_{ij} - \hat{d}_{ij})}{\hat{d}_{ij}} - \sum_{j=1}^{n} \frac{(\delta_{ij} - \hat{d}_{ij})}{\hat{d}_{ij}} y_{jp} = 0$$

Entonces, el sistema de ecuaciones resultante derivando para los np valores de las coordenadas principales puede expresarse como

$$FX = 0$$

siendo F una matriz cuadrada y simétrica de orden n cuyos coeficientes son:

$$f_{ij} = -\frac{\delta_{ij} - \hat{d}_{ij}}{\hat{d}_{ij}}, \quad i \neq j$$

$$f_{ii} = \sum_{j=1, j \neq i}^{n} f_{ij}, \quad i = j$$

2.2.4. Unfolding

La técnica denominada *Unfolding* es un caso especial de escalado multidimensional para el análisis de datos de preferencia. Surge como una técnica derivada del MDS para posicionar en un mismo subespacio los conjuntos de individuos y de objetos o estímulos observados en base a ciertos juicios de preferencia. En esta técnica se considera que la información de partida consiste en una matriz $n\ x\ p$ que contiene las preferencias de n individuos acerca de p estímulos u objetos. Un ejemplo muy clarificador lo constituye el conjunto de datos *breakfast* incluido en el paquete *smacof* de R que contiene las preferencias de 42 individuos acerca de 15 alimentos (Green y Rao, 1972) que pueden formar parte del desayuno tales como diferentes tipos de tostadas, pastas, etc... Se solicitó a cada uno de los 42 individuos que ordenara estos 15 elementos según su preferencia, otorgando el valor 1 al alimento preferido y el valor 15 al alimento menos preferido para su desayuno.

```
##     toast butoast engmuff jdonut cintoast bluemuff hrolls toastmarm
## 1      13      12       7      3        5        4      8        11
## 2      15      11       6      3       10        5     14         8
## 3      15      10      12     14        3        2      9         8
## 4       6      14      11      3        7        8     12        10
## ...   ...     ...     ...    ...      ...      ...    ...       ...
## 39      6       1      12      5       15        9      2         7
## 40     14       1       5     15        4        6      3         8
## 41     10       3       2     14        9        1      8        12
## 42     13       3       1     14        4       10      5        15
```

Las preferencias pueden entonces concebirse como proximidades entre los elementos de dos conjuntos, individuos y objetos de elección, lo que a su vez se traduce en un caso especial de MDS en el que faltan las proximidades dentro de cada uno de los conjuntos de tal forma que los individuos se representan como puntos "ideales" en el espacio MDS en los que las distancias de cada punto ideal a cada uno de los objetos son las correspondientes puntuaciones de preferencia. En el ejemplo anterior el alimento para desayunar preferido por el individuo 1 es *danpastry* (pastas danesas), que será por tanto el más *próximo* a él, mientras que el último de su preferencia es *toastmarg* (pan tostado con margarina), que deberá ser por tanto el mas *distante*.

Por lo tanto, las filas y las columnas de la matriz de proximidades no coinciden, como sucede en MDS, sino que se trata de una matriz rectangular que contiene medidas de proximidad entre elementos de conjuntos distintos. La "transformación" que se realiza para aplicar MDS sobre este tipo de matrices es considerar que se trata de una matriz cuadrada *(n+p) x (n+p)* con valores perdidos. Concretamente, tanto las filas como las columnas representan a la totalidad de individuos y estímulos y se consideran como datos perdidos las proximidades "internas" en cada uno de estos conjuntos, es decir, por un lado las proximidades entre individuos y por otro lado las proximidades entre estímulos u objetos.

Sea $I = \{I_1, \ldots, I_n\}$ el conjunto de individuos y $O = \{O_1, \ldots, O_p\}$ el conjunto de estímulos, y supongamos las preferencias entre individuos y estíumulos $p_{ij},\ i = (1, \ldots, n),\ j = (1, \ldots, p)$:

	I_1	$\ldots$	I_n	O_1	$\ldots$	O_p
I_1				p_{11}	$\ldots$	p_{1p}
$\vdots$				$\vdots$	$\ddots$	$\vdots$
I_n				p_{n1}	$\ldots$	p_{np}
O_1	p_{11}	$\ldots$	p_{n1}			
$\vdots$	$\ddots$	$\vdots$				
O_p	p_{1p}	$\ldots$	p_{np}			

La solución aportada por *unfolding* consiste en la representación de los individuos y los estímulos como puntos en un mismo espacio multidimensional, es decir, a partir de los datos de la matriz de preferencias se obtendrá un punto para cada individuo y un punto para cada estímulo. Los puntos de individuos se denominan puntos *ideales*, en el sentido de que un punto de estímulo situado en las mismas coordenadas sería *ideal* para ese individuo. De esta forma las preferencias son representadas por las distancias entre los puntos *ideales* (de los individuos) y los puntos de los estímulos. Así, las preferencias de cada individuo se modelan relacionando los puntos *ideales* con los puntos que representan los estímulos de forma que, para un individuo determinado, los estímulos con mas preferencia son los que están más cerca del punto *ideal* mientras que los estímulos mas alejados son los menos preferidos por el individuo en cuestión.

Algebraicamente, la versión métrica de este modelo puede representarse como:

$$\delta_{ij} = f(d_{ij}) \; ; \quad d_{ij} = \sqrt{\sum_{t=1}^{r}(x_{jt} - y_{jt})^2}$$

donde δ_{ij} es el valor de preferencia del individuo i acerca del estímulo j, x_{jt} es la proyección del punto del estímulo j sobre la dimensión t, y_{jt} es la proyección del punto *ideal* i sobre la dimensión t y f es un función continua, paramétrica y monótona.

La versión no métrica se definirá de manera equivalente pero siendo $\delta_{ij} = g(d_{ij})$ con g una transformación arbitraria que únicamente obedece a la restricción monótona $\delta_{ij} < \delta_{rs} \;\Rightarrow\; g(\delta_{ij}) \leq g(\delta_{rs})$.

2.3. Análisis de correspondencias.

El análisis de correspondencias es una técnica descriptiva enfocada al estudio y representación de tablas de contingencia.

Una tabla de contingencia consiste, en general, en un conjunto de números positivos dispuestos en forma matricial de forma que el valor de cada casilla representa la frecuencia absoluta observada para la combinación de las dos variables correspondientes a la fila y columna en cuestión, es decir, se tiene una matriz de valores numéricos k_{ij} que representan el número de individuos pertenecientes a la clase i de la característica I y a la clase j de la característica J, donde tanto I como J clasifican o particionan la población objeto de estudio. En este tipo de tablas no tendrá sentido, por tanto, distinguir entre variables e individuos, jugando ambos un papel simétrico.

El análisis de correspondencias es una técnica equivalente al análisis de componentes principales específicamente desarrollada para el estudio de variables cualitativas, siendo la información de partida la tabla de contingencia que recoge las frecuencias absolutas observadas de dos variables cualitativas (representadas, respectivamente, en las filas y en las columnas) en n elementos.

Se trata de un procedimiento que permite resumir la información de una tabla de contingencia obteniendo una representación de las variables en un espacio de dimensión menor, como en el análisis de componentes principales, pero realizando unas determinadas

transformaciones sobre la tabla de contingencia inicial y utilizando la distancia χ^2 en lugar de la distancia euclídea que no es apropiada para la interpretación con este tipo de datos.

La tabla de contingencia inicial necesita ser transformada de acuerdo con su naturaleza y características particulares. Para ello, se trabaja con la matriz F de frecuencias relativas en la que cada casilla de la tabla inicial se divide por el número total de elementos observados, n, con lo que las casillas de F verifican:

$$\sum_{i=1}^{I}\sum_{j=1}^{J} f_{ij} = 1$$

Es decir, dada una tabla inicial de contingencia:

	1	...	j	...	J	
1	n_{11}	...	n_{1j}	...	n_{1J}	$n_{1.}$
...	...	...	...	...	...	...
i	n_{i1}	...	n_{ij}	...	n_{iJ}	$n_{i.}$
...	...	...	...	...	...	...
I	n_{I1}	...	n_{Ij}	...	n_{IJ}	$n_{I.}$
	$n_{.1}$	...	$n_{.j}$	...	$n_{.J}$	n

consideraremos la tabla transformada conocida como tabla o matriz de correspondencias:

	1	...	j	...	J
1	$f_{11} = \frac{n_{11}}{n}$	...	$f_{1j} = \frac{n_{1j}}{n}$	...	$f_{1J} = \frac{n_{1J}}{n}$
...	...	...	...	...	...
i	$f_{i1} = \frac{n_{i1}}{n}$	...	$f_{ij} = \frac{n_{ij}}{n}$	...	$f_{iJ} = \frac{n_{iJ}}{n}$
...	...	...	...	...	...
I	$f_{I1} = \frac{n_{I1}}{n}$	...	$f_{Ij} = \frac{n_{Ij}}{n}$	...	$f_{IJ} = \frac{n_{IJ}}{n}$

Se realiza esta transformación para eliminar el efecto que, sobre el cálculo de distancias entre puntos-fila (o puntos-columna) tiene el efectivo total de cada fila (o columna).

2.3.1. Proyección de las filas.

Las I filas pueden tomarse como I puntos en el espacio $\mathbb{R}^J$ y se busca una representación en un espacio de dimensión menor que permita apreciar las distancias relativas entre ellos. Para ello es necesario tener en cuenta que, por las características intrínsecas de las tablas de contingencia, las filas no tienen el mismo peso puesto que el número de datos de cada una puede ser diferente, es decir, cada fila de F tiene una frecuencia relativa específica $f_{i.} = \sum_{j=1}^{J} f_{ij}$. En esta situación la distancia euclídea no es una buena medida de la proximidad puesto que aunque las frecuencias relativas de dos filas sean muy distintas esto puede ser debido únicamente al distinto número de elementos *contados* en cada fila siendo una de ellas simplemente producto de la otra por un determinado escalar. La distancia euclídea entre dos filas de este tipo arrojará un valor alto que no será reflejo de diferencias en la estructura de las filas sino únicamente de su distinta frecuencia relativa. Para evitar esta distorsión se divide cada casilla de la matriz por la frecuencia relativa de su fila, $f_{i.}$, obteniendo una nueva matriz transformada en la que los valores de cada casilla representan ahora la frecuencia relativa de la variable columna condicionada a la variable fila. De esta forma, las dos filas *linealmente dependientes* anteriores resultarán ahora idénticas y la distancia euclídea entre ellas será 0 reflejando que no existen diferencias en la estructura de ambas aun cuando sus frecuencias relativas sean muy distintas.

Si llamamos R a esta matriz de frecuencias relativas condicionadas a los totales por filas y D_f a la matriz diagonal de dimensiones I x I con las frecuencias relativas de las filas, $f_{i.}$, podemos escribir:

$$R = D_f^{-1} F$$

Ahora cada fila de R representa la distribución de la variable en columnas condicionada al atributo correspondiente a la fila.

Las filas, r_i' de R pueden considerarse puntos de $\mathbb{R}^J$ y puesto que $\sum_{j=1}^{J} r_{ij}' = 1, \;\; i = 1, \ldots, I$, todos los puntos r_i' se encuentran en un espacio de dimensión $J - 1$.

Se trata ahora de proyectar estos puntos en un espacio de dimensión menor que preserve las distancias relativas entre filas en el sentido de que las filas que tengan estructuras similares estén próximas y las que tengan estructuras diferentes se encuentren alejadas. Para ello se debe definir una medida de distancia entre los puntos-fila pero la distancia euclídea

no resulta apropiada puesto que considera de la misma forma todos los componentes de cada punto-fila y es necesario tener en cuenta la magnitud de las frecuencias de los atributos representados por las columnas en los cambios relativos entre estos atributos, para lo que se ponderan las diferencias en frecuencia relativa entre dos atributos de manera inversamente proporcional a la frecuencia de cada atributo, es decir, en lugar de calcular la distancia euclídea entre dos puntos r_a y r_b:

$$d_{ab} = \sum_{j=1}^{J} (r_{aj} - r_{bj})^2$$

se calculará la distancia χ^2 definida como:

$$D^2(r_a, r_b) = \sum_{j=1}^{J} \frac{(r_{aj} - r_{bj})^2}{f_{.j}}$$

o, en expresión matricial, $D^2(r_a, r_b) = (r_a - r_b)' D_c^{-1} (r_a - r_b)$, siendo D_c la matriz diagonal con términos $f_{.j}$.

Esta distancia equivale a la distancia euclídea entre los vectores transformados $y_i = D_c^{-1/2} r_i$ lo que permite simplificar el problema definiendo una nueva matriz de datos transformada de forma que tenga sentido aplicar la distancia euclídea:

$$Y = RD_c^{-1/2} = D_f^{-1} F D_c - 1/2$$

cuyos términos representan las frecuencias relativas condicionadas por filas y estandarizadas por su variabilidad, que serán directamente comparables entre sí y que serán de la forma:

$$y_{ij} = \frac{f_{ij}}{f_{i.}\sqrt{f_{.j}}}, \quad i = 1, \ldots, I, \ j = i, \ldots, J$$

La distancia χ^2 tiene una propiedad importante conocida como *principio de equivalencia distribucional* que viene a decir que si dos filas (o análogamente dos columnas) tiene la misma estructura relativa,

$$\frac{f_{ij}}{f_{i.}} = \frac{f_{kj}}{f_{k.}}, \quad j = 1, \ldots, J$$

y se agregan en una nueva fila única, las distancias entre las restantes filas permanecen invariables.

En el Análisis de Correspondencias esta propiedad permite que si existe una suficiente proximidad entre los perfiles de dos filas (o columnas) puedan sustituirse por una única fila (o columna) como agregación de las anteriores sin que los resultados se vean alterados sustancialmente.

Ahora bien, para calcular la proyección de Y no será del todo correcto considerarla como una matriz de variables continuas tal y como se hace para encontrar las *componentes principales* puesto que, una vez mas, hay que tener en cuenta que al tratarse de una tabla de contingencia las filas tienen una distinta frecuencia relativa y por tanto deben tener distinto peso. Es decir, es necesario que las filas con mayor número de elementos estén bien representadas aunque ello suponga una peor representación de las filas con pocos elementos. Para ello se otorga a cada fila un peso proporcional al número de elementos que contiene, lo cual se lleva a cabo maximizando la suma de cuadrados ponderada:

$$m = a'Y'D_fYa$$

sujeto a $a'a = 1$, lo que equivale a: $m = a'D_c^{-1/2}F'D_f^{-1}FD_c^{-1/2}a$.

Alternativamente, se puede construir la matriz que estandariza las frecuencias relativas en cada casilla por el producto de las raíces cuadradas de las frecuencias relativas totales de la fila y columna:

$$Z = D_f^{-1/2}FD_c^{-1/2}$$

que tendrá sus componentes de la forma

$$z_{ij} = \frac{f_{ij}}{\sqrt{f_{i.}f_{.j}}}, \quad i = 1, \ldots, I, \ j = i, \ldots, J$$

Se tratará entonces de encontrar el vector a que maximice $m = a'Z'Za$ sujeto a $a'a = 1$, es decir, encontrar las *componentes principales* de Z y, por tanto, su solución será:

$$D_c^{-1/2}F'D_f^{-1}FD_c^{-1/2}a = \lambda a$$

donde a es un vector propio de $Z'Z$ y λ su valor propio asociado.

Dado que la matriz $Z'Z$ tiene como mayor valor propio siempre 1, se descarta esta solución trivial que no aporta información acerca de la estructura de las filas y se toma

el mayor valor propio menor que 1 y su vector propio asociado, proyectando Y sobre la dirección encontrada para obtener la mejor representación, en una dimensión, de las filas de la tabla de contingencia original:

$$y_f(a) = Ya = D_f^{-1} F D_c^{-1/2} a$$

Análogamente, al extraer el vector propio ligado al siguiente mayor valor propio de $Z'Z$ se obtiene la segunda coordenada para cada fila de su mejor representación en un espacio de dimensión dos. De esta forma, las coordenadas de la mejor representación bidimensional de las filas vendrán dadas por las filas de la matriz:

$$C_f = YA_2 = D_f^{-1} F D_c^{-1/2} A_2$$

donde $A_2 = [a_1 a_2]$ es la matriz que contiene en columnas los dos vectores propios de $Z'Z$ asociados a los dos mayores valores propios menores que la unidad.

Este procedimiento se puede extender para obtener la mejor representación de las filas en mas dimensiones mediante el cálculo de los vectores propios asociados a los siguientes valores propios de $Z'Z$ en orden decreciente.

2.3.2. Proyección de las columnas.

Dado que en las tablas de contingencia, en general, no tiene sentido diferenciar entre individuos y variables, es decir, entre filas y columnas, debido a que cada una de estas dimensiones representa los atributos de una determinada variable, es posible aplicar a las columnas un análisis equivalente al descrito en la sección anterior para las filas, considerando ahora las J columnas como J puntos en $\mathbb{R}^I$.

Si llamamos $c = F'1$ al vector de frecuencias relativas de las columnas y D_c a la matriz diagonal que las contiene, la búsqueda de la mejor representación de los puntos-columna en un espacio de dimensión menor conducirá, aplicando la distancia χ^2 a estudiar la matriz $D_c^{-1} F' D_f^{-1/2}$, problema idéntico al de la sección anterior intercambiando el papel de las matrices D_c y D_f por lo que las direcciones de proyección serán ahora los vectores propios de la matriz:

$$ZZ' = D_f^{-1/2} F D_c^{-1} F' D_f^{-1/2}$$

siendo Z la matriz definida en la sección anterior.

Dado que $Z'Z$ y ZZ' tienen los mismos valores propios no nulos, esta última matriz tendrá también un valor propio unidad ligado al vector propio **1** y, llamando b al mayor valor propio menor que 1 de ZZ', la mejor representación unidimensional de los puntos-columna vendrá dada por:

$$y_c(b) = Y'b = D_c^{-1}F'D_f^{-1/2}b$$

Análogamente, la mejor representación bidimensional de las columnas viene dada por las coordenadas definidas por las filas de la matriz:

$$C_c = Y'B_2 = D_c^{-1}F'D_f^{-1/2}B_2$$

siendo $B_2 = [b_1 b_2]$ la matriz que contiene en columnas los vectores propios ligados a los dos valores propios mayores de ZZ' menores que la unidad.

2.3.3. Análisis conjunto.

El carácter simétrico del problema estudiado con el análisis de correspondencias hace que resulte de especial interés la representación conjunta de las filas y columnas de la matriz.

Las matrices $Z'Z$ y ZZ' tienen los mismos valores propios no nulos y, si a_i es un vector propio de $Z'Z$ ligado al valor propio λ_i: $Z'Za_i = \lambda_i a_i \Rightarrow ZZ'(Za_i) = \lambda_i(Za_i) \Rightarrow b_i = Za_i$, donde b_i es un vector propio de ZZ' ligado a λ_i.

En consecuencia, un método para obtener los vectores propios es calcular directamente los correspondientes a la matriz de dimensión menor, $Z'Z$ o ZZ', y a partir de estos obtener los restantes como Za_i o $Z'b_i$

En el análisis de correspondencias, como en el caso del análisis de componentes principales, la proporción de variabilidad explicada por cada dimensión se calcula descartando el valor propio igual a 1 y tomando la proporción que representa cada valor propio en relación a los restantes.

La interpretación de los ejes se apoyará sobre los elementos que presentan una fuerte contribución a la construcción de cada uno de ellos junto al análisis de sus coordenadas y la calidad de su representación. La contribución indica, eje a eje, la inercia aportada por cada perfil-fila (y respectivamente, cada perfil-columna) mientras que la calidad de la representación se suele calcular mediante el coseno cuadrado entre los perfiles y cada eje. De esta forma, las contribuciones, que sumarán 100 sobre cada eje, representan la importancia porcentual del perfil en la construcción del eje, y los cosenos cuadrados que para cada perfil suman 1 sobre todos los ejes, permiten apreciar la importancia de cada eje en la representación del perfil o, lo que es lo mismo, respecto de que eje es mejor la representación de cada perfil.

En cuanto a la interpretación de los planos formados por pares de ejes, permite conjugar la de cada uno de los ejes implicados dando lugar a interpretaciones complementarias. Aunque en el análisis de correspondencias se considera apropiado representar las filas y las columnas sobre un mismo gráfico, la interpretación en este caso se debe realizar con mucha precaución, puesto que se justifica en las relaciones de transición que enlazan las coordenadas de una fila con las de todas las palabras y, respectivamente, las que enlazan las coordenadas de una columna con las de todas las filas, pero en ningún caso justificarán la interpretación de la proximidad entre una fila y una columna, sino únicamente entre una fila y el conjunto de todas las palabras (y respectivamente, entre una columna y todas las filas).

En definitiva, el análisis de correspondencias de una tabla de contingencia de dimensiones I x J se lleva a cabo en los siguientes pasos:

1. Se obtiene la matriz de frecuencias relativas, F y se transforma en la matriz estandarizada de frecuencias relativas en la que cada celda se divide por la raíz de los totales de su fila y columna, $z_{ij} = \frac{f_{ij}}{\sqrt{f_{i.}f_{.j}}}, \quad i = 1, \ldots, I, \quad j = 1, \ldots, J$
2. Se calculan los h vectores propios ligados a valores propios mayores distintos de la unidad de la matriz de menor dimensión entre $Z'Z$ y ZZ', y se obtienen los restantes de la siguiente forma: si $Z'Z$ es la matriz de menor dimensión se calculan directamente sus vectores propios a_i y a continuación los restantes aplicando $b_i = Za_i$, mientras que si la dimensión de ZZ' es menor se calcular sus vectores propios b_i y posteriormente los vectores propios de $Z'Z$ como $a_i = Z'b_i$

Las I filas y las J columnas se representarán como puntos en $\mathbb{R}^h$ con coordenadas dadas, respectivamente, por

$$C_f = D_f^{-1/2} Z A_h$$

$$C_c = D_c^{-1/2} Z' B_h$$

donde A_h contiene en columnas los h vectores propios de $Z'Z$ correspondientes a los h valores propios distintos a la unidad en orden decreciente, y B_h los correspondientes h vectores propios asociados de ZZ'.

En general se suele considerar $h = 2$ para obtener una representación bidimensional aunque se puede establecer otro valor de h en función de las particularidades del problema concreto para lo que se suele fijar una cota inferior para la proporción acumulada de variabilidad explicada y considerar el valor de h que la supere.

