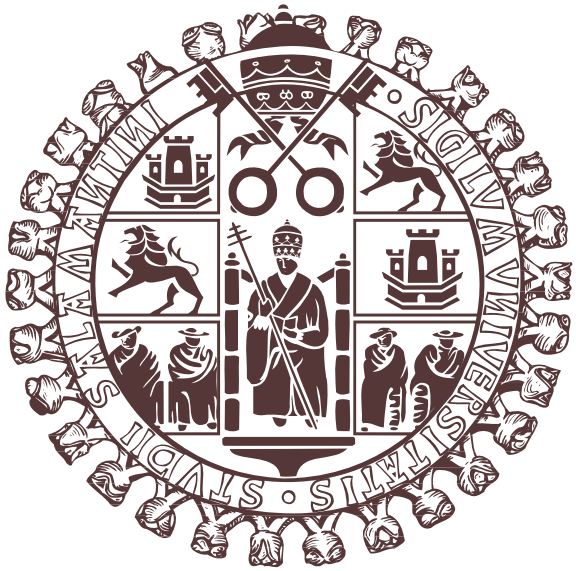
UNIVERSIDAD DE SALAMANCA

DEPARTAMENTO DE ESTADÍSTICA



BIPLOT LOGÍSTICO PARA DATOS NOMINALES Y ORDINALES

JULIO CÉSAR HERNÁNDEZ SÁNCHEZ

2016

2016