Capítulo 3

Técnicas estadístcas de clasificación. Análisis discriminante.

El problema de la clasificación (o discriminación) es común en muchas áreas de las ciencias, tanto experimentales como sociales, desde la biología o la medicina hasta la sociología o la economía. En ingeniería ha sido ampliamente estudiado para diseñar máquinas capaces de clasificar de forma automática, por ejemplo billetes y monedas, sonidos, imágenes, etc. También en el ámbito financiero tiene aplicación desde hace mucho tiempo para la clasificación de riesgo crediticio donde, a partir de una serie de variables conocidas sobre la persona que solicita un crédito (ingresos, profesión, miembros de la unidad familiar, patrimonio, edad, etc) se aplica para decidir, de manera automatizada, acerca de su concesión. Existen otros muchos ejemplos de aplicación: en el diagnóstico de enfermedades, en procesos de control de calidad para procesos de fabricación o, como en el caso de este trabajo, para asignar un texto a uno de varios autores a partir del análisis de las frecuencias de utilización de las palabras.

En términos generales el planteamiento estadístico de este tipo de problemas es el siguiente: se cuenta con un conjunto de elementos pertenecientes a dos o mas poblaciones distintas sobre los que se ha observado una variable aleatoria p-dimensional, x, y se desea clasificar un nuevo elemento, a partir de sus valores conocidos para la variable x, en una de las poblaciones.

En las situaciones en las que se cuenta a priori con una serie de clases o categorías predefinidas que particionan la población las técnicas para construir un modelo que asigne

los elementos a la clase correspondiente se conocen también como técnicas de *clasificación automática supervisada.* En contraposición, las técnicas en las que el objetivo es construir un modelo para particionar la población en grupos homogéneos pero sin partir de ningún tipo de información previa acerca del número y características de estos, y en las que se aplica la metodología del *análisis cluster* o *análisis de conglomerados, se denominan técnicas* de *clasificación automática no supervisada* .

Entre las técnicas de clasificación supervisada mas utilizadas encontramos algoritmos probabilísticos (entre los que destaca el conocido algoritmo de Naive-Bayes que consiste en estimar la probabilidad de que un documento pertenezca a una categoría en función de la probabilidad de poseer una serie de características conocida para cada uno de los elementos que pertenecen a la categoría en cuestión), el algoritmo del vecino mas próximo, extensivo a los k vecinos mas próximos (en el que se calcula la similitud entre el elemento a clasificar y cada uno de los elementos del conjunto de entrenamiento, asumiendo que la categoría del elemento coincide con la del mas similar entre estos últimos), algoritmos basados en redes neuronales (una de las aplicaciones mas extendidas de las redes neuronales es precisamente el reconocimiento de patrones) o algoritmos basados en árboles de clasificación.

Aunque existen muchos algoritmos de clasificación supervisada la idea básica es la misma en todos ellos: construir un patrón para cada una de las clases que, mediante la aplicación de alguna función, permita estimar el parecido o similitud entre cada elemento a clasificar y los patrones de las categorías. Para construir los patrones se utiliza un conjunto de individuos de los que se conoce previamente su clase y que se denomina conjunto de entrenamiento, conociéndose como *entrenamiento* o aprendizaje el proceso de formación de los patrones de cada clase a partir de estos individuos conocidos.

Los conceptos y técnicas de clasificación utilizados con variables de tipo numérico son aplicables a variables textuales sin mas que tener en cuenta que en este último caso debe realizarse una tarea previa de procesamiento de los datos textuales no estructurados que proporcione una matriz de datos estructurados sobre la que sea posible su aplicación. Por lo tanto, las dos etapas de las técnicas de clasificación: construcción del clasificador y clasificación de nuevos documentos, vendrán precedidas en el caso de la variable textual del preprocesamiento de los datos textuales.

Así, la clasificación supervisada de textos se puede definir como la tarea de aproximar una función de asignación de categoría desconocida $F : D\ x\ C\ \rightarrow\ \{0, 1\}$, donde D es el conjunto de documentos de texto y C es el conjunto de categorías predefinidas. El valor de $F(d, c)$ es 1 si el documento d pertenece a la categoría c mientras que de otra manera el valor es 0.

En este trabajo se aplicará el análisis discriminante clásico a un problema de clasificación de textos para la identificación de los autores. Se pretende obtener un modelo discriminante que clasifique los textos en función de su autor. Para reducir la dimensionalidad de los datos y, al mismo tiempo, contar con un conjunto de variables continuas bajo la hipótesis de normalidad multivariante (idónea para la aplicación del análisis discriminante), se aplicarán técnicas del escalado multidimensional a la matriz de *Documentos* x *Términos*.

3.1. Análisis Discriminante Lineal (LDA).

El enfoque para la resolución del problema de clasificación propuesto por Fisher bajo la denominación de *análisis discriminante* está basado en la normalidad multivariante de los datos considerados y es óptimo bajo este supuesto. En los casos en que todas las variables son continuas, aun cuando los datos originales no estén normalmente distribuidos es posible transformarlos para obtener normalidad y con ello posibilitar la aplicación de esta técnica, pero cuando en el conjunto de variables exista alguna de tipo discreto, la aceptación de la hipótesis de normalidad es poco realista y será preferible aplicar otro tipo de técnicas.

Se denomina *regla de decisión* a cualquier partición del espacio muestral E_x en regiones: $A_1, \ldots, A_n$ tales que $\bigcap_{i \neq j} A_i A_j = 0,\ i, j \in (1, \ldots, n)$ y $E_x = \bigcup_{i=1}^{n} A_i$, y de manera que si $x_0 \in A_i \Rightarrow d_i,\ i = (1, \ldots, n)$, donde d_i representa la decisión de clasificar x_0 en la población P_i

Consideremos ahora una variable x p-variante y absolutamente continua y dos poblaciones, P_1 y P_2 en las que x tiene funciones de densidad conocidas f_1 y f_2, y planteemos el problema de clasificar una nueva observación x_0.

Si conocemos las probabilidades *a priori* de que x_0 provenga de cada una de las poblaciones, $P[x_0 \in P_1] = \pi_1$ y $P[x_0 \in P_2] = \pi_2$, con $\pi_1 + \pi_2 = 1$, entonces $f(x) = \pi_1 f_1(x) + \pi_2 f_2(x)$, y observado x_0 será posible calcular, por el teorema de Bayes, las probabilidades *a posteriori* de que haya sido generado por cada una de las poblaciones:

$$P[P_1|x_0] = \frac{\pi_1 P[x_0|P_1]}{\pi_1 P[x_0|P_1] + \pi_2 P[x_0|P_2]} = \frac{\pi_1 f_1(x_0)}{\pi_1 f_1(x_0) + \pi_2 f_2(x_0)}$$

$$P[P_2|x_0] = \frac{\pi_2 P[x_0|P_2]}{\pi_1 P[x_0|P_1] + \pi_2 P[x_0|P_2]} = \frac{\pi_2 f_2(x_0)}{\pi_1 f_1(x_0) + \pi_2 f_2(x_0)}$$

Entonces, clasificaremos x_0 en la población mas probable *a posteriori*, es decir, clasificaremos x_0 en P_2 si $\pi_2 f_2(x_0) > \pi_1 f_1(x_0)$. Nótese que en el caso en que las probabilidades a priori fueran iguales, $\pi_1 = \pi_2$, la regla de clasificación anterior se reduce a clasificar x_0 en P_2 si $f_2(x_0) > f_1(x_0)$.

En muchas ocasiones, los errores de clasificación tienen distintas consecuencias que se pueden cuantificar, en cuyo caso, planteándolo como un problema bayesiano de decisión, podemos incluir estas consecuencias en la solución. Sea $c(i|j)$ el coste de clasificar en P_i un elemento que pertenece a P_j, con $c(i|i) = 0$ y $c(i|j)$ conocido para todo $i \neq j$. Entonces, se trata ahora de minimizar el coste esperado.

En el caso de dos poblaciones, los costes esperados de las decisiones d_1 y d_2 serán:

$$E(d_1) = c(1|1)P[P_1|x_0] + c(1|2)P[P_2|x_0] = c(1|2)P[P_2|x_0] = c(1|2)\pi_2 f_2(x_0)$$

$$E(d_2) = c(2|1)P[P_1|x_0] + c(2|2)P[P_2|x_0] = c(2|1)P[P_1|x_0] = c(2|1)\pi_1 f_1(x_0)$$

Entonces, clasificaremos x_0 en P_1 si su coste esperado es menor, es decir, si $c(1|2)\pi_2 f_2(x_0) < c(2|1)\pi_1 f_1(x_0)$ o, equivalentemente, si:

$$\frac{\pi_2 f_2(x_0)}{c(2|1)} < \frac{\pi_1 f_1(x_0)}{c(1|2)}$$

y puede comprobarse que este criterio es equivalente minimizar la probabilidad total de error en la clasificación.

3.1.1. Función lineal discriminante.

Supongamos que f_1 y f_2 son distribuciones normales con distintos vectores de medias, μ_1 y μ_2, pero igual matriz de covarianzas, V.

La partición óptima, como acabamos de ver, es la que clasificará los nuevos elementos en P_1 si

$$\frac{\pi_2 f_2(x_0)}{c(2|1)} < \frac{\pi_1 f_1(x_0)}{c(1|2)}$$

Tomando logaritmos y sustituyendo f_i por sus expresiones, esto equivale a:

$$-\frac{1}{2}(x_0-\mu_2)'V^{-1}(x_0-\mu_2)+log\frac{\pi_2}{c(2|1}>-\frac{1}{2}(x_0-\mu_1)'V^{-1}(x_0-\mu_1)+log\frac{\pi_1}{c(1|2}$$

que se puede reescribir a su vez como:

$$D_1^2 - log\frac{\pi_1}{c(1|2} > D_2^2 - log\frac{\pi_2}{c(2|1}$$

siendo D_i^2 la distancia de Mahalanobis entre el elemento x_0 y la media de P_i, μ_i: $D_i^2 = (x_0-\mu_i)'V^{-1}(x_0-\mu_i)$

En el caso de que los costes y las probabilidades *a priori* sean iguales, ($c(1|2) = c(2|1)$ y $\pi_1 = \pi_2$) la regla anterior se reduce a clasificar x_0 en P_1 si $D_1^2 < D_2^2$. Nótese además que si las variables son incorreladas, $V = I\sigma^2$, D_i^2 coincide con la distancia euclídea.

La generalización de lo anterior para G poblaciones es inmediata. Supongamos los costes de clasificación constantes e independientes de la población en que se clasifique la observación, en cuyo caso A_g será la región definida por los puntos con probabilidad máxima de ser generados por P_g, es decir, la región en la que el producto de la probabilidad *a priori* y la verosimilitud sea máximo:

$$A_g = \{x \in E_x | \pi_g f_g(x) > \pi_i f_i(x);\ \forall i \neq g\}$$

Cuando las probabilidades *a priori* son iguales ($\pi_i = \frac{1}{G},\ i = (1,\ldots,G)$ y $f_i(x) \rightsquigarrow N(\mu_i, V)$, la definición anterior equivale a calcular la distancia de Mahalanobis desde el elemento al centro de cada población y clasificarlo en la población que la haga mínima.

Poblaciones desconocidas.

Consideremos la matriz de datos X, n x p, particionada en G matrices correspondientes a las G poblaciones. Entonces los elementos de X serán de la forma x_{ijg}, donde i representa la fila, j la columna y g la submatriz. Si n_g es el número de elementos que pertenecen a la submatriz g, entonces $n = \sum_{g=1}^{G} n_g$, y si $x'_{ig} = (x_{i1g}, \ldots, x_{ipg})$, entonces el vector de medias de cada población y la correspondiente matriz de varianzas y covarianzas serán

$$\bar{x}_g = \frac{1}{n_g} \sum_{i=1}^{n_g} x_{ig}$$

$$\hat{S}_g = \frac{1}{n_g - 1} \sum_{i=1}^{n_g} (x_{ig} - \bar{x}_g)(x_{ig} - \bar{x}_g)'$$

Si suponemos que todas las poblaciones tienen la misma matriz de varianzas y covarianzas, su mejor estimación centrada con todos los datos será una combinación lineal de las estimaciones centradas de las matrices de varianzas y covarianzas en cada población con peso proporcional a su precisión:

$$\hat{S}_w = \sum_{g=1}^{G} \frac{n_g - 1}{n - G} \hat{S}_g$$

Llamaremos W a la matriz de sumas de cuadrados dentro de las clases

$$W = \sum_{g=1}^{G} \sum_{i=1}^{n_g} (x_{ig} - \bar{x}_g)(x_{ig} - \bar{x}_g)' = (n - G)\hat{S}_w$$

Ahora, para obtener las funciones discriminantes se utiliza $\bar{x}_g$ como estimador de μ_g y $\hat{S}_w$ como estimación de V. Así, suponiendo iguales las probabilidades *a priori* y los costes de clasificación ($c(i|j) = k,\ \pi_i = 1/G,\ \forall i \in \{1, \ldots, G\},\ i \neq j$) y llamando $\hat{w}_g = \hat{S}_w^{-1} \bar{x}_g$ se clasificará x_0 en la población g que verifique:

$$min[(x_0 - \bar{x}_g)' \hat{S}_w^{-1} (x_0 - \bar{x}_g)] = min[w'_g(\bar{x}_g - x_0)]$$

lo que equivale a calcular las variables indicadoras

$$z_{g,g+1} = \hat{w}'_{g,g+1}x_0,\ g = 1, \ldots, G-1$$

donde $\hat{w}_{g,g+1} = \hat{S}_w^{-1}(\bar{x}_g - \bar{x}_{g+1}) = \hat{w}_g - \hat{w}_{g+1}$ y la regla de decisión será clasificar en g frente a $g+1$ si

$$|z_{g,g+1} - \hat{m}_g| < |z_{g,g+1} - \hat{m}_{g+1}|$$

siendo $\hat{m}_g = \hat{w}'_{g,g+1}\bar{x}_g$

3.1.2. Probabilidades de error y validación cruzada.

El método mas inmediato para aproximar el error asociado a una regla de clasificación consiste en aplicar la función discriminante a los n elementos observados para obtener su clasificación, llamando n_{ij} al número de elementos de P_i clasificados en P_j, de forma que el error aparente se calculará como el cociente entre el número de elementos mal clasificados y el número de elementos clasificados correctamente, es decir:

$$\epsilon = \frac{\sum_{i \neq j}^{G} n_{ij}}{\sum_{i=1} n_{ii}}$$

El método anterior tiende a subestimar las probabilidades de error al emplear los mismos datos en la estimación de los parámetros y en la evaluación de la regla de clasificación resultante. Por este motivo es preferible aplicar la técnica denominada *validación cruzada* consistente en clasificar cada elemento con una regla construida prescindiendo de dicho elemento. Así, se construyen n funciones discriminantes con los n conjuntos de $n-1$ elementos cada uno resultantes de eliminar uno a uno cada elemento del conjunto original, para a continuación clasificar cada uno de los elementos con la regla construida sin él.

Cuando el número de elementos es muy elevado, el método de validación cruzada es computacionalmente muy costoso y se suele dividir el conjunto de datos en k subconjuntos de igual tamaño con los que se aplica la *validación cruzada* pero prescindiendo de cada uno de los subconjuntos para obtener las reglas de clasificación.

3.2. Discriminación cuadrática (QDA).

Si aun presumiendo la normalidad de las variables observadas no fuera posible admitir la hipótesis de igualdad de varianzas, el problema se resolverá clasificando cada nueva observación en la población con máxima probabilidad *a posteriori*, es decir, clasificar x_0 en la población que minimice la siguiente función

$$\frac{1}{2}log|V_g| + \frac{1}{2}(x_0 - \mu_g)'V_g^{-1}(x_0 - \mu_g) - log(C_g\pi_g), \quad g = 1, \ldots, G$$

Si V_g y μ_g no son conocidas se estimarán de la forma habitual mediante S_g y $\bar{x}_g$.

En este caso el término $x_0'V_g^{-1}x_0$ no se puede anular puesto que depende de la población, y las funciones discriminantes no serán lineales al contar con un término de segundo grado.

El número de parámetros a estimar en este caso cuadrático se mucho mayor que en el lineal lo que hace que, excepto en el caso de muestras muy grandes, la discriminación cuadrática sea bastante inestable y, aun con matrices de covarianzas muy diferentes, es frecuente obtener mejores resultados con la clasificación lineal. Un problema adicional es la extremada sensibilidad de la función discriminante cuadrática a desviaciones de la normalidad de las variables observadas.

En general, la evidencia indica que la discriminación lineal es mas robusta que la discriminación cuadrática.

3.3. Métricas de evaluación.

Existen diversas medidas para evaluar la "bondad" de los modelos de clasificación obtenidos que, en general, están basadas en el concepto de *matriz de confusión.*

Esta matriz es una tabla:

	C_1'	$\ldots$	C_p'	
C_1	n_{11}	$\ldots$	n_{1p}	$n_{1.}$
$\vdots$		$\ddots$		$\vdots$
C_p	n_{p1}	$\ldots$	n_{pp}	$n_{p.}$
	$n_{.1}$	$\ldots$	$n_{.p}$	

Las filas, C_i, representan las clases o categorías reales y las columnas, C'_i, las clases o categorías predichas por el modelo. Los valores n_{ij} corresponden con el número de elementos de la clase real i que han sido predichos por el modelo en la clase j.

Por lo tanto el tamaño del conjunto de test es igual a la suma de todos los elementos de la matriz es, $n_{..} = \sum_{i=1}^{p} \sum_{j=1}^{p} n_{ij}$, y el número total de elementos bien clasificados vendrá determinado por la suma de la diagonal principal,$\sum_{i=1}^{p} n_{ii}$.

3.3.1. Exactitud, precisión y recuperación.

La proporción de elementos clasificados correctamente por el modelo corresponderá entonces con el cociente entre la suma de la diagonal principal y la suma de todos los elementos de la matriz de confusión. Este valor, en forma de porcentaje, se conoce como *exactitud* (o *accuracy*) y se utiliza como medida global del "acierto" del modelo pues representa la proporción total de elementos de todas las categorías clasificados correctamente.

$$Exactitud = 100\frac{\sum_{1=1}^{p} n_{ii}}{\sum_{i=1}^{p} \sum_{j=1}^{p} n_{ij}} = 100\frac{\sum_{1=1}^{p} n_{ii}}{n_{..}}$$

Además se definen algunas otras medidas individuales para evaluar el comportamiento del modelo respecto de cada una de las categorías. Las mas habituales son la *precisión* y la *recuperación* (o *recall*).

La *precisión* representa la proporción de elementos de una categoría que el modelo clasifica correctamente y se calcula como el cociente de cada elemento de la diagonal principal entre la suma de la fila a la que pertence.

$$Precisión_i = 100\frac{n_{ii}}{\sum_{j=1}^{p} n_{ij}} = \frac{n_{ii}}{n_{i.}}$$

La *recuperación* representa la proporción de los elementos clasificado por el modelo en una determinada categoría que realmente pertenecen a la misma y se obtiene calculando el cociente entre cada elemento de la diagonal principal y la suma de la columna a la que pertenece.

$$Recuperación_i = 100\frac{n_{jj}}{\sum_{i=1}^{p} n_{ij}} = 100\frac{n_{jj}}{n_{.j}}$$

3.3.2. Adjusted Rand Index (ARI)

Se trata de una medida definida en el área del análisis cluster para comparar dos particiones de entre el conjunto finito de particiones posibles de una población determinada. Fue propuesta por Hubert y Arabie en 1985 a partir del índice de Rand, una de las medidas mas populares para comparar particiones en aquel momento.

Supongamos una población de n objetos, $S = \{o_1, ..., o_n\}$ y dos particiones de S, $U = \{u_1, ..., u_p\}$ y $V = \{v_1, ..., v_q\}$. En el análisis clúster "clásico" el número de conjuntos que forman cada una de las particiones puede no coincidir ($p \neq q$) pero en la utilización de estos indices para evaluar modelos de clasificación supervisada este número representa el de las clases y enfrentaremos la partición correspondiente a las clases reales de los elementos del conjunto de test frente a las clases predichas por el modelo, por lo que siempre se verificará $p = q$.

Consideramos ahora todos los posibles pares de objetos de S, (o_i, o_j), $i = 1, \ldots, n$; $i \neq j$ y definimos:

- a: número de pares (o_i, o_j) que se encuentran en el mismo subconjunto de U y en el mismo subconjunto de V, es decir, $o_i, o_j \in U_h$; $o_i, o_j \in V_k$
- b: número de pares (o_i, o_j) que se encuentran en distintos subconjuntos de U y en distintos subconjuntos de V, es decir, $o_i \in U_h, o_j \in U_k \neq U_h$; $o_i \in V_m, o_j \in V_n \neq V_m$
- c: número de pares (o_i, o_j) que se encuentran en el mismo subconjunto de U y en distintos subconjuntos de V, es decir, $o_i, o_j \in U_h$; $o_i \in V_m, o_j \in V_n \neq V_m$
- d: número de pares (o_i, o_j) que se encuentran en distintos subconjuntos de U y en el mismo subconjunto de V, es decir, $o_i \in U_h, o_j \in U_k \neq U_h$; $o_i, o_j \in V_k$

De forma que diremos que se produce un *acuerdo* entre U y V cuando el par de objetos (o_i, o_j) se encuentra en los casos a o b, es decir, si ambas particiones incluyen a los dos objetos en un mismo subconjunto o ambas particiones incluyen a cada uno de los objetos en distintos subconjuntos.

Al contrario diremos que se produce un *desacuerdo* entre U y V cuando el par de objetos (o_i, o_j) se encuentra en los casos c o d, es decir, si una partición incluye a los dos objetos en un mismo subconjunto y la otra partición incluye a cada uno de los objetos en distintos subconjuntos.

La suma de posibles *acuerdos* y *desacuerdos* $(a + b + c + d)$ será entonces el número total de pares posibles, $\binom{n}{2}$.

El índice de Rand representa la proporción de acuerdos entre U y V:

$$RI = \frac{a+b}{a+b+c+d} = \frac{a+b}{\binom{n}{2}}$$

El *índice de Rand ajustado* propuesto por Hubert y Arabie es la versión corregida por el azar del índice de Rand. Su expresión a partir de una matriz de confusión es la siguiente:

$$ARI = \frac{\sum_{i=1}^{p}\sum_{j=1}^{q}\binom{n_{ij}}{2} - \frac{\sum_{i=1}^{p}\binom{n_{i.}}{2}\sum_{j=1}^{q}\binom{n_{.j}}{2}}{\binom{n_{..}}{2}}}{\frac{\sum_{i=1}^{p}\binom{n_{i.}}{2}+\sum_{j=1}^{q}\binom{n_{.j}}{2}}{2} - \frac{\sum_{i=1}^{p}\binom{n_{i.}}{2}\sum_{j=1}^{q}\binom{n_{.j}}{2}}{\binom{n_{..}}{2}}}$$

En los problemas de clasificación supervisada, en los que se utiliza este índice para comparar la "partición" obtenida (clases predichas) con la "partición" perfecta (clases reales), ARI=0 cuando la clasificación obtenida se corresponde con la esperada para un modelo puramente aleatorio, mientras que ARI=1 cuando la clasificación del modelo es perfecta, es decir coincide con la clasificación real del conjunto de test. Se da la circunstancia de que este índice puede tomar valores negativos cuando la clasificación obtenida es peor que la esperada para un modelo aleatorio puro.

Capítulo 4

Análisis del corpus *Reuters Corpus Volume I.*

La aplicación de los métodos y técnicas estadísticas presentadas en este trabajo se reailizará sobre un subconjunto del *dataset Reuters Corpus Volume I* (Lewis, Yang, Rose, & and Li, 2004). Este conjunto de datos contiene más de 800.000 noticias periodísticas de la agencia Reuters clasificadas por autores. El análisis exploratorio se realizará sobre 100 noticias de cada uno de los 4 autores siguientes:

```
##
##  AlexanderSmith BenjaminKangLim     BradDorfman     DavidLawder
##             100             100             100             100
```

Se realiza el siguiente preprocesamiento del corpus:

- Convertir todas las palabras a minúsculas
- Eliminar los números
- Eliminar espacios en blanco innecesarios
- Eliminar los signos de puntuación
- Eliminar las *stopWords* del idioma inglés
- Eliminar las palabras de 1 y 2 caracteres
- *Lematizar* todas las palabras resultantes

4.1. Matriz *Documentos* x *Palabras*.

Con este corpus transformado se construye la matriz de Documentos x Palabras, obteniendo así una primera aproximación a la información de interés, estadísticamente hablando, de los textos objeto de análisis.

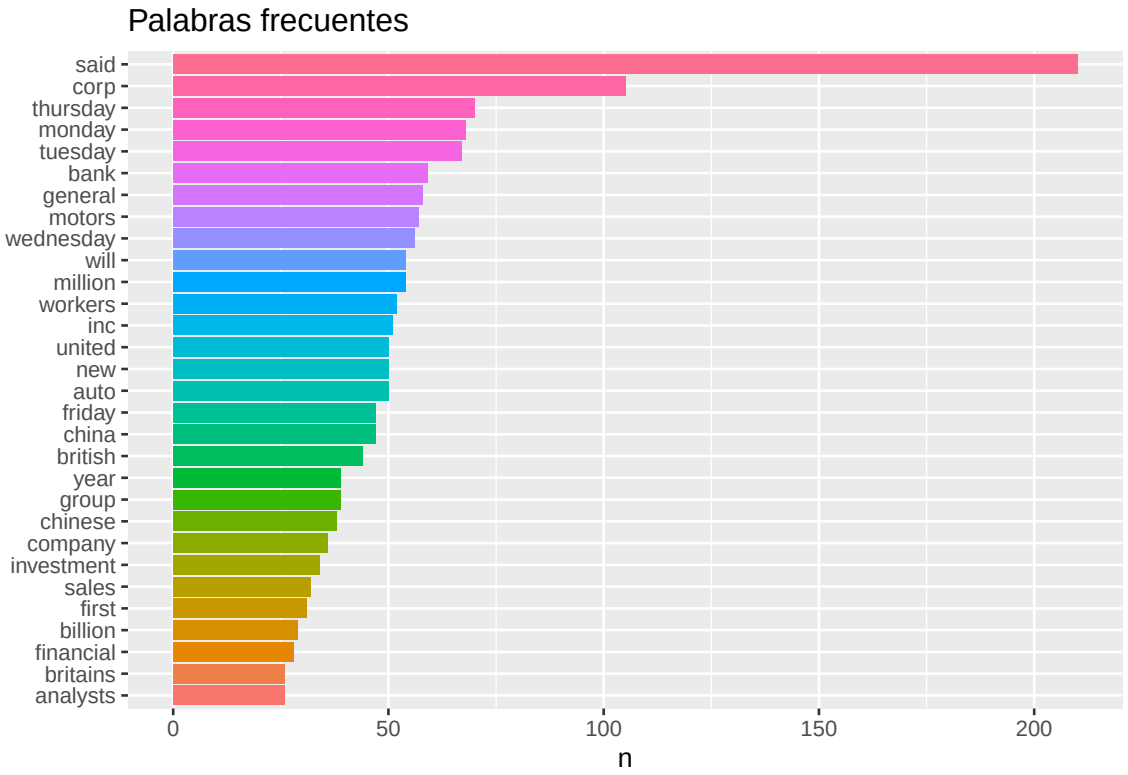
```
## <<DocumentTermMatrix (documents: 400, terms: 1990)>>
## Non-/sparse entries: 7781/788219
## Sparsity           : 99%
## Maximal term length: 21
## Weighting          : term frequency (tf)
```

La simple transformación en una matriz numérica proporciona una primera información sobre el corpus objeto de estudio. Esta matriz está compuesta por 400 filas (que corresponden a las noticias que forman el corpus) y 1990 columnas (que corresponden a las palabras conservadas tras el preprocesamiento). Se trata de una matriz muy dispersa en la que el 99 % de los elementos son ceros (788219 elementos iguales a cero frente a 7781 elementos mayores que cero) y la palabra de mayor longitud tiene 21 caracteres.

4.2. Análisis exploratorio

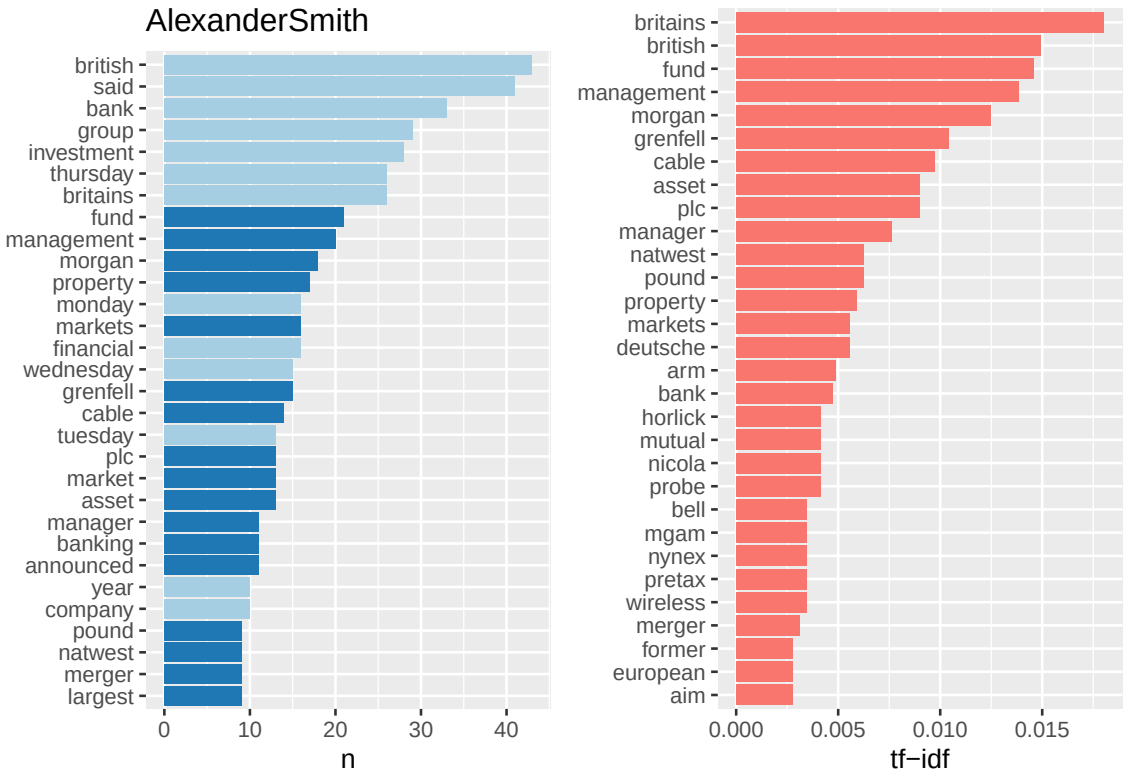
En primer lugar es habitual identificar las palabras más frecuentes a nivel de *corpus* y, en su caso, a nivel de cada una de las categorías que clasifican los documentos del corpus (en nuestro caso los autores) así como aplicar la función tfIdf a fin de identificar los términos más representativos de cada categoría.

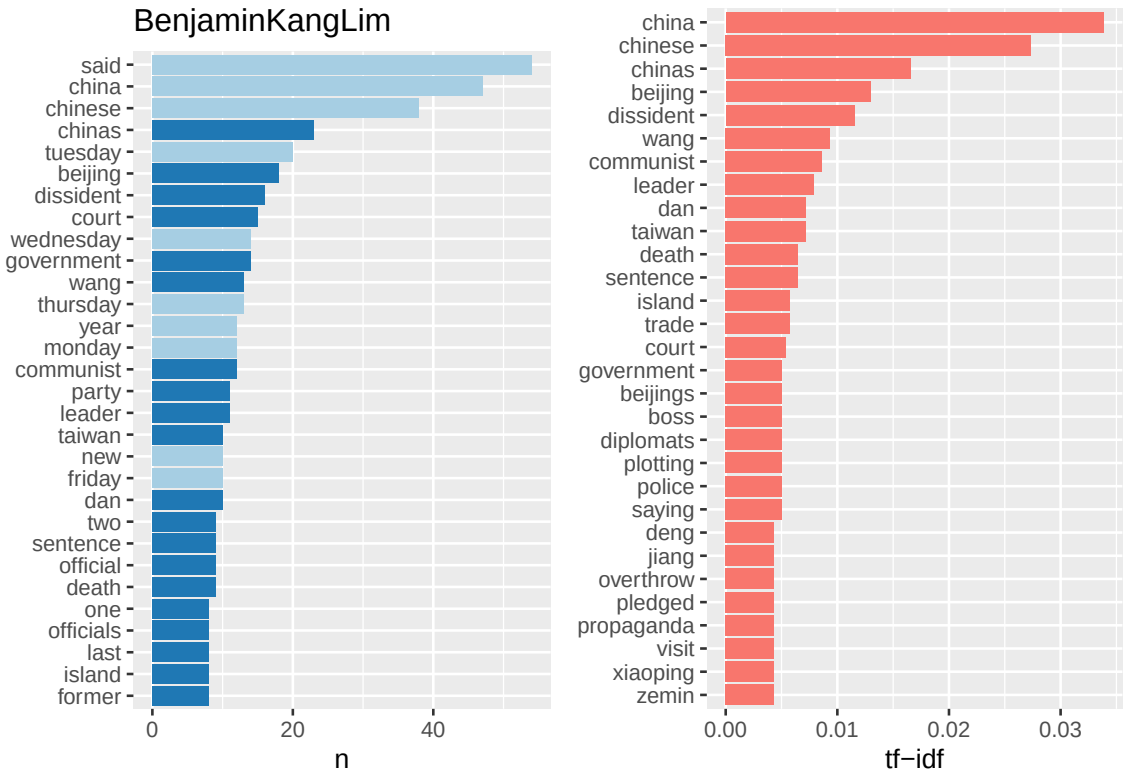
4.2.1. Palabras frecuentes, *tfIdf* y distribución del vocabulario.

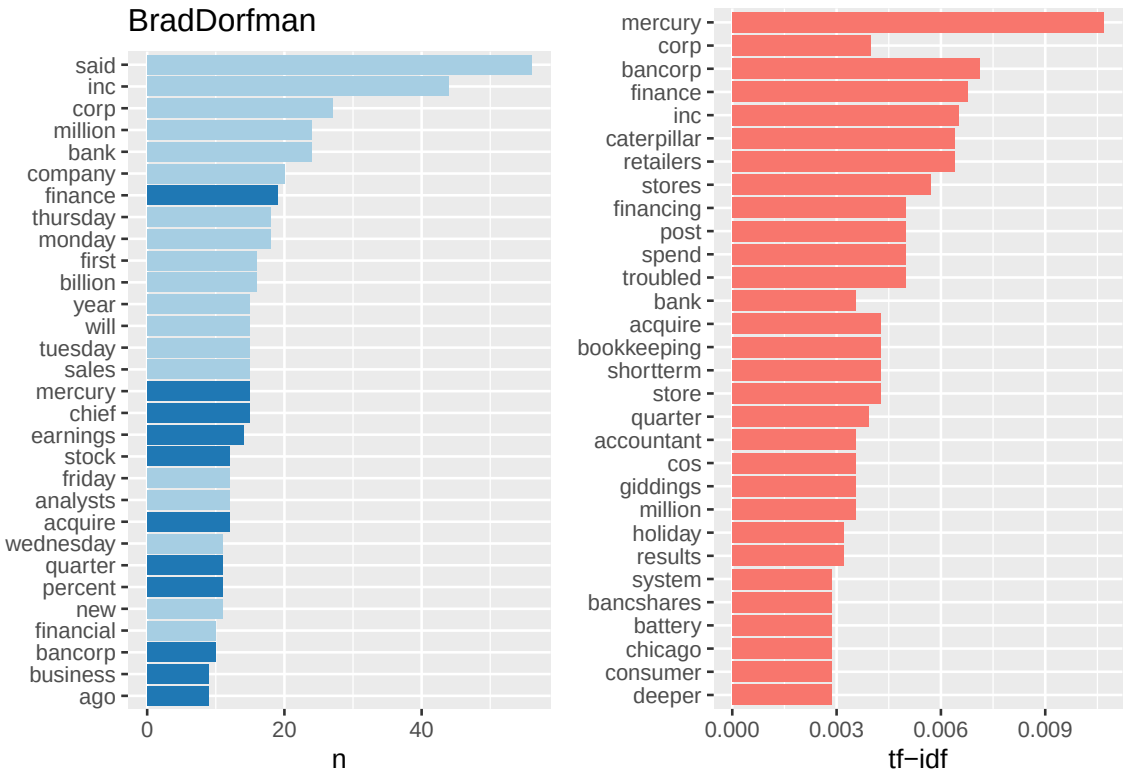


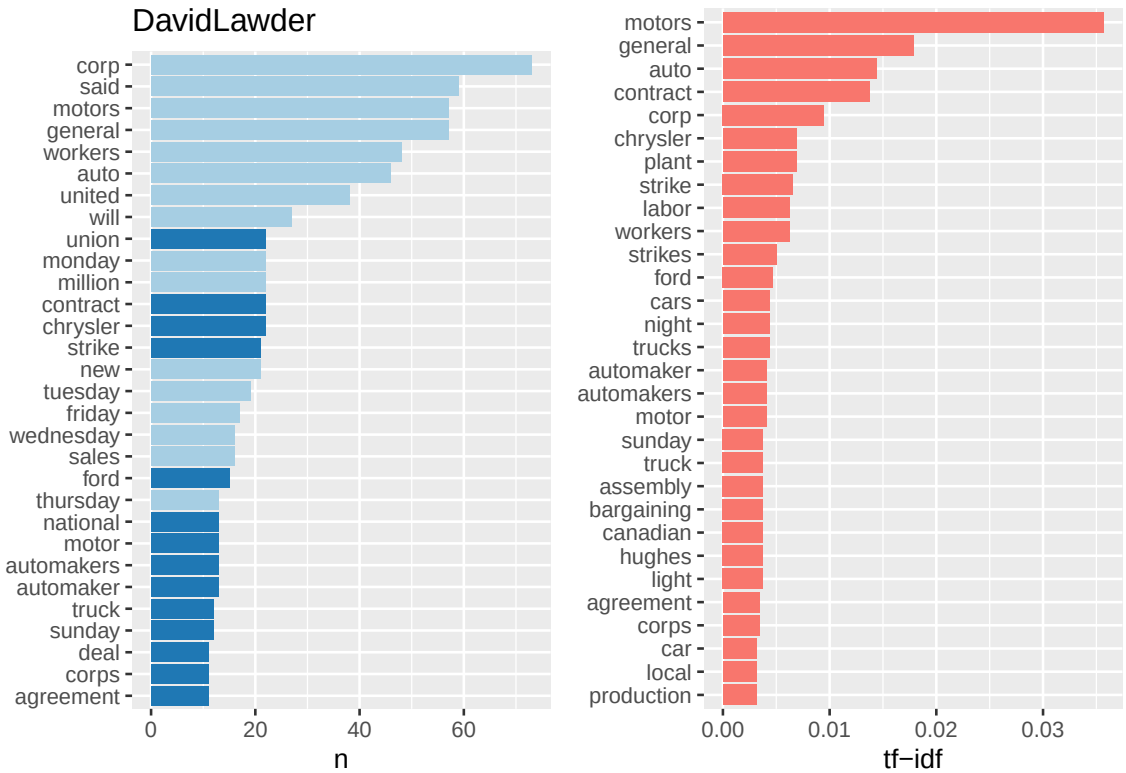
Este gráfico nos permite detectar las temáticas predominantes en las noticias que forman el corpus: temas empresariales, financieros, banca, sector automovilístico, china, etc.

Observar las palabras más frecuentes de cada autor facilitará la identificación de posibles diferencias entre ellos en cuanto al vocabulario que emplean o los temas que subyacen en sus documentos. Por este motivo estudiaremos a continuación, para cada uno de los cuatro autores, las palabras más frecuentes y las que mejor los caracterizan (las de mayor valor tfIdf). En los gráficos de palabras más frecuentes se muestran en una tonalidad más clara aquellas que también se encuentran presentes entre las más frecuentes a nivel de corpus, frente a una tonalidad más oscura para las palabras más frecuentes en cada autor que no encuentran correspondencia a este nivel.


AlexanderSmith
british
said
bank
group
investment
thursday
britains
fund
management
morgan
property
monday
markets
financial
wednesday
grenfell
cable
tuesday
plc
market
asset
manager
banking
announced
year
company
pound
natwest
merger
largest
0
10
20
30
40
n
britains
british
fund
management
morgan
grenfell
cable
asset
plc
manager
natwest
pound
property
markets
deutsche
arm
bank
horlick
mutual
nicola
probe
bell
mgam
nynex
pretax
wireless
merger
former
european
aim
0.000
0.005
0.010
0.015
tf−idf


BenjaminKangLim
said
china
chinese
chinas
tuesday
beijing
dissident
court
wednesday
government
wang
thursday
year
monday
communist
party
leader
taiwan
new
friday
dan
two
sentence
official
death
one
officials
last
island
former
0
20
40
n
china
chinese
chinas
beijing
dissident
wang
communist
leader
dan
taiwan
death
sentence
island
trade
court
government
beijings
boss
diplomats
plotting
police
saying
deng
jiang
overthrow
pledged
propaganda
visit
xiaoping
zemin
0.00
0.01
0.02
0.03
tf−idf

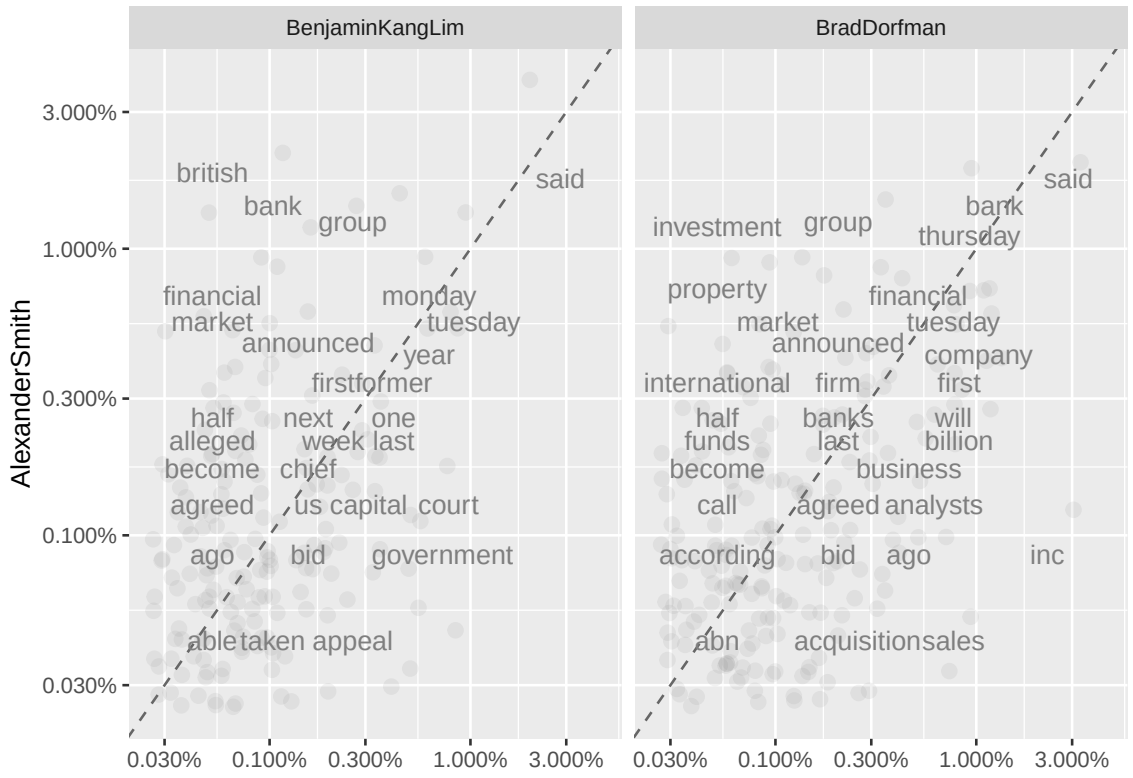

BradDorfman
said
inc
corp
million
bank
company
finance
thursday
monday
first
billion
year
will
tuesday
sales
mercury
chief
earnings
stock
friday
analysts
acquire
wednesday
quarter
percent
new
financial
bancorp
business
ago
0
20
40
n
mercury
corp
bancorp
finance
inc
caterpillar
retailers
stores
financing
post
spend
troubled
bank
acquire
bookkeeping
shortterm
store
quarter
accountant
cos
giddings
million
holiday
results
system
bancshares
battery
chicago
consumer
deeper
0.000
0.003
0.006
0.009
tf−idf


DavidLawder
corp
said
motors
general
workers
auto
united
will
union
monday
million
contract
chrysler
strike
new
tuesday
friday
wednesday
sales
ford
thursday
national
motor
automakers
automaker
truck
sunday
deal
corps
agreement
0
20
40
60
n
motors
general
auto
contract
corp
chrysler
plant
strike
labor
workers
strikes
ford
cars
night
trucks
automaker
automakers
motor
sunday
truck
assembly
bargaining
canadian
hughes
light
agreement
corps
car
local
production
0.00
0.01
0.02
0.03
tf−idf

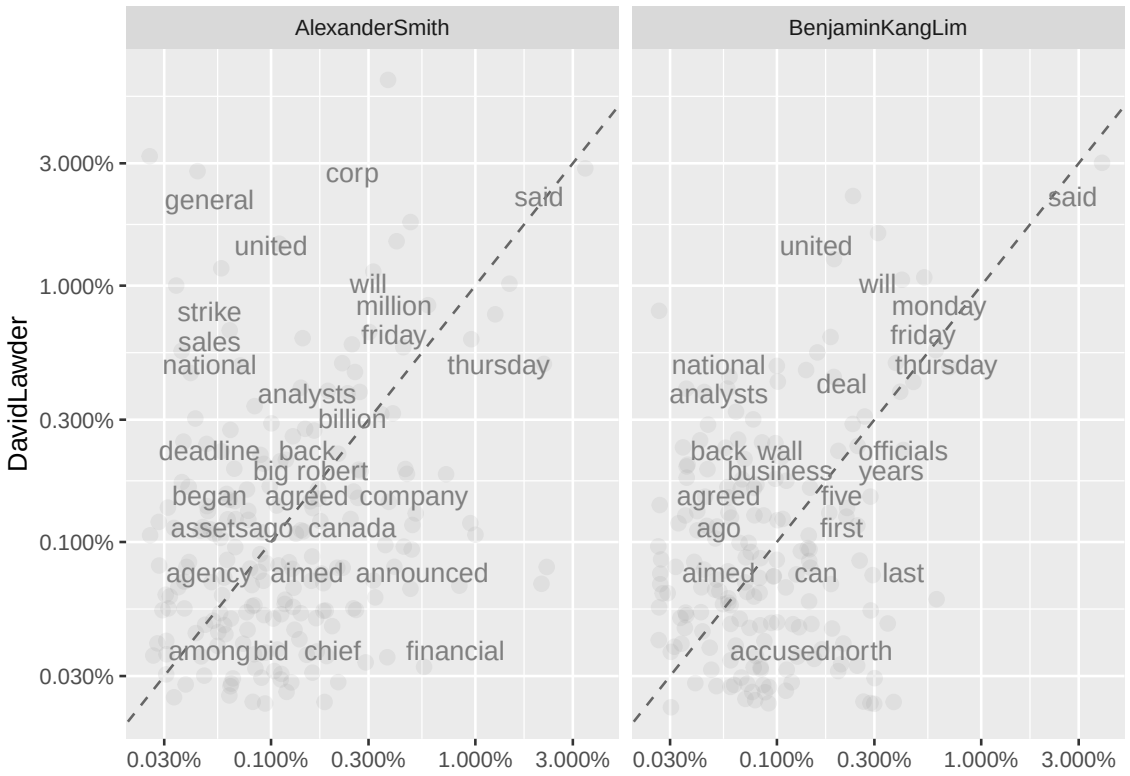
La observación detallada de los gráficos anteriores permite profundizar en el vocabulario empleado y en la temática propia de cada autor. Así, en Alexander Smith destaca la palabra *british*, siendo además la segunda con mayor valor tfIdf (lo que indica que es un término, en comparación, muy poco empleado por el resto de autores); en Benjamin Kang Lim predomina un vocabulario muy relacionado con China y la política de este país; en Brad Dorfman se observa una cierta preferencia por términos relacionados con temática bancaria y financiera así como la palabra *chicago* entre las de valores tfIdf altos; y, por último, las palabras que más destacan para David Lawder tienen que ver con la industria automovilística y, entre las palabras de valores tfIdf altos, encontramos *canadian*.

El análisis de la frecuencia de las palabras tanto a nivel del corpus completo como a nivel de cada categoría aporta mucha información sobre el vocabulario. Esta información se completa y enriquece con la observación de las palabras que arrojan valores altos de tfIdf. Un ejemplo de ello es el término *british*. Se trata de una palabra muy frecuente a nivel de corpus (en la posición 19 de las más frecuentes) lo que podría llevar a pensar que es un término transversal utilizado frecuentemente por varios autores. Al individualizar el análisis de las frecuencias para los autores resulta ser la palabra más utilizada por *Alexander Smith* y, observando el gráfico del valor tfIdf para este autor, resulta ser la segunda palabra con mayor valor en este índice (estando además la primera, *britains*, muy relacionada), lo que implica que, comparativamente, es mucho más utilizada por Alexander Smith que por el resto de autores. Además, se da la circunstancia de que esta palabra no aparece entre las 30 más frecuentes ni con mayores valores tfIdf para ningún otro autor, por lo que, en definitiva, aun siendo una palabra muy frecuente a nivel de corpus, puede resultar muy indicativa para identificar los textos de *Alexander Smith*.

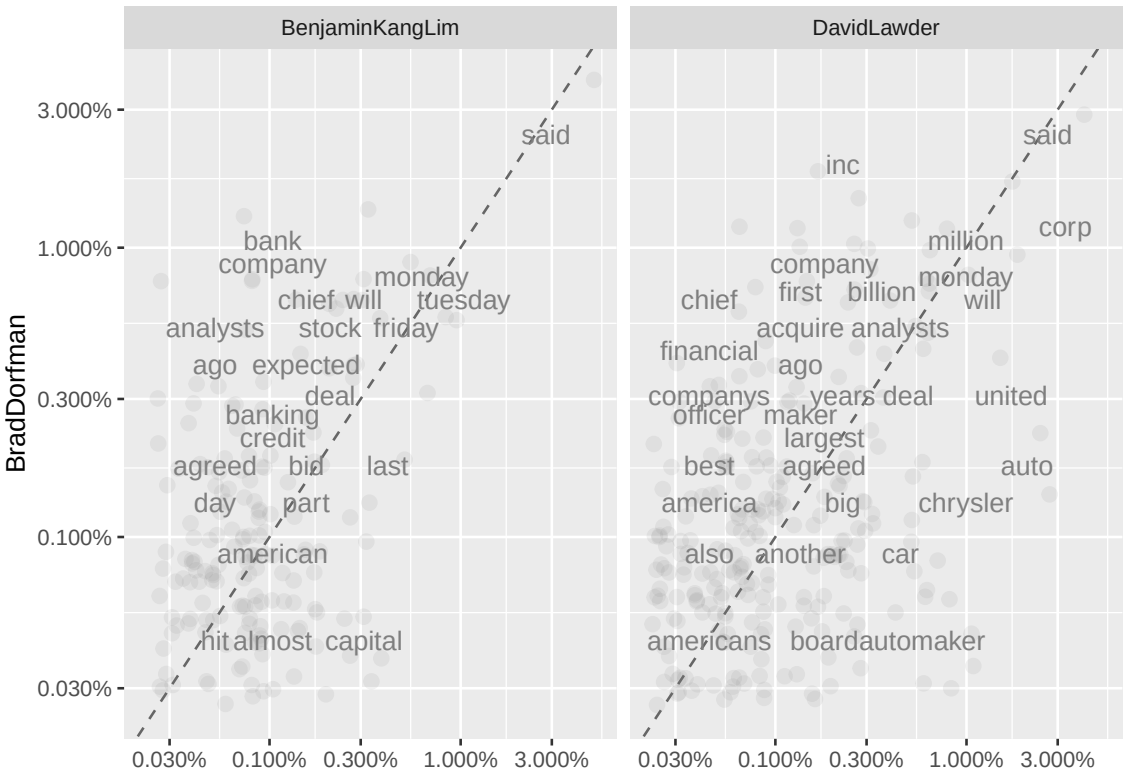
Tras esta aproximación inicial tenemos una primera idea acerca de la temática propia de las noticias de cada uno de los autores así como del vocabulario más representativo de cada uno. También puede resultar interesante realizar una comparativa de la distribución del vocabulario compartido por distintos autores, es decir, analizar las diferencias y similitudes en las frecuencias de uso de las palabras que aparecen en varios autores como se muestra a continuación.



Estas representaciones pueden aportar información adicional interesante. Por ejemplo, en el primer gráfico observamos que la palabra *government* es utilizada por Alexander Smith con una frecuencia que no alcanza el 0,1 % mientras que en las noticias de Benjamin Kang Lim aparece con una frecuencia que supera el 0.3 %, más de tres veces superior, mientras que la frecuencia con que Alexander Smith emplea la palabra *british* es muy superior a la frecuencia con la que la utiliza Benjamin Kang Lim. Por otra parte las palabras *chief*, *weed*, *monday* o *tuesday* se encuentran prácticamente sobre la diagonal lo que indica que son empleadas con frecuencias muy similares por ambos autores. En el segundo gráfico destaca como Brad Dorfman emplea la palabra *inc* con una frecuencia muy superior a Alexander Smith, mientras que se da el caso contrario respecto de *investment*, *group* o *property*.



En la parte superior izquierda del primer gráfico observamos las palabras *general*, *corp* y *united* lo que indica que son mucho más empleadas por Alexander Smith que por David Lawder mientras las palabras *financial* y *announced* son más empleadas por este último. Respecto de la comparativa entre Benjamin Kang Lim y David Lawder este último emplea las palabras *united*, *national* y *analysts* con mayor frecuencia que el primero, dándose la situación inversa con *north*, *last* o *accused*.



Por último observamos que Benjamin Kang Lim emplea *capital* con mayor frecuencia que Brad Dorfman, mientras que este último emplea más frecuentemente *bank*, *company* o *analysts* y, por otra parte, Brad Dorfman hace un uso más frecuente de *inc* que David Lawder que lo supera en cuanto al empleo de *automaker*, *car*, *united*, *auto* o *chrysler*.

En definitiva, hemos comprobado que el vocabulario más empleado en el conjunto del corpus difiere sustancialmente del vocabulario más frecuente restringido a cada uno de los autores, es decir, cada autor utiliza un conjunto de palabras más frecuentes diferente, en térmminos generales, al empleado por los demás autores. Además hemos visto como el empleo de la frecuencia de aparición de los términos ponderada por la frecuencia inversa de los documentos (aplicando la función *TfIdf* que minora considerablemente las frecuencias asociadas a los términos que aparecen con una frecuencia elevada de manera transversal en todo el corpus) facilita la identificación de los términos que, teniendo o no frecuencias elevadas, son representativos de cada uno de los autores.

4.2.2. Análisis de correspondencias.

El análisis de las frecuencias, los valores tfIdf y la comparativa de las frecuencias de utilización del vocabulario nos ha permitido detectar las palabras más relevantes, tanto a nivel de corpus como para cada uno de los autores, así como la presencia de una temática subyacente en cada autor. El siguiente paso, para profundizar en el análisis de las diferencias entre autores, será acometer el estudio de los documentos agrupados por autor para lo que construiremos la correspondiente tabla léxica agregada que da lugar, una vez aplicadas las tareas de preprocesado, a la siguiente matriz de *Categorías* x *Palabras*:

```
## <<DocumentTermMatrix (documents: 4, terms: 1990)>>
## Non-/sparse entries: 2914/5046
## Sparsity           : 63%
## Maximal term length: 21
## Weighting          : term frequency (tf)
```

El análisis de correspondencias de esta matriz nos permitirá la exploración y visualización de la relación entre sus filas y columnas, es decir, entre los autores y el vocabulario. La siguiente salida proporciona los valores propios y porcentajes de inercia calculados al aplicar el análisis de correspondencias a esta matriz.

```
##       eigenvalue percentage of variance cumulative percentage of variance
## dim 1  0.6372633              38.58889                           38.58889
## dim 2  0.5583585              33.81089                           72.39978
## dim 3  0.4557946              27.60022                          100.00000
```

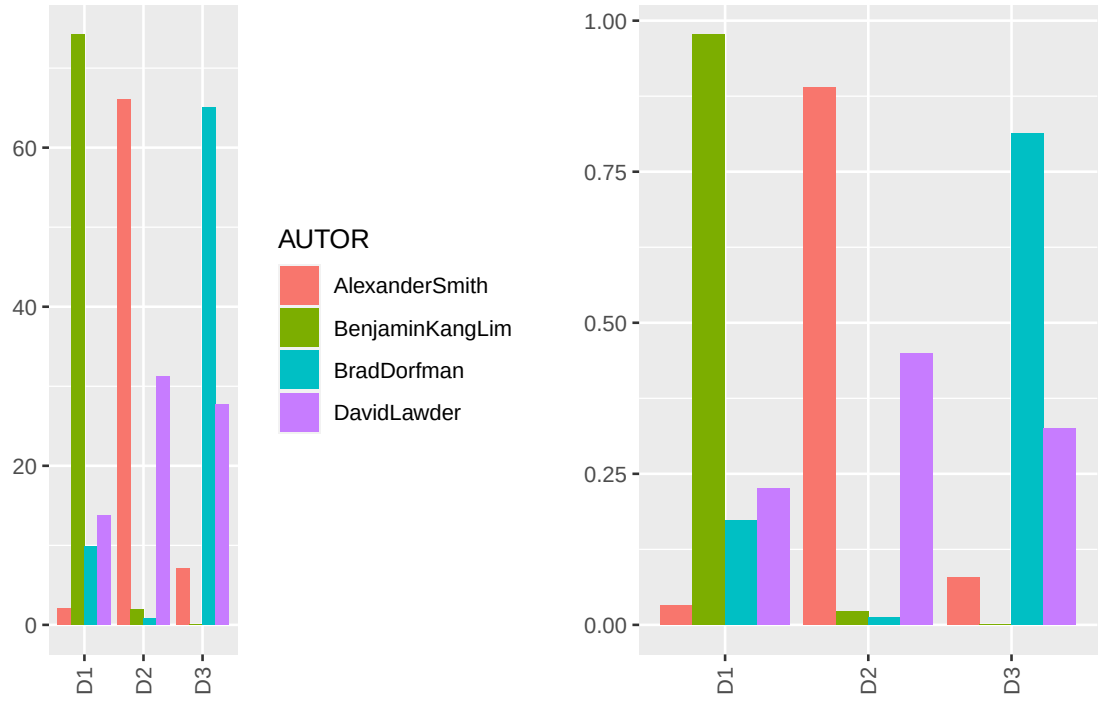
La segunda columna indica el porcentaje de la inercia que conserva cada uno de los ejes obtenidos. El primer eje conserva un 38.59 % de la inercia total, el segundo conserva el 33.81 % y el tercero el 27.6 %. Observamos un moderado predominio del primer eje, que acumula un porcentaje de variabilidad superior al que le correspondería proporcionalmente en detrimento del tercer eje.

A continuación analizaremos la calidad de la representación de las filas (autores) así como su contribución en la construcción de cada uno de los ejes:

```
## % CONTRIBUCIÓN A LA CONSTRUCCIÓN DE LOS EJES

##                      Dim 1     Dim 2       Dim 3
## AlexanderSmith    2.087152 66.023678  7.11512102
## BenjaminKangLim 74.264525  1.894197  0.05769461
## BradDorfman      9.928563  0.845777 65.10779492
## DavidLawder     13.719761 31.236348 27.71938946

## CALIDAD DE LA REPRESENTACIÓN EN CADA EJE (cos^2)

##                       Dim 1      Dim 2       Dim 3
## AlexanderSmith  0.03209773 0.88963993 0.078262342
## BenjaminKangLim 0.97760921 0.02184758 0.000543212
## BradDorfman     0.17346361 0.01294709 0.813589299
## DavidLawder     0.22523008 0.44929784 0.325472078
```

Benjamin Kang Lim es el autor que contribuye mayoritariamente a la construcción del primer eje con un 74,26 % siendo este el eje bajo el que está mejor representado ($cos^2 = 0.98$), *Brad Dorfman* es el autor que contribuye mayoritariamente a la construcción del tercer eje con un 65,11 % y *Alexander Smith* es el que aporta más en la construcción del segundo eje con un 66,02 %, mientras que *David Lawder* es el autor que contribuye en segunda posición a la construcción del primer eje, aunque tan solo en un 13,72 %, pero tiene una aportación importante en la construcción de los ejes segundo y tercero ocupando en ambos la segunda posición con un 31,24 % y un 27,72 % respectivamente.

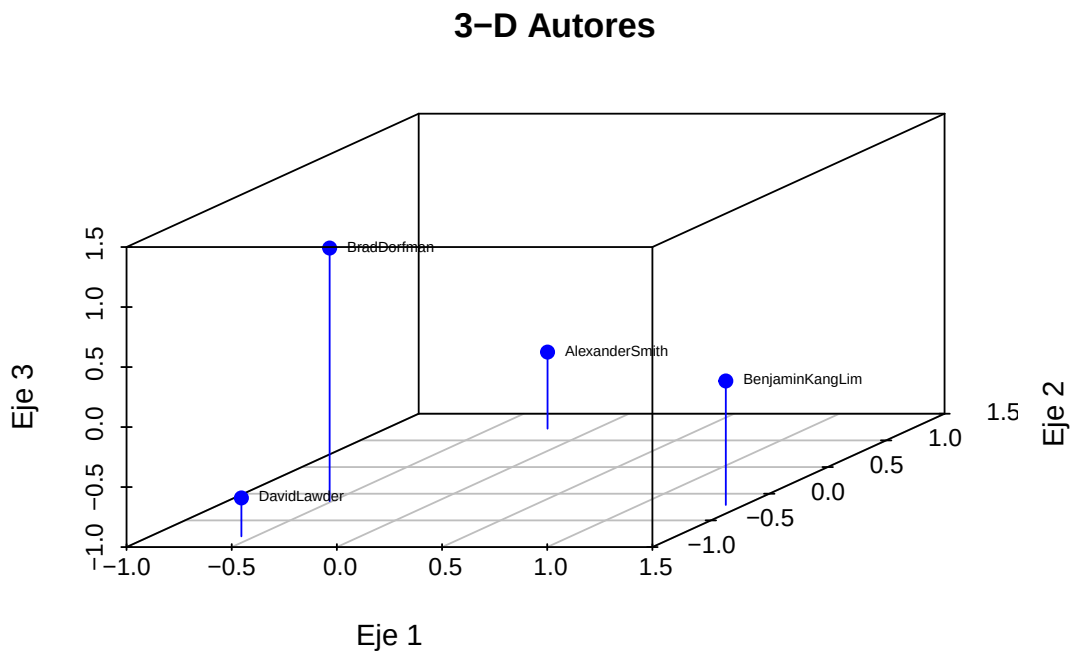
A continuación se presenta una representación gráfica de estos resultados:

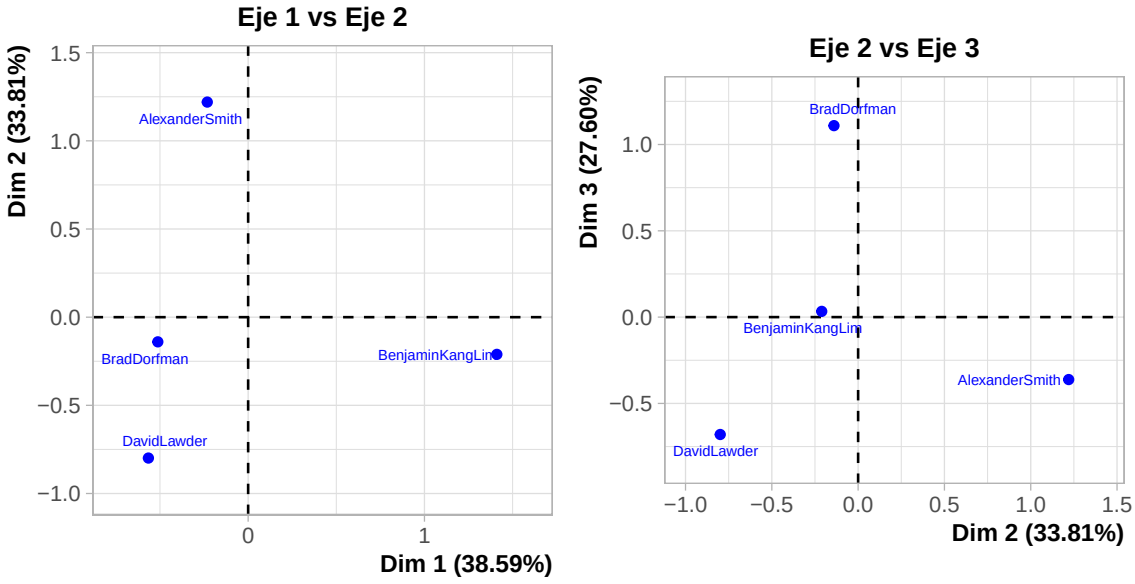


En otras palabras, el primer eje se ha construido mayoritariamente con la información de *Benjamin Kang Lim*, el segundo eje con la información de *Alexander Smith* y, en menor medida, de *David Lawder*, y el tercer eje principalmente con la información de *Brad Dorfman* y, en menor medida, de *David Lawder*. La mejor representación de *Alexander Smith* y de *David Lawder* la proporciona el segundo eje ($cos^2 = 0{,}89$ y $0{,}45$ respectivamente) mientras que la mejor representación de *Brad Dorfman* se obtiene con el tercer eje ($cos^2 = 0.81$).

Por tanto, el gráfico bidimensional de los autores sobre los ejes 1 y 2 proporcionará la mejor representación bidimensional de *Benjamin Kang Lim*, mientras que en el gráfico sobre los ejes 2 y 3 obtenemos la mejor representación bidimensional de *Alexander Smith* y *David Lawder*, así como la mejor representación unidimensional (sobre el eje 3) de *Brad Dorfman*.

Se muestran a continuación la representación tridimensional seguida de los gráficos bidimensionales sobre los ejes 1-2 y 2-3.

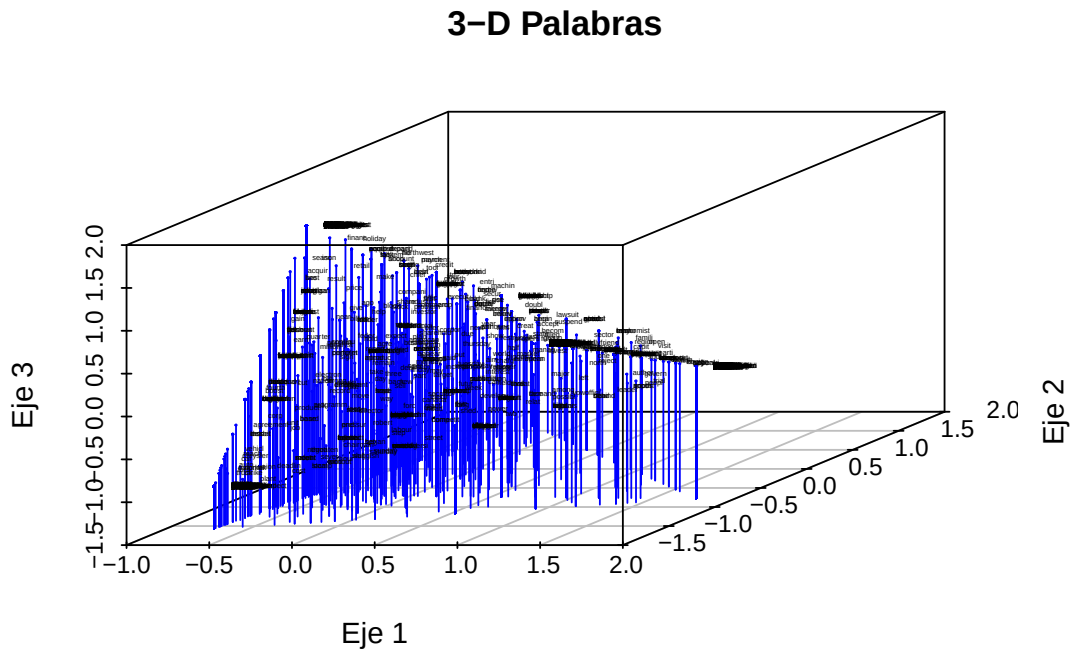




Benjamin Kang Lim se situa en el extremo superior del primer eje frente al resto de autores que ocupan una posición opuesta cercana al extremo contrario de este eje. El segundo eje contrapone a *Alexander Smith* y a *David Lawder* mientras que situa a *Benjamin Kang Lim* y a *Brad Dorfman* en una posición cercana a valores nulos. Por

último el tercer eje reproduce una forma similar al segundo eje contraponiendo en este caso a *Brad Dorfman* y a *David Lawder*.

La representación tridimensional de las columnas (palabras) sobre el subespacio generado por las tres direcciones obtenidas es la siguiente:



Como en el caso de las filas, nos interesará identificar las palabras, entre todas las que forman el corpus, que más han contribuido a la construcción de cada uno de los ejes. Recordemos que el corpus está formado por 1990 palabras diferentes y, por lo tanto, una contribución proporcional de cada una de ellas supondría un 0,05 % aproximadamente.

Tras análizar las aportaciones he decidido conservar las 151 palabras cuya aportación a la construcción de cualquiera de los ejes es $\geq 0.25\,\%$. Las primeras 50 son las siguientes:

```
##  [1] "acquir"     "acquisit"   "agreement"  "alleg"      "appeal"
##  [6] "arm"        "asset"      "auto"       "automak"    "bancorp"
## [11] "bancshar"   "bank"       "batteri"    "beij"       "bell"
## [16] "bookkeep"   "boss"       "britain"    "british"    "cabl"
## [21] "car"        "caterpillar" "chen"      "chicago"    "chief"
## [26] "china"      "chines"     "chrysler"   "communist"  "compani"
## [31] "consum"     "contract"   "corp"       "cos"        "countri"
## [36] "court"      "creat"      "credit"     "dan"        "death"
## [41] "deeper"     "deng"       "depart"     "detain"     "deutsch"
```

```
## [46] "diplomat"    "dissid"      "door"        "european"    "expand"
```

Veamos, respecto de cada eje, las diez palabras que más contribuyen a su construcción:

```
##  [1] "china"     "chines"    "beij"      "dissid"    "corp"      "sentenc"
##  [7] "taiwan"    "wang"      "communist" "court"
```

En el primer eje destacan palabras relacionadas con China que, como ya sabemos son muy representativas de las noticias de *Benjamin Kang Lim*, autor predominante en la construcción de este eje.

```
##  [1] "british" "manag"   "motor"   "fund"    "britain" "corp"    "bank"
##  [8] "general" "invest"  "auto"
```

En el segundo eje la palabra más importante es *british*, fuertemente asociada a *Alexander Smith*. Entre las demás se encuentran palabras características de este mismo autor (como *fund* o *britain*) así como de *David Lawder* (como *motor* o *auto*).

```
##  [1] "inc"     "motor"   "financ"  "general" "mercuri" "store"   "worker"
##  [8] "auto"    "retail"  "bancorp"
```

En el tercer eje la palabra más importante es *inc*, una de las palabras más frecuente y de mayor tfIdf de *Brad Dorfman*. El resto son palabras igualmente representativas de este autor (como *financ* o *mercuri*) y de *David Lawder* (como *motor* o *general*).

En las primeras fases de exploración del corpus identificamos, mediante el estudio de las frecuencias y de los valores *tfIdf*, las palabras más relevantes para cada autor. Este conocimiento puede utilizarse ahora para facilitar la exploración de los datos y su interpretación representando únicamente las palabras más relevantes para cada uno.

Así, una vez identificadas las palabras que más han contribuido a la construcción de los ejes buscamos conservar aquellas más representativas (atendiendo al valor tfIdf y a la frecuencia) para cada uno de los autores.

La siguiente tabla resume numéricamente el resultado de cruzar las palabras que cuentan con una aportación $\geq 0{,}25\,\%$ a cualquiera de los ejes con el mismo número de palabras (151) con mayor frecuencia y mayor valor *tfIdf* de cada autor:

```
##
##                    1  2  3
##   AlexanderSmith   3 26  6
##   BenjaminKangLim 31  3  4
##   BradDorfman      4  7 30
##   DavidLawder      7 15 16
```

Esta tabla muestra como se reproduce la influencia de los autores sobre cada uno de los ejes en la contribución de las palabras. El autor que más contribuye a la construcción del primer eje es *Benjamin Kang Lim* y 31 palabras de las que más contribuyen a la construcción de este eje se encuentran entre las más relevantes para este autor frente a 7 de *David Lawder*, 4 de *Brad Dorfman* y 3 de *Alexander Smith*. En cuanto al segundo eje, los autores que más contribuyen a su construcción son *Alexander Smith* y *David Lawder* y observamos como, respectivamente, 26 y 15 palabras de las que más contribuyen se encuentran entre las más representativas de estos autores frente a 7 de *Brad Dorfman* y 3 de *Benjamin Kang Lim*. Por último, en la construcción del tercer eje la mayor contribución corresponde a *Brad Dorfman* y *David Lawder* y en la contribución de las palabras observamos, respectivamente, 30 y 16 palabras representativas de estos autores frente a 6 de *Alexander Smith* y 4 de *Benjamin Kang Lim*.

Cada una de estas palabras pueden contribuir a la construcción de uno o más ejes y hacerlo por parte de uno o más autores por lo que del total de 152 palabras que refleja la tabla resultan en realidad las siguientes 120 que, siendo relevantes para alguno de los autores, aportan al menos un $0{,}25\,\%$ a la construcción de alguno de los ejes:

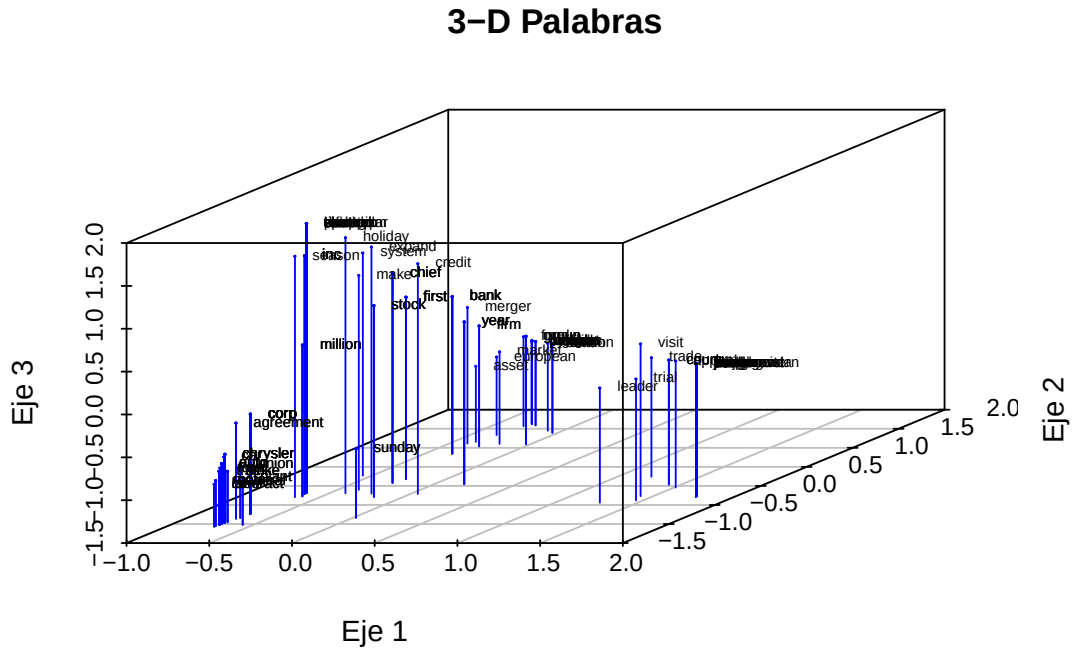
```
##   [1] "china"        "inc"          "british"      "motor"        "motor"
##   [6] "general"      "fund"         "britain"      "corp"         "bank"
##  [11] "general"      "auto"         "group"        "morgan"       "corp"
##  [16] "store"        "market"       "auto"         "taiwan"       "wang"
##  [21] "plc"          "bancorp"      "communist"    "court"        "strike"
##  [26] "motor"        "caterpillar"  "contract"     "strike"       "dan"
```

```
##  [31] "island"       "holiday"      "pound"        "union"        "post"
##  [36] "contract"     "general"      "chief"        "death"        "chrysler"
##  [41] "natwest"      "spend"        "asset"        "auto"         "firm"
##  [46] "leader"       "truck"        "union"        "inc"          "shortterm"
##  [51] "boss"         "son"          "tibet"        "truck"        "mutual"
##  [56] "pension"      "plant"        "million"      "car"          "plant"
##  [61] "system"       "merger"       "season"       "cos"          "ford"
##  [66] "deng"         "jail"         "jiang"        "mother"       "overthrow"
##  [71] "propaganda"   "zemin"        "corp"         "horlick"      "nicola"
##  [76] "probe"        "appeal"       "european"     "chrysler"     "arm"
##  [81] "london"       "trial"        "british"      "first"        "trade"
##  [86] "expand"       "ford"         "stock"        "chen"         "human"
##  [91] "korean"       "least"        "northwestern" "paramount"    "riot"
##  [96] "xinjiang"     "bell"         "mgam"         "nynex"        "pretax"
## [101] "wireless"     "chicago"      "deeper"       "revamp"       "roebuck"
## [106] "tcf"          "tire"         "trust"        "worth"        "strike"
## [111] "agreement"    "labor"        "visit"        "car"          "make"
## [116] "credit"       "sunday"       "bank"         "labor"        "year"
```

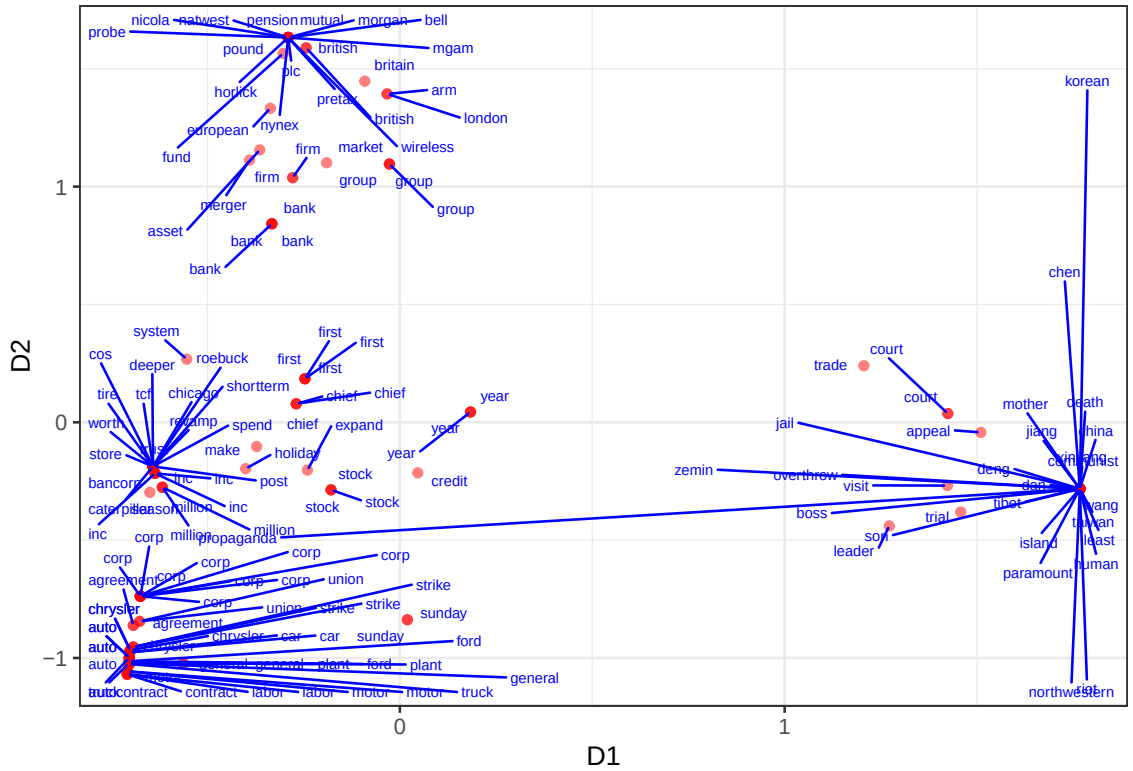
```
##    PALABRA DIMENSION CONTRIBUCION
## 39   china         1     4.246520
## 99     inc         3     2.497152
## 8  british         2     2.463342
```

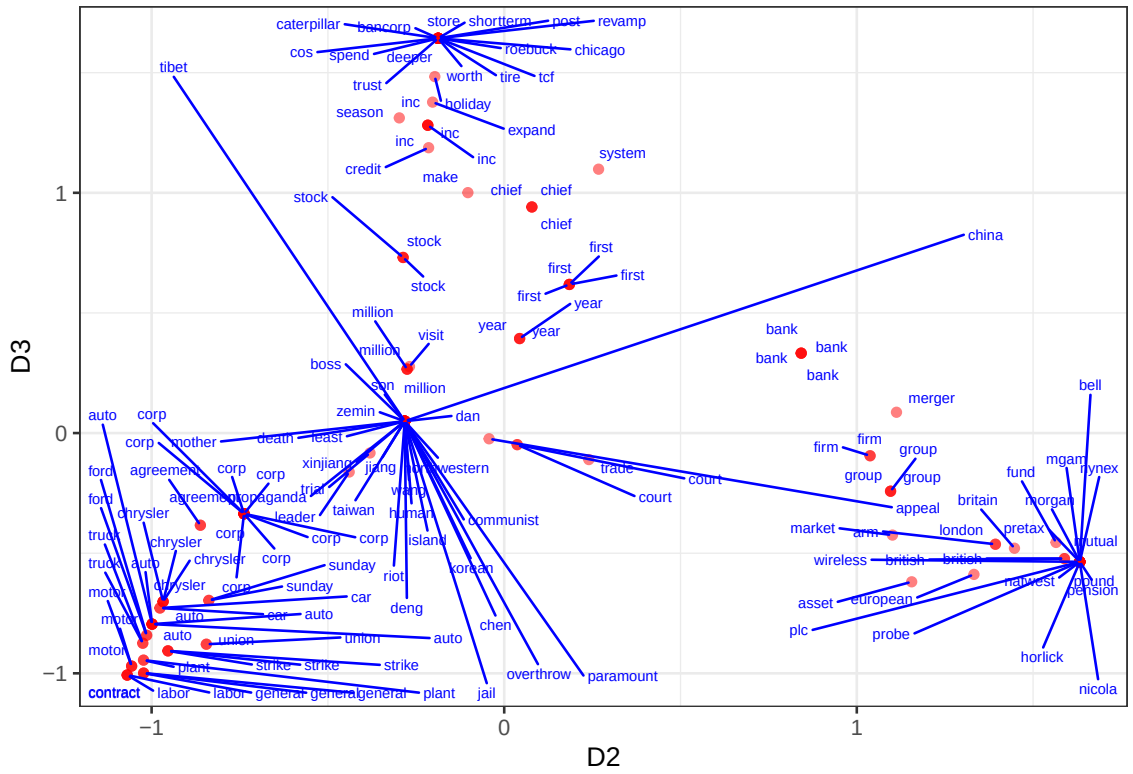
La palabra que más ha aportado a la construcción del primer eje ha sido *china* con un 4,25 %. La de mayor contribución al segundo eje ha sido *british* con un 2,46 % y respecto del tercer eje la primera posición la ha ocupado *inc* con un 2,5 %.

La representación tridimensional de estas palabras es la siguiente:



Las correspondientes representaciones bidimensionales sobre los ejes 1-2 y 2-3 son las siguientes:





La representación conjunta de filas (autores) y columnas (palabras), aunque es posible en análisis de correspondencias y no carece de interés, no permite realizar interpretaciones acerca de la proximidad de determinadas palabras sobre determinadas autores. Sin embargo esta interpretación de las proximidades entre autores y palabras es posible mediante la técnica de escalado multidimensional denominada *unfolding* que será aplicada más adelante.

4.2.3. Escalado multidimensional.

Para aplicar estas técnicas es necesario contar con una matriz inicial de disimilaridades. En el caso de la tabla léxica agregada, aplicando la distancia del coseno, se obtiene una medida de la similaridad que se transforma fácilmente en disimilaridad resultando la siguiente matriz:

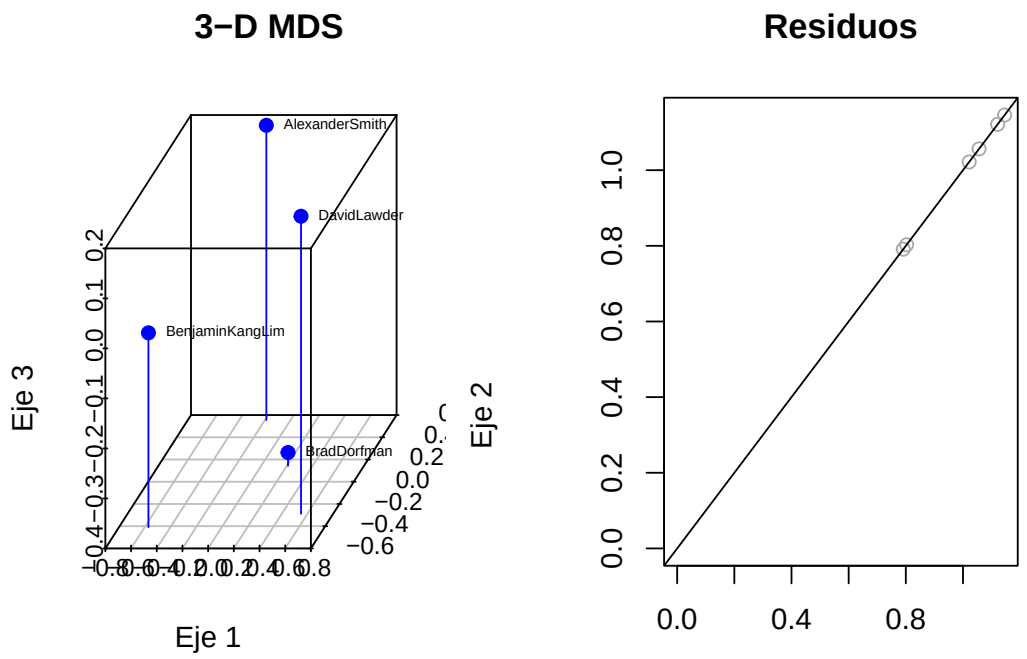
```
##                 AlexanderSmith BenjaminKangLim BradDorfman DavidLawder
## AlexanderSmith        0.000000        0.696476    0.529259    0.739267
## BenjaminKangLim       0.696476        0.000000    0.673665    0.755474
## BradDorfman           0.529259        0.673665    0.000000    0.521640
## DavidLawder           0.739267        0.755474    0.521640    0.000000
```

Sabemos que la dimensión del subespacio para el que la solución aportada por MDS explica el 100 % de los datos es, como máximo, $n-1$, siendo n el rango de la matriz anterior, y además, que el porcentaje explicado de los datos se puede obtener a partir de la matriz resultante al aplicar MDS ya que la suma de sus valores propios positivos representa el 100 %.

```
## [1]  47.153  85.768 100.000 100.000
```

En efecto, la configuración de dimensión $n-1=3$ explica el 100 % de los datos de la matriz de disimilaridades, mientras que la configuración bidimensional obtenida mediante MDS *clásico* explica el 85,77 %.

Por lo tanto en el caso tridimensional, en el que la configuración obtenida explica el 100 % de las disimilaridades de partida, *S-stres=0* y los residuos son nulos, por lo que el algoritmo *Smacof* iterará una única vez resultando equivalente, tanto en su modalidad métrica como no métrica, a la aplicación del método MDS *clásico*.



Vamos ahora a aplicar el algoritmo iterativo de minimización del stress *Smacof*, en sus modalidades métrica y no métrica, para obtener las configuraciones en 2 dimensiones.

Obtenemos los siguientes resultados respecto de la bondad del ajuste (1-S_stress) y el número de iteraciones realizadas hasta alcanzar el criterio de convergencia (0,000001) en la minimización del S-stress.

```
## S-Stress (MDS Métrico): 0.097654
```

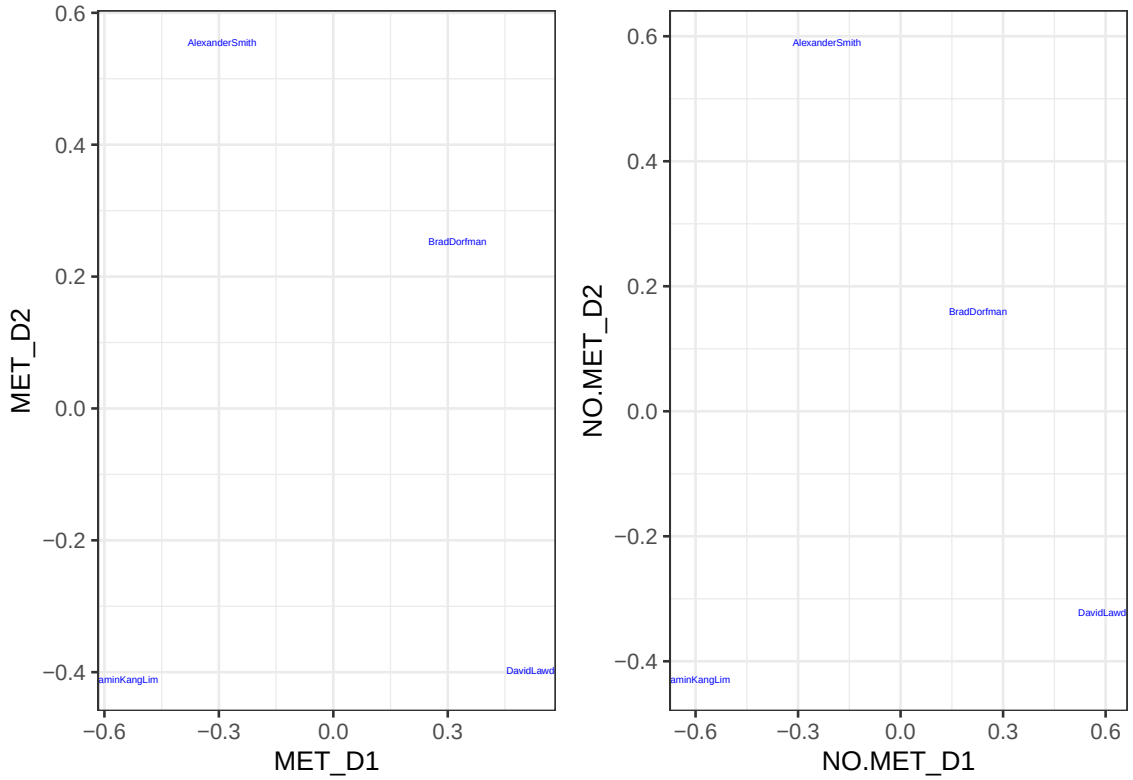
```
## S-Stress (MDS No Métrico): 0.0008265046
```

```
##     TIPO    BONDAD ITERACIONES
## 1    MET 0.9023460          22
## 2 NO.MET 0.9991735           8
```

Con el método métrico el algoritmo necesita 22 iteraciones para alcanzar el criterio de convergencia y obtenemos un valor S-stress=0,0976 (el modelo ajustado explica un 90,23 % de las disimilaridades observadas) mientras que el método no métrico alcanza el criterio de convergencia con tan solo 8 iteraciones explicando el 99,92 % de las disimilaridades (S-stress=0,0008).

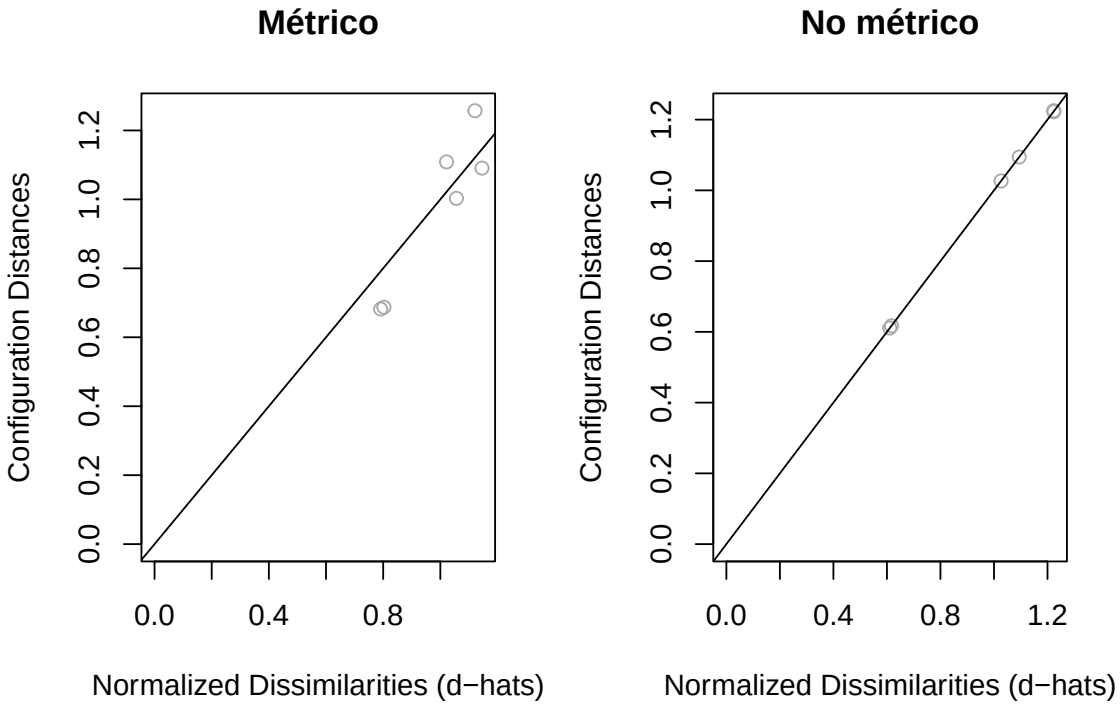
Las configuraciones obtenidas con cada uno de estos ajustes son las siguientes

```
##                      MET_D1     MET_D2  NO.MET_D1  NO.MET_D2
## AlexanderSmith  -0.2914495  0.5552867 -0.2152174  0.5903159
## BenjaminKangLim -0.5624844 -0.4104166 -0.6146073 -0.4284357
## BradDorfman      0.3257546  0.2531377  0.2268681  0.1596627
## DavidLawder      0.5281793 -0.3980078  0.6029566 -0.3215428
```



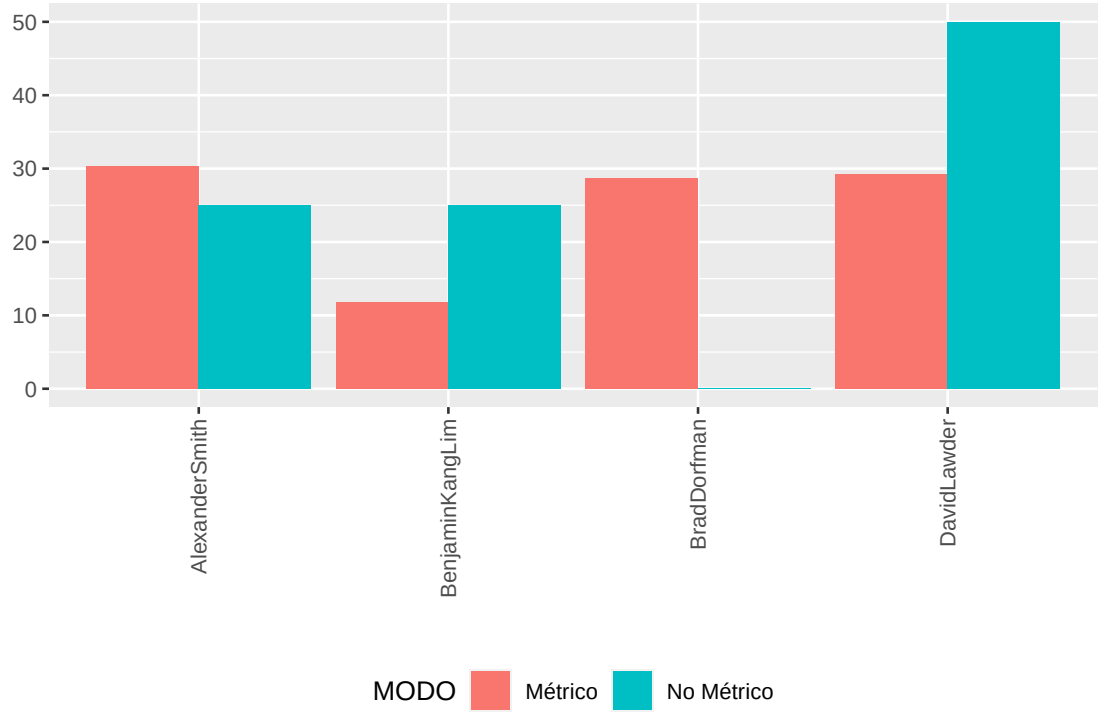
En la configuración obtenida mediante el método no métrico *Brad Dorfman* se encuentra más *alineado* con *Alexander Smith* y *David Lawder* y recordemos que este método proporciona un mejor ajuste (99,92 % explicado frente al 90,23 % del método métrico).

```
## Residuos
```



Los gráficos de residuos ponen de manifiesto la diferencia cualitativa a favor del método no métrico en cuanto a la bondad del ajuste. Veamos una comparativa de la aportación al stress de cada autor en ambos métodos:

```
## Contribución al stress:

##              AlexanderSmith DavidLawder BradDorfman BenjaminKangLim
## Métrico             30.30229    29.24228    28.70841         11.74701
## No métrico          49.99998    25.03706    24.96295          0.00002
```



En el caso métrico, el autor que contribuye al stress en menor medida es *Benjamin Kang Lim* (11,75 %), mientras que los restantes autores se reparten el 88,25 % restante casi a partes iguales. Sin embargo en el caso no métrico la contribución al stress es prácticamente nula para *Brad Dorfman* (0,00002 %) mientras que a *David Lawder* le corresponde el 50 % del stress total seguido de *Benjamin Kang Lim* (25,04 %) y Alexander Smith (24,96 %).

Al aplicar el análisis de correspondencias habíamos identificado el conjunto de palabras cuya aportación a la construcción de cualquiera de los ejes era $\geq 0{,}25\,\%$. A partir de estas palabras construimos la correspondiente matriz de distancias y determinamos el número de dimensiones idóneo para aplicar el escalado multidimensional.

```
##   [,1]            [,2]            [,3]            [,4]
## h "46.511"        "79.423"        "99.899"        "100"
##   "..."       ..." "..."       ..." "..."       ..." "..."       ..."
## t "100"           "100"           "100"           "100"
```

La representación bidimensional conservará el 79,42 % de la variabilidad, alcanzándose el 99,9 % en el caso tridimensional. Aplicaremos de nuevo los métodos métrico y no métrico comenzando por el caso tridimensional.

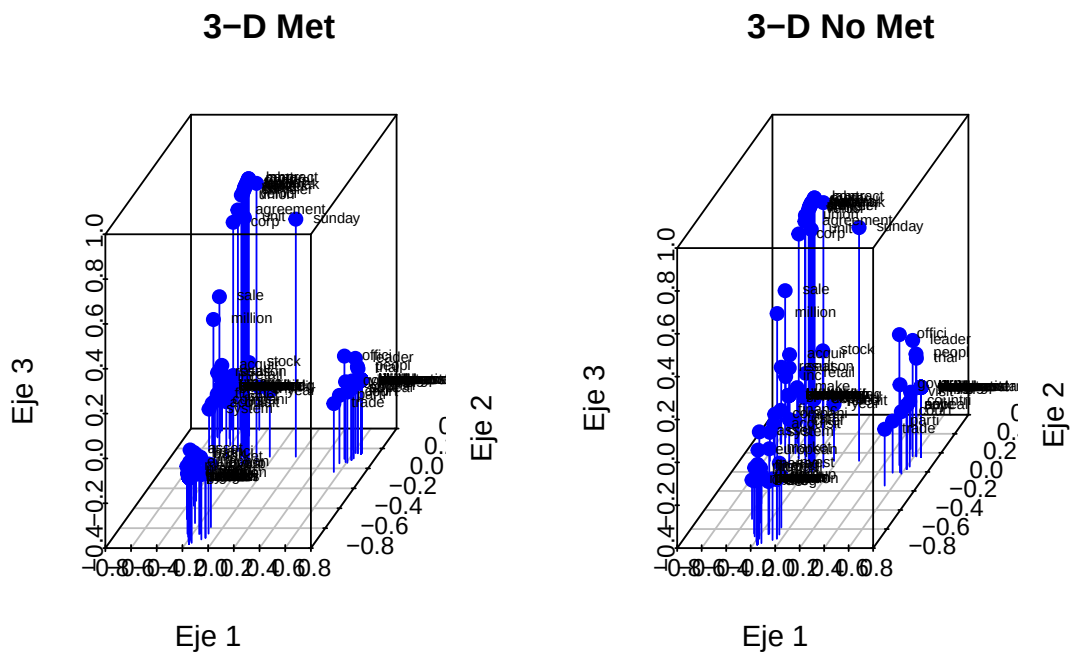
```
## S-Stress (MDS Métrico): 0.07692117
```

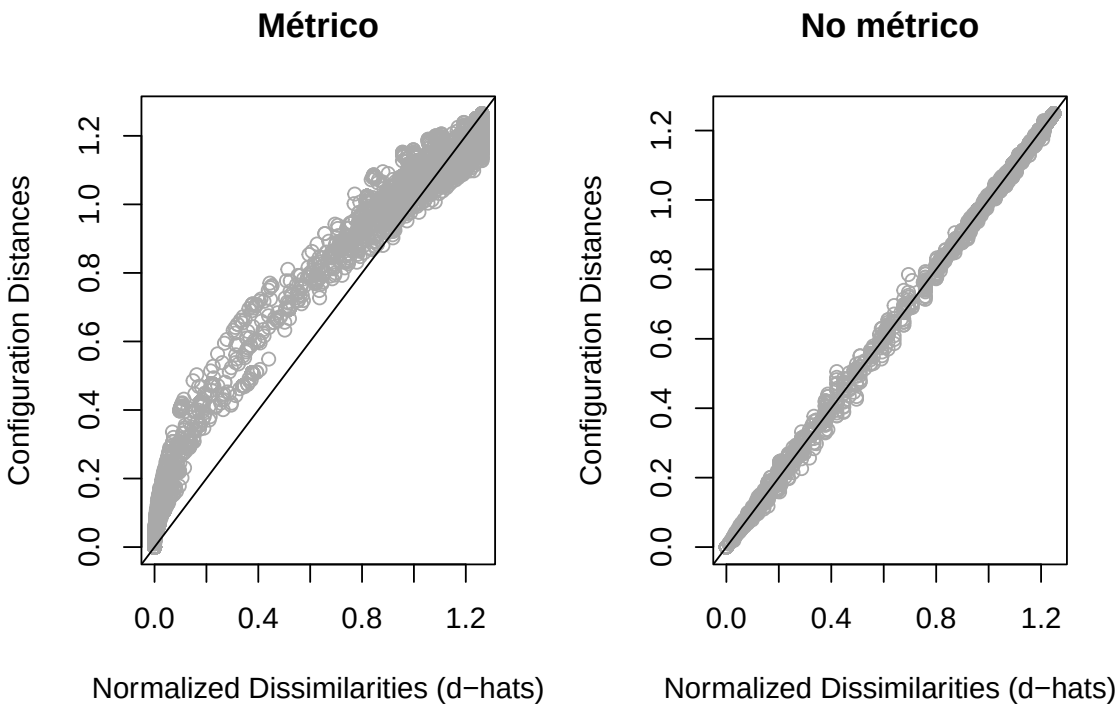
```
## S-Stress (MDS No Métrico): 0.006743885
```

```
##     TIPO    BONDAD ITERACIONES
## 1    MET 0.9230788          10
## 2 NO.MET 0.9932561          14
```

Con el método métrico el algoritmo necesita 10 iteraciones para alcanzar el criterio de convergencia y explica un 92,3 % de las disimilaridades observadas mientras que el método no métrico necesita 14 iteraciones pero explica el 99,3 %.

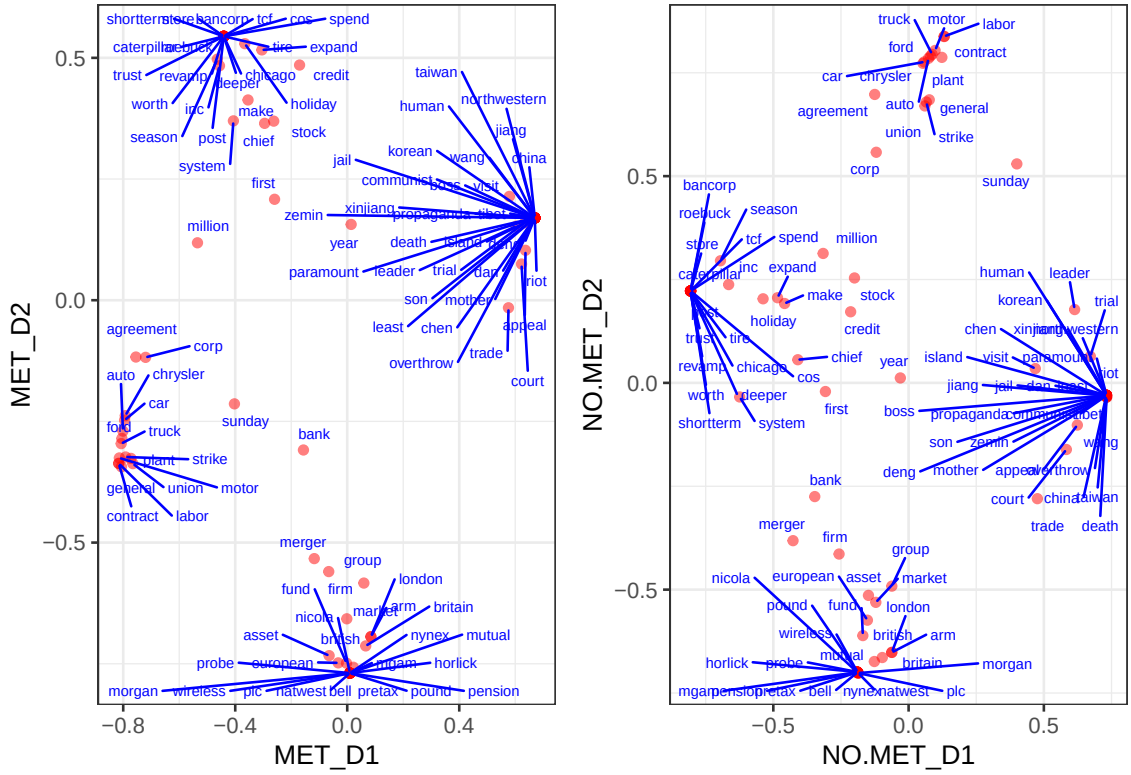
A continuación se muestran las representaciones de las palabras y los gráficos de residuos (que de nuevo manifiestan el mejor ajuste obtenido con el método no métrico):





Los resultados en el caso bidimensional son los siguientes:

```
##      TIPO    BONDAD ITERACIONES
## 1     MET 0.8277487          66
## 2 NO.MET 0.9159765          90
```



4.2.4. Unfolding.

La técnica de *unfolding* se utiliza para obtener una configuración en el mismo subespacio tanto para individuos como para variables cuando los datos representan *preferencias*.

Siguiendo el procedimiento empleado hasta ahora nos limitaremos a las palabras relevantes y, puesto que en nuestra tabla léxica agregada los individuos se corresponden con los autores, las variables con las palabras y los valores observados son frecuencias podemos realizar la transformación "mayor frecuencia relativa = mayor preferencia". Para ello, teniendo en cuenta que para aplicar el algoritmo el valor más pequeño corresponderá a la mayor preferencia y cuanto mayor sea el valor menor será la preferencia, debemos aplicar sobre las frecuencias relativas la función inversa para lo cual, dado que tenemos valores de frecuencia relativa nulos, los transformaremos previamente en valores cercanos a cero.

Es decir, tomando como ejemplo las 10 primeras palabras de la tabla léxica agregada:

```
##            Docs
## Terms       AlexanderSmith BenjaminKangLim BradDorfman DavidLawder
##   abandon                0               2           0           0
##   abil                   2               0           0           0
##   abl                    1               1           0           0
##   abn                    1               0           1           0
##   abroad                 0               1           0           0
##   absent                 0               1           0           0
##   academ                 0               1           0           0
##   acceler                0               0           1           0
##   accept                 3               1           1           0
##   accomplic              0               1           0           0
```

transformamos las frecuencias absolutas en frecuencias relativas por autores para eliminar la influencia del distinto número total de palabras de cada autor

```
##            Docs
## Terms       AlexanderSmith BenjaminKangLim BradDorfman DavidLawder
##   agreement         0.0000          0.0000      0.0021      0.0059
##   appeal            0.0005          0.0036      0.0000      0.0000
```

```
##   arm              0.0035          0.0005      0.0000      0.0000
##   asset            0.0070          0.0000      0.0000      0.0014
##   auto             0.0000          0.0000      0.0021      0.0208
##   bancorp          0.0000          0.0000      0.0051      0.0000
##   bank             0.0250          0.0021      0.0175      0.0000
##   bell             0.0025          0.0000      0.0000      0.0000
##   boss             0.0000          0.0036      0.0000      0.0000
##   britain          0.0140          0.0016      0.0000      0.0000
```

Ahora, sustituimos los ceros por un valor mucho más pequeño que la menor frecuencia relativa distinta de cero. La mínima frecuencia relativa distinta de cero en la matriz es:

```
## [1] 0.0004531038
```

Sustituimos los ceros, por ejemplo, por 0,000001 y a continuación aplicamos la función inversa para que las palabras de mayor frecuencia relativa sean las que tengan valores más pequeños, y por tanto interpretadas por el algoritmo como de mayor preferencia.

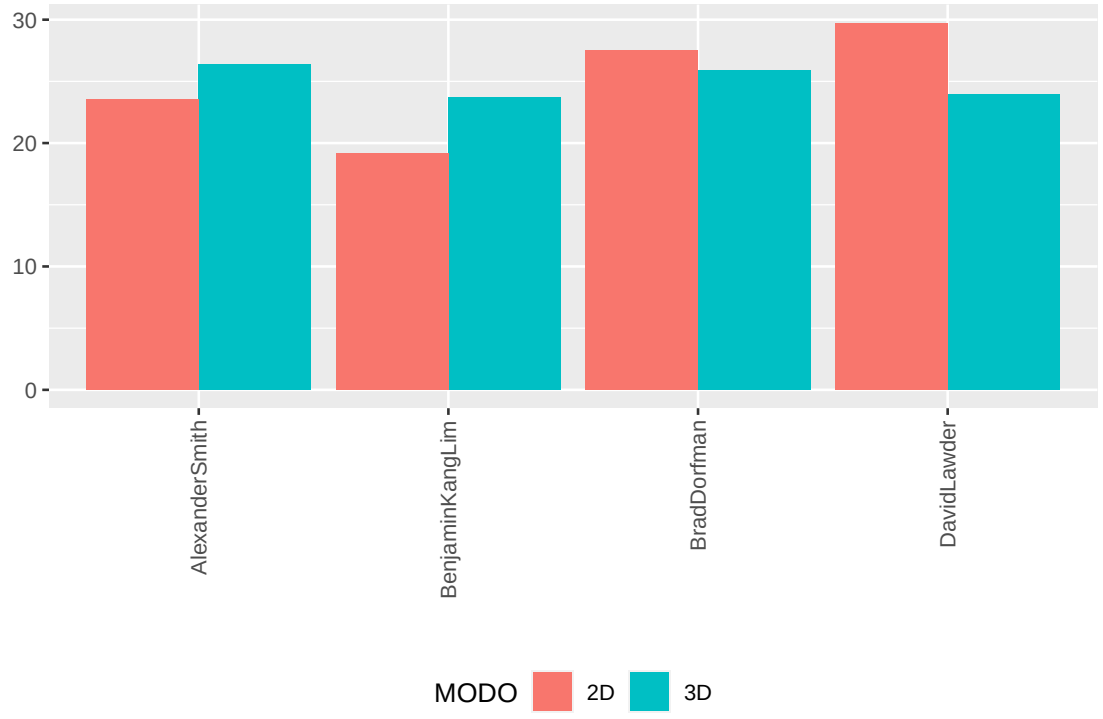
```
##            Docs
## Terms       AlexanderSmith BenjaminKangLim BradDorfman DavidLawder
##   agreement        1000000          1000000         487         170
##   appeal              2001              274     1000000     1000000
##   arm                  286             1921     1000000     1000000
##   asset                143          1000000     1000000         736
##   auto             1000000          1000000         487          48
##   bancorp          1000000          1000000         195     1000000
##   bank                  40              480          57     1000000
##   bell                 400          1000000     1000000     1000000
##   boss             1000000              274     1000000     1000000
##   britain               71              640     1000000     1000000
```

Así, entre las 10 primeras palabras tomadas como ejemplo, la de mayor preferencia para *Alexander Smith* es *bank*, para *Benjamin Kang Lim* son *appeal* y *boss*, para *Brad Dorfman* es *bank*, y para *David Lawder* es *auto*.

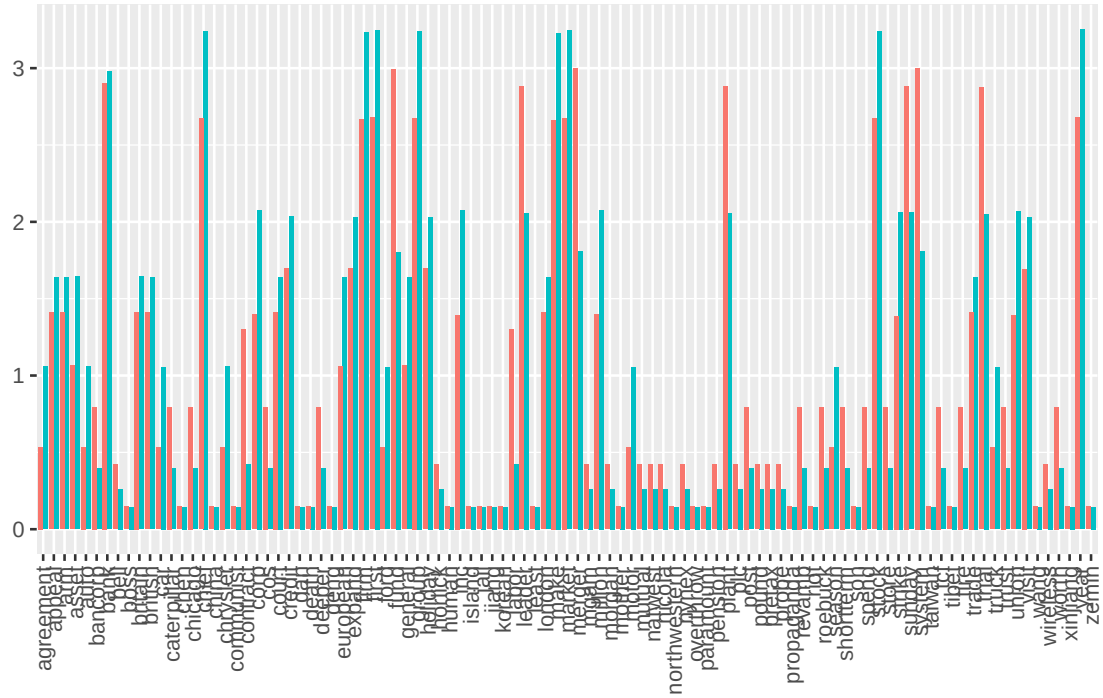
Al aplicar *unfolding* a 2 y 3 dimensiones sobre esta matriz de preferencias obtenemos los siguientes resultados acerca de la bondad del ajuste:

```
##   TIPO    BONDAD ITERACIONES
## 1   3D 0.6697214          19
## 2   2D 0.6498887          21
```

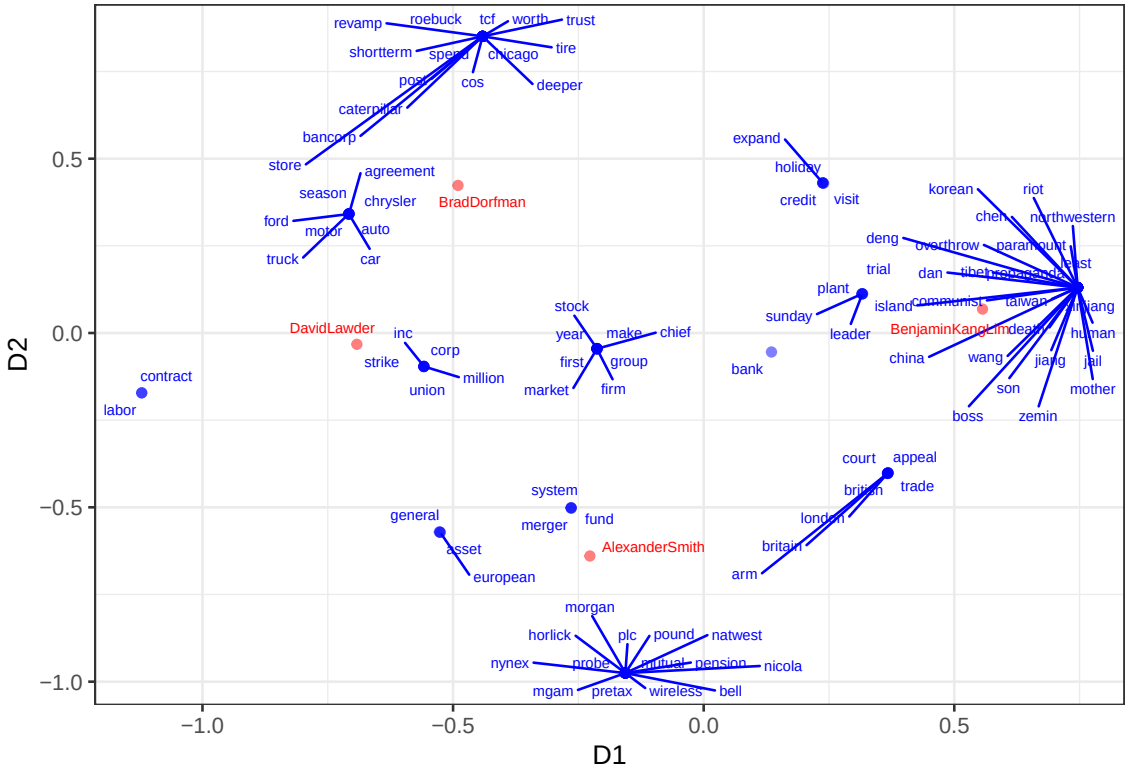
Contribución al stress de los autores:



Contribución al stress de las palabras:



La representación conjunta de los autores y las palabras más relevantes nos permite sacar conclusiones en base a las distancias entre autores y palabras de las preferencias de aquellos por estas últimas:



4.3. Clasificación mediante Análisis Discriminante Lineal.

Hasta aquí hemos mostrado una cierta metodología para abordar el análisis exploratorio de un conjunto de documentos de texto que, básicamente, podemos resumir en:

- Identificación de las palabras más relevantes, tanto a nivel global como por categorías (en nuestro caso: autores), mediante medidas estadísticas relacionadas con la frecuencia de uso de cada una de las palabras (frecuencia y tfIdf)
- Aplicación de técnicas multivariantes sobre la matriz léxica agregada para la visualización, interpretación y comparación de las categorías así como de las palabras más relevantes

Para facilitar esta interpretación/visualización se ha restringido el análisis realizado hasta el momento a los documentos pertenecientes a 4 autores. Para ilustrar la resolución del problema de clasificación asociado vamos a ampliar el conjunto de datos a los siguientes 10 autores:

```
##              NOMBRE ABREVIATURA
## 1     AaronPressman          AP
## 2        AlanCrosby          AC
## 3    AlexanderSmith          AS
## 4   BenjaminKangLim         BKL
## 5     BernardHickey          BH
## 6       BradDorfman          BD
## 7  DarrenSchuettler          DS
## 8       DavidLawder          DL
## 9     EdnaFernandes          EF
## 10      EricAuchard          EA
```

La matriz de documentos-palabras para este nuevo corpus ampliado tiene las siguientes características:

```
## <<DocumentTermMatrix (documents: 1000, terms: 3372)>>
## Non-/sparse entries: 19280/3352720
## Sparsity           : 99%
```

```
## Maximal term length: 21
## Weighting          : term frequency (tf)
```

Se trata de una matriz de 1000 filas (documentos) y 3372 columnas (palabras) con una dispersión del 99 % (19280 elementos distintos de 0 y 3352720 elementos iguales a cero).

Sabemos que el análisis discriminante lineal resulta óptimo bajo el supuesto de normalidad multivariante de las variables aunque en la práctica son habituales las situaciones en que no cumpliéndose está condición proporciona igualmente buenos resultados, motivo por el que lo aplicaremos inicialmente sin comprobar este supuesto.

Por lo tanto, aplicaremos en primer lugar el análisis discriminante lineal directamente sobre la matriz de documentos-palabras aunque, como venimos haciendo, "relativizaremos" previamente las frecuencias por documentos para evitar las diferencias debidas al distinto número total de palabras de cada uno.

A continuación aplicaremos el escalado multidimensional métrico y no métrico sobre dicha matriz de frecuencias relativas como paso previo a la aplicación del análisis discriminante.

Para determinar el modelo que consigue el mejor ajuste aplicaremos la metodología de validación cruzada "*k-folds cross validation*" con $k = 10$ mediante el siguiente procedimiento:

- En primer lugar dividimos el conjunto de 1000 documentos en un conjunto de entrenamiento (900 documentos) y un conjunto de test (los 100 documentos restantes)
- A continuación dividimos el conjunto de entrenamiento en 10 subconjuntos (k=10) con igual número de documentos de cada autor y realizamos 10 veces el análisis discriminante considerando cada vez uno de estos subconjuntos como conjunto de test y la unión de los nueve restantes como conjunto de entrenamiento.
- Sobre la suma de las 10 matrices de confusión obtenidas calculamos la precisión, recuperación e índice de Rand ajustado (ARI)

Repitiendo lo anterior y comparando los resultados para cada uno de los métodos de análisis discriminante analizados (sobre la matriz de frecuencias relativas, con previo MDS métrico y con previo MDS no métrico) determinaremos el que proporciona el mejor ajuste.

Finalmente generaremos el modelo correspondiente al mejor método sobre el subconjunto completo de entrenamiento y aplicaremos para predecir sobre el conjunto de test evaluando así la bondad del ajuste.

4.3.1. Matriz de frecuencias relativas por documentos.

Al aplicar *10-fold cross validation* sobre la matriz de frecuencias relativas obtenemos la siguiente matriz de confusión:

```
##
##        AC AP AS BD BH BKL DL DS EA EF
##    AC  44 12  3  8  3   3  2  5  1  9
##    AP   4 32  4  5 10   4  9  7  3 12
##    AS   4  8 22  8 16   6  7  7  2 10
##    BD   6  5  3 29 18   3  7  5  4 10
##    BH   4  6  4 11 45   1  7  2  2  8
##    BKL  5  8  3  4 17  25  6  9  3 10
##    DL   5 11  4  9  6   7 29  7  4  8
##    DS   4  5  3  8 10   3  4 38  3 12
##    EA   8  7  2  6 18   4  6  9 20 10
##    EF   5  8  4 10  9   3  6  5  1 39
```

y los siguientes resultados acerca de la bondad del ajuste:

```
## [1] "Precision: 0.3589"
```

```
## [1] "Adjusted Rand Index: 0.0812"
```

Teniendo en cuenta que contamos con diez autores y el mismo número de documentos de cada uno, un modelo de clasificación aleatorio puro tendrá una precisión esperada del 10 % así que la precisión del 35,89 % indica que el modelo obtenido es mejor que un modelo de clasificación al azar. Sin embargo, si consideramos que para el índice ARI un valor igual a 0 equivale a una clasificación aleatoria y un valor igual a 1 a una clasificación perfecta, el valor obtenido (ARI=0,0812) indica que el modelo obtenido no es mucho mejor que el azar.

A partir de esta matriz de confusión se obtienen los siguientes indicadores para cada autor:

```
##          AC   AP   AS   BD  BH  BKL   DL   DS   EA   EF
## Precis 0.49 0.36 0.24 0.32 0.5 0.28 0.32 0.42 0.22 0.43
## Recup  0.49 0.31 0.42 0.30 0.3 0.42 0.35 0.40 0.47 0.30
```

La primera columna contiene el índice de *precisión* (proporción de documentos de cada autor correctamente clasificados), mientras que en la segunda columna se encuentra el índice de *recuperación* (proporción de documentos clasificados en un autor que pertenecen al mismo). En este sentido, el modelo es más preciso con *Bernard Hickey*, mientras que la mejor *recuperación* se da con *Alan Crosby*.

4.3.2. Configuraciones obtenidas mediante escalado multidimensional.

Se trata ahora de encontrar, mediante escalado multidimensional, una configuración basada en la matriz original de documentos-palabras (no agregada) sobre la que aplicar la técnica del análisis discriminante y determinar si mejora los resultados obtenidos en el apartado anterior.

Para ello, calculamos en primer lugar las similaridades entre documentos con la distancia del coseno y aplicamos la transformación $1 - cos(x)$ para obtener una matriz de disimilaridades. A partir de los valores propios de esta matriz determinamos el número de dimensiones apropiado que explique un porcentaje suficiente de las disimilaridades.

```
## Rtados para dims 381 a 390 :
##  98.908 98.932 98.956 98.98 99.003 99.026 99.049 99.071 99.094 99.116
```

Con 385 dimensiones se supera el porcentaje de disimilaridades explicado del 99 %.

4.3.2.1. MDS métrico.

Realizamos a continuación el escalado multidimensional métrico mediante el algoritmo Smacof y aplicamos el análisis discriminante lineal sobre la configuración formada por las 385 primeras coordenadas. Obtenemos la siguiente matriz de confusión y los siguientes resultados sobre la evaluación del modelo resultante.

```
##
##       AC AP AS BD BH BKL DL DS EA EF
##   AC  89  0  1  0  0   0  0  0  0  0
##   AP   0 82  2  3  1   1  0  0  0  1
##   AS   0  3 75  3  1   0  0  2  1  5
##   BD   0  2  0 71  2   0  4  0 10  1
##   BH   1  3  4  2 76   0  0  0  0  4
##   BKL  0  0  1  1  0  88  0  0  0  0
##   DL   0  0  2  5  0   0 78  0  3  2
##   DS   0  0  0  0  1   0  1 87  1  0
##   EA   1  0  2  5  0   0  0  0 82  0
##   EF   0  0  8  2  2   0  0  0  2 76
```

```
## [1] "Precision: 0.8933"
```

```
## [1] "Adjusted Rand Index: 0.7823"
```

La precisión ha aumentado hasta el 89,33 % y el índice ARI hasta 0,78.

Los indicadores individuales por autores resultan:

```
##          AC   AP   AS   BD   BH  BKL   DL   DS   EA   EF
## Precis 0.99 0.91 0.83 0.79 0.84 0.98 0.87 0.97 0.91 0.84
## Recup  0.98 0.91 0.79 0.77 0.92 0.99 0.94 0.98 0.83 0.85
```

La precisión oscila entre 0,79 (*Brad Dorfman*) y 0,99 (*Alan Crosby*), y la recuperación entre 0,77 (*Brad Dorfman*) y 0,99 (*Benjamin Kang Lim*). Recordemos que, en el apartado anterior, al realizar análisis discriminante a partir de la matriz de frecuencias relativas el índice de precisión mas alto por autores era 0,5 y el de recuperación era 0,49.

4.3.2.2. MDS no métrico.

En cuanto a la aplicación del escalado multidimensional en su versión no métrica obtenemos los siguientes resultados

```
##
##        AC AP AS BD BH BKL DL DS EA EF
##   AC   90  0  0  0  0   0  0  0  0  0
##   AP    0 81  4  2  1   0  0  0  1  1
##   AS    0  1 77  1  0   0  1  0  1  9
##   BD    0  2  0 77  2   0  3  0  6  0
##   BH    0  1  1  3 80   0  0  0  3  2
##   BKL   0  0  0  3  0  86  0  0  1  0
##   DL    0  2  2  7  1   0 73  0  0  5
##   DS    0  0  0  4  1   1  1 82  1  0
##   EA    0  2  1 10  0   0  0  0 76  1
##   EF    0  0  6  0  2   0  1  0  3 78

## [1] "Precision: 0.8889"

## [1] "Adjusted Rand Index: 0.7709"
```

La precisión global de la matriz de confusión y el índice ARI son ligeramente inferiores a los obtenidos con el método métrico (88,89 frente a 89,33 % y 0,77 frente a 0,78).

Veamos los resultados por autores:

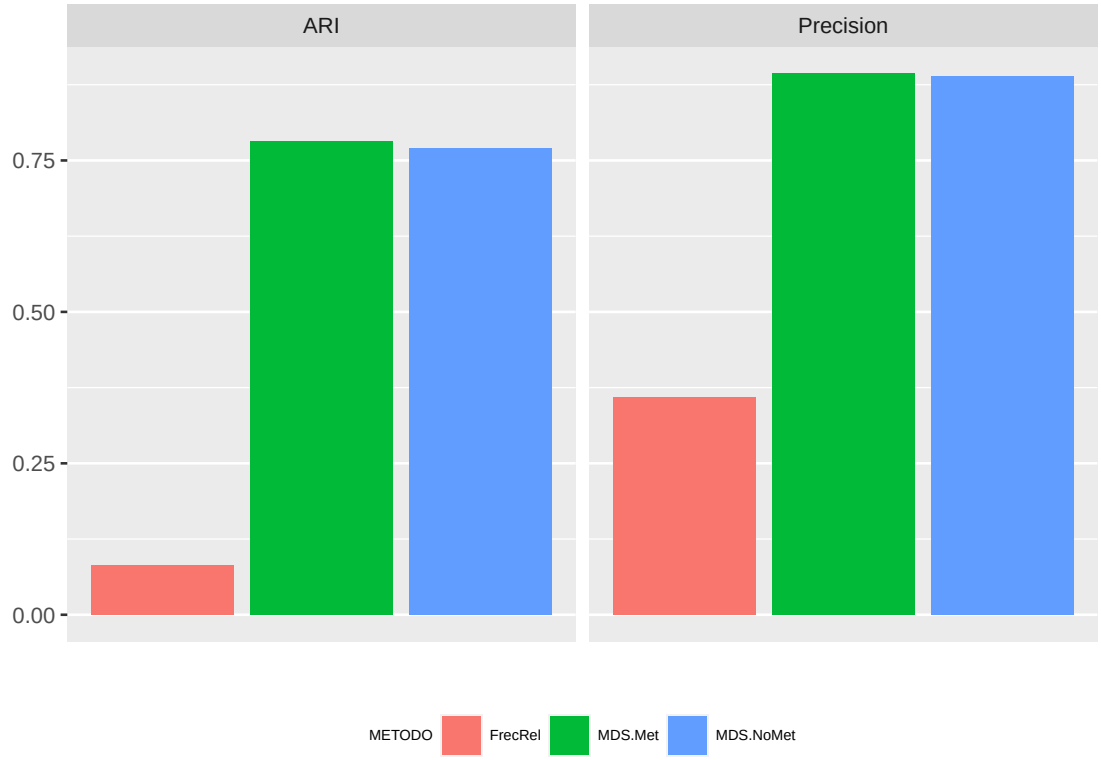
```
##         AC   AP   AS   BD   BH  BKL   DL   DS   EA   EF
## Precis   1 0.90 0.86 0.86 0.89 0.96 0.81 0.91 0.84 0.87
## Recup    1 0.91 0.85 0.72 0.92 0.99 0.92 1.00 0.83 0.81
```

El modelo no métrico alcanza el 100 % de precisión para los documentos de *Alan Crosby* y, aunque la recuperación mínima es inferior que en el caso métrico (72 frente a 77 %), alcanza el 100 % en dos autores (*Alan Crosby* y *Darren Schuettler*).

4.3.3. Resultados globales del ajuste.

A nivel global se aprecia claramente la mejora producida al aplicar el análisis discriminante sobre la configuración resultante de un previo escalado multidimensional frente a su aplicación directa sobre la matriz de frecuencias. En este último caso el índice ARI muestra que el modelo obtenido apenas difiere de un modelo aleatorio puro, mientras que mediante el escalado multidimensional previo este índice supera el valor 0,77 tanto con el

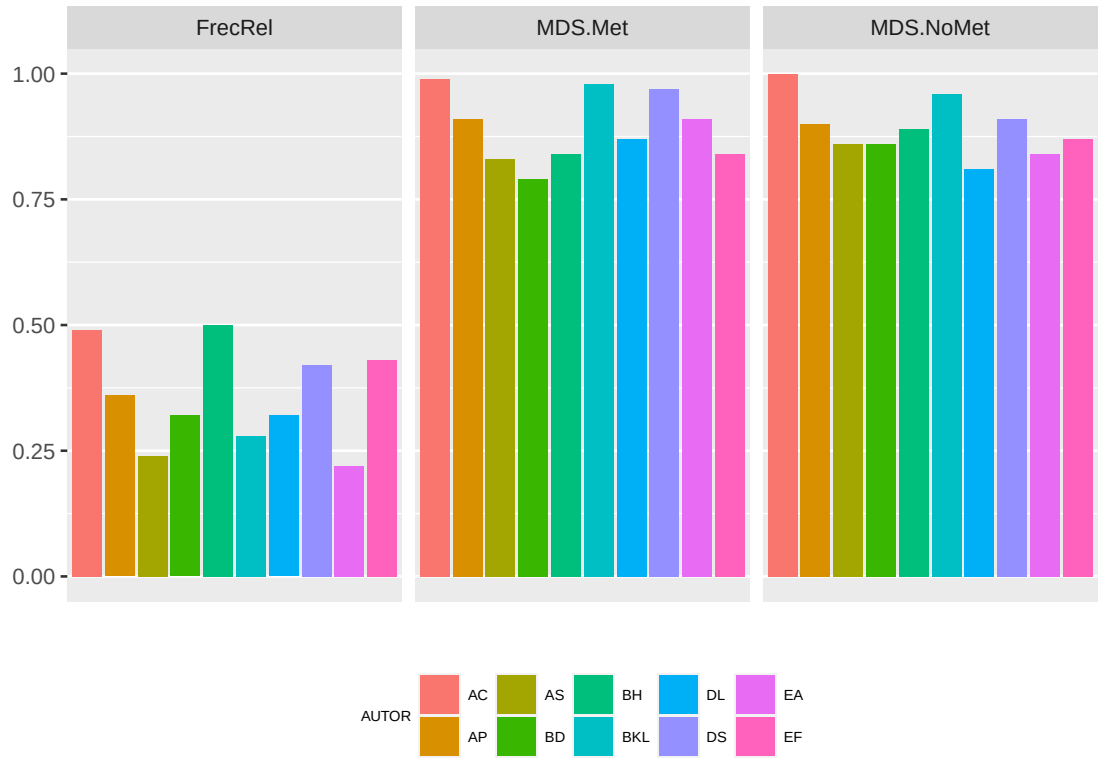
método métrico como con el no métrico con un porcentaje de documentos clasificados correctamente superior al 88 % en ambos casos.



4.3.3.1. Resultados del índice de precisión por autores.

```
##             AC   AP   AS   BD   BH  BKL   DL   DS   EA   EF
## FrecRel   0.49 0.36 0.24 0.32 0.50 0.28 0.32 0.42 0.22 0.43
## MDS.Met   0.99 0.91 0.83 0.79 0.84 0.98 0.87 0.97 0.91 0.84
## MDS.NoMet 1.00 0.90 0.86 0.86 0.89 0.96 0.81 0.91 0.84 0.87
```

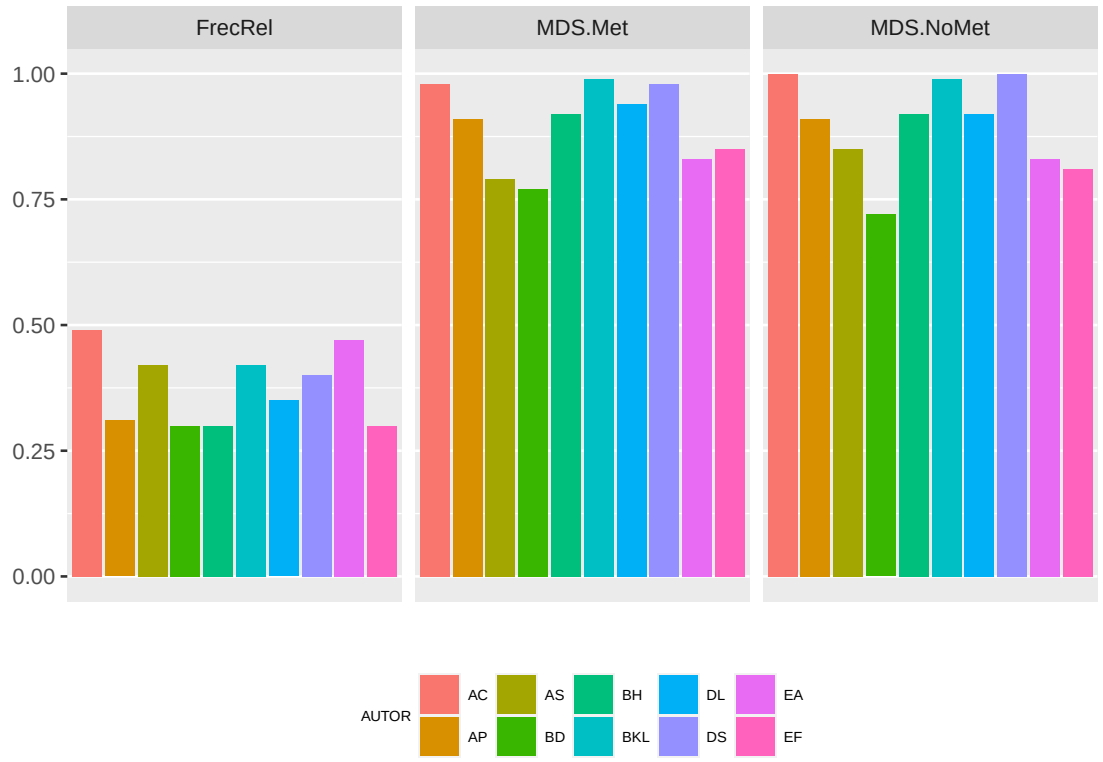
En todos los casos el autor con mayor proporción de sus documentos clasificados correctamente es *Alan Crosby*. El porcertaje de documentos clasificados correctamente para cada uno de los autores oscila entre el 22 y el 50 % en el caso de la matriz de frecuencias frente a 79 y 99 % para MDS métrico u 81 y 100 % con MDS no métrico.



4.3.3.2. Resultados del índice de recuperación por autores.

```
##              AC   AP   AS   BD   BH   BKL   DL   DS   EA   EF
## FrecRel    0.49 0.31 0.42 0.30 0.30 0.42 0.35 0.40 0.47 0.30
## MDS.Met    0.98 0.91 0.79 0.77 0.92 0.99 0.94 0.98 0.83 0.85
## MDS.NoMet 1.00 0.91 0.85 0.72 0.92 0.99 0.92 1.00 0.83 0.81
```

Los índices de recuperación oscilan entre 30 y 49 % con la matriz de frecuencias, 77 y 99 % con MDS métrico y entre 72 y 100 % en MDS no métrico



En definitiva podemos concluir que la aplicación de la metodología del análisis discriminante clásico para obtener un modelo que permita clasificar los textos del corpus objeto de estudio en función de sus autores no es apropiada cuando se aplica directamente sobre la matriz de frecuencias directamente, mientras que la aplicación a esta matriz de las técnicas de escalado multidimensional proporciona una configuración muy reducida (alrededor del 10 % de las dimensiones de la matriz original) que produce una mejora sustancial de los resultados, superándose el 88 % de documentos clasificados correctamente con un índice de Rand ajustado mayor que 0,77.

4.3.4. Aplicación a los datos reservados para test.

Los resultados obtenidos muestran la conveniencia de aplicar MDS sobre la matriz de frecuencias con carácter previo al análisis discriminante. Al haber obtenido resultados muy similares tanto en el caso métrico como no métrico los aplicaremos de nuevo ambos para comparar los resultados.

En los problemas de clasificación aplicados en la vida real nos podemos encontrar con dos situaciones:

- En el momento de entrenar el modelo se dispone de la información de toda la población aunque únicamente se conozca la clase de un subconjunto de la misma que se utilizará como conjunto de entrenamiento.
- En el momento de entrenar el modelo se dispone de la información de un conjunto de elementos, incluida la clase a la que pertenece cada uno, mientras que la información acerca de los elementos a clasificar se va conociendo en momentos posteriores al entrenamiento del modelo.

El análisis estadístico de textos tiene una particularidad: los individuos son los documentos y las variables medidas son las palabras que forman parte de cada uno. La fase de preprocesamiento finaliza con la construcción de la matriz de documentos-palabras sobre la que aplicaremos las técnicas de escalado multidimensional. Por lo tanto, supondremos que se trata de resolver un problema de clasificación en el que, en el momento inicial, todos los documentos son conocidos ya que en otro caso, cuando se prevea que los documentos a clasificar se conocerán en momentos posteriores a la construcción del modelo, la aplicación de las técnicas MDS con carácter previo al análisis discriminante requerirá calcular las disimilaridades entre el documento a clasificar y todos los del conjunto de entrenamiento para a continuación obtener las coordenadas del nuevo punto y, mediante el modelo de análisis discriminante construido, determinar la clase a la que pertenece. Esta situación constituye un caso que excede los objetivos de este TFM y requiere un estudio pormenorizado de los métodos disponibles que pudieran resultar más apropiados por lo que se plantea como una futura línea de investigación.

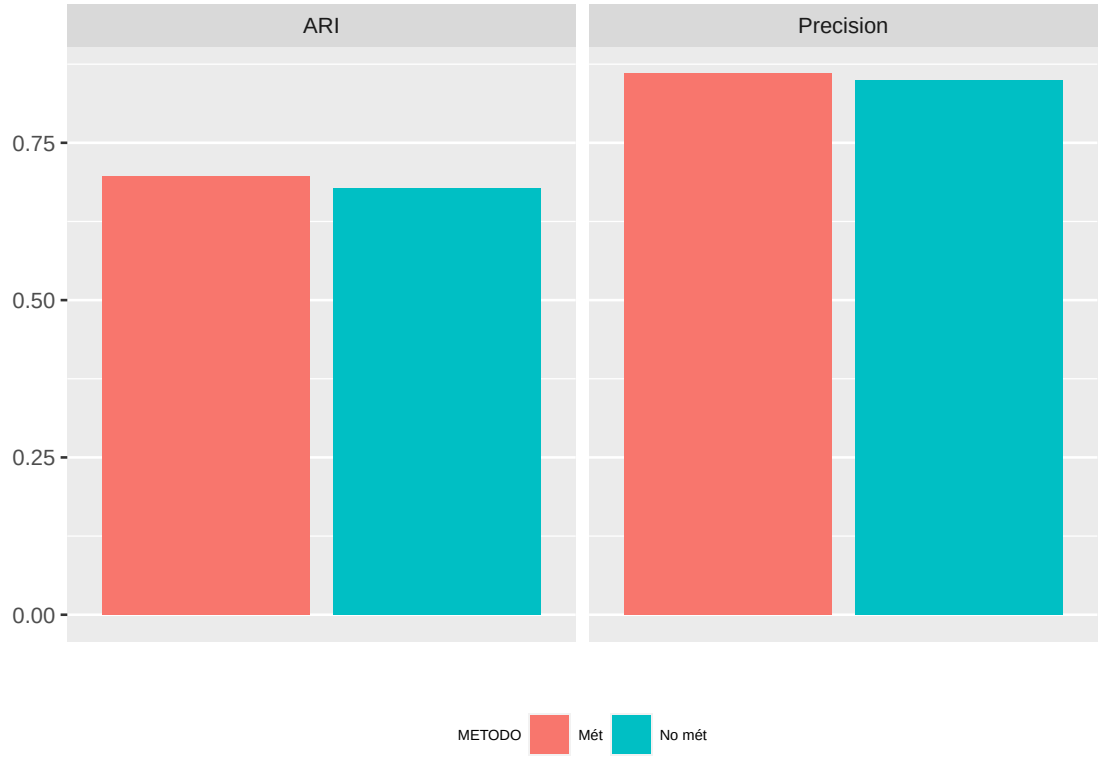
Así, consideramos que en el momento de abordar la construcción del modelo se dispone de todos los documentos a clasificar y, por lo tanto:

- Incorporamos estos 100 documentos a la matriz de documentos-palabras
- Calculamos las disimilaridades entre todos los documentos del corpus
- Obtenemos las correspondientes configuraciones MDS
- Aplicamos LDA sobre las configuraciones MDS de los 900 documentos del conjunto de entrenamiento

- Con el modelo obtenido determinamos, sobre sus configuraciones MDS, las clases de los 100 documentos de test, es decir sus autores

Obtenemos los siguientes resultados:

```
##             PRECISION    ARI
## MDS Met          0.86 0.6963
## MDS No met       0.85 0.6782
```



El método métrico clasifica correctamente 86 de los 100 documentos desconocidos por el modelo mientras que el método no métrico clasifica correctamente 85 de estos documentos. En cuanto al índice de Rand ajustado obtenemos igualmente valores muy similares en ambos métodos (0,6963 para el método métrico frente a 0'6782 para el no métrico).

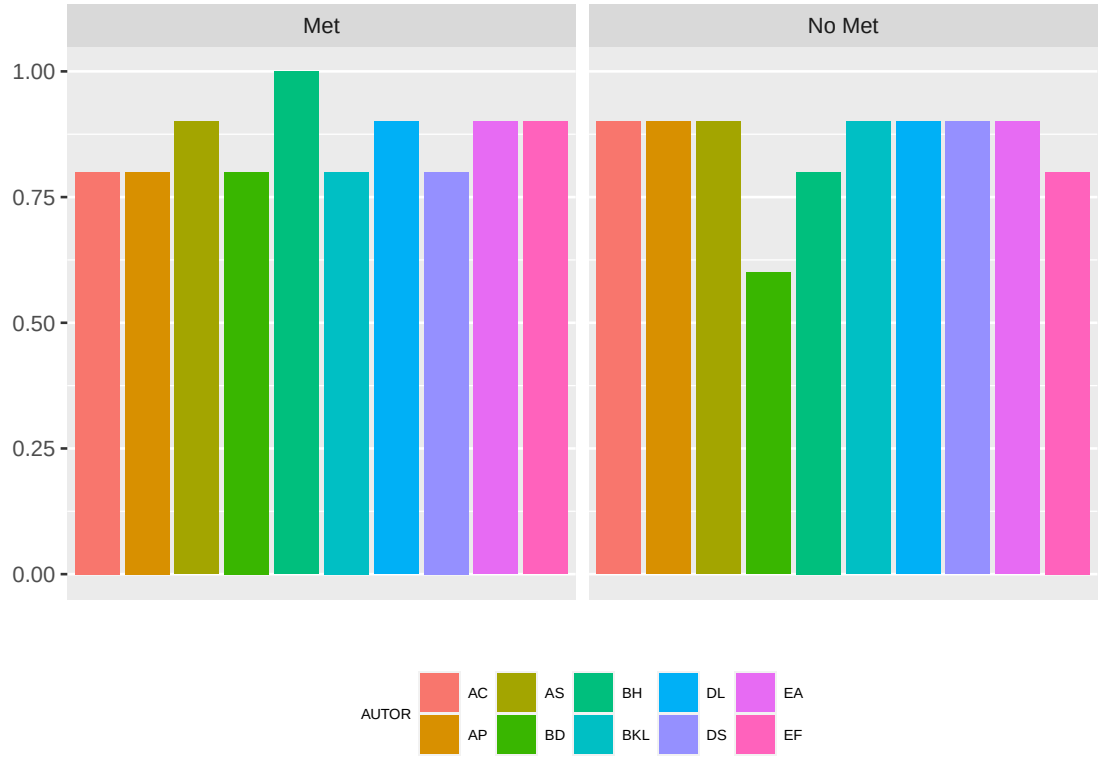
```
## Precisión por autores:
```

```
##              AC  AP  AS  BD  BH BKL  DL  DS  EA  EF
## MDS.Met     0.8 0.8 0.9 0.8 1.0 0.8 0.9 0.8 0.9 0.9
## MDS.No.met 0.9 0.9 0.9 0.6 0.8 0.9 0.9 0.9 0.9 0.8
```

El método métrico alcanza el 100 % de documentos clasificados correctamente para *Bernard Hickey*, mientras que los autores para los que el modelo es menos preciso (con un

80 %) son *Alan Crosby*, *Aaron Pressman*, *Brad Dorfman*, *Benjamin Kang Lim* y *Darren Schluettler*.

En el método no métrico la precisión mas alta es del 90 % y el autor para el que el modelo es menos preciso es *Brad Dorfman* con un 60 % de documentos clasificados correctamente.

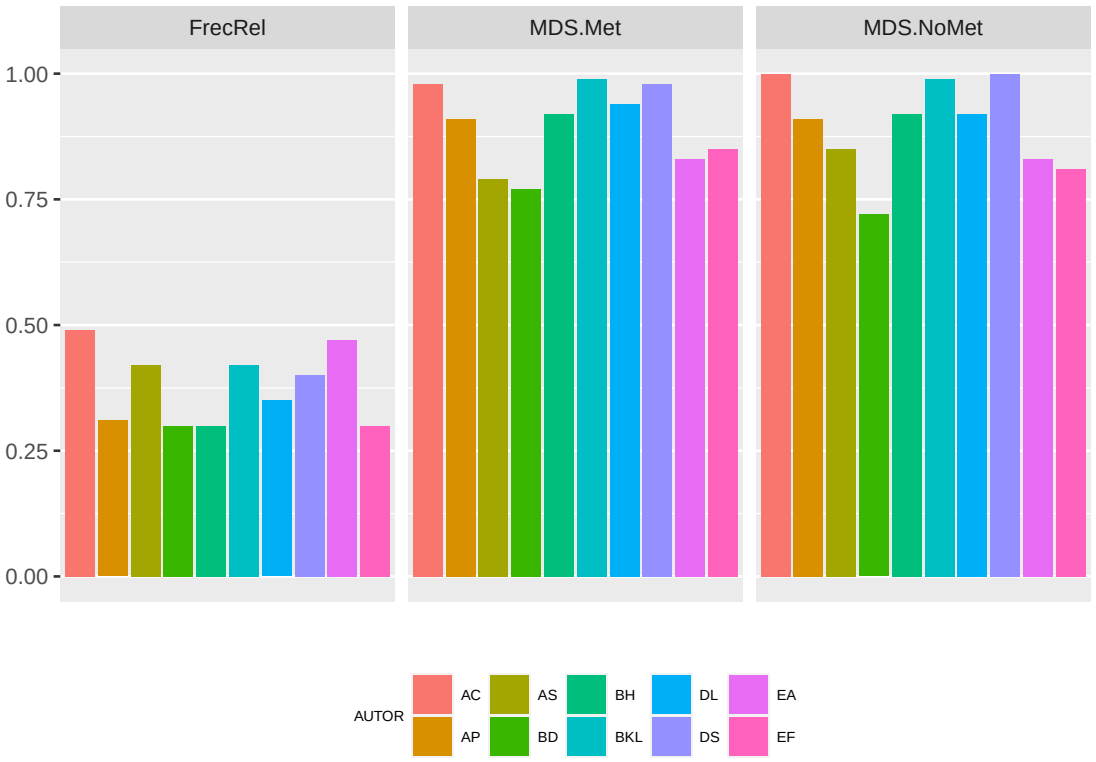



```
## Recuperación por autores:
```

	AC	AP	AS	BD	BH	BKL	DL	DS	EA	EF
MDS.Met	1.0	1	0.75	0.62	0.91	0.89	0.82	1	0.90	0.90
MDS.No.met	0.9	1	0.69	0.86	0.89	1.00	0.75	1	0.69	0.89

En cuanto al índice de recuperación, el método métrico alcanza el 100 % en 3 autores: *Alan Crosby*, *Aaron Pressman* y *Darren Schuettler*, mientras que el autor para el que el modelo es menos preciso es *Brad Dorfman* con un 62 %

En el método no métrico, se alcanza el 100 % de recuperación con *Aaron Pressman*, *Benjamin Kang Lim* y *Darren Schuettler*, mientras que el valor mas pequeño es compartido por *Alexander Smith* y *Eric Auchard* con un 69 %.


FrecRel
MDS.Met
MDS.NoMet
1.00
0.75
0.50
0.25
0.00
AUTOR
AC
AP
AS
BD
BH
BKL
DL
DS
EA
EF

Apéndice A

Apéndice: Implementación con R

La implementación con el lenguaje de programación R correspondiente al tratamiento y resultados mostrados en el capítulo 4 se encuentra en los siguientes enlaces de mi repositorio en gitHub:

- librerias Yfunciones.R
- codigo.R
- C50.zip

Bibliografía

Karmele Fernández Aguirre. Análisis textual: generación y aplicaciones. *Metodología de encuestas*, 5(1):55–66, 2003.

Mönica Bécue-Bertaut. *Minería de textos. Aplicación a preguntas abiertas en encuestas.* Cuadernos de Estadística, 2010.

Daniel Peña Sánchez de Rivera. *Estadística. Modelos y métodos. (1. Fundamentos).* Alianza Universidad Textos, 1997.

Daniel Peña Sánchez de Rivera. *Estadística. Modelos y métodos. (2. Modelos lineales y series temporales).* Alianza Universidad Textos, 1999.

Daniel Peña Sánchez de Rivera. *Análisis de Datos Multivariantes.* McGraw Hill, 2002.

José Manuel Ramírez Hurtado Flor María Guerrero Casas. El análisis de escalamiento multidimensional: una alternativa y un complemento a otras técnicas multivariantes. 03 2019.

J. C. Gower. Adding a point to vector diagrams in multivariate analysis. *Biometrika*, 55 (3):582–585, 1968.

Marcial Terrádez Gurrea. Frecuencias léxicas del español coloquial: Análisis cuantitativo y cualitativo. *Facultat de Filología, Universitat de València*, pages 37–42, 2001.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classifications*, (2):193–218, 1985.

F. Torres R. Gutiérrez, A. González. *Técnicas de Análisis de datos multivariable. Tratamiento computacional.* Alianza Universidad Textos, 1997.

Jorge M Santos and Mark Embrechts. On the use of the adjusted rand index as a metric for evaluating supervised classification. *Artificial Neural Networks - ICANN 2009*, pages 175–184, 2009.

Kari Torkkola. Linear discriminant analysis in document classification. *IEEE TextDM 2001*, 12 2001.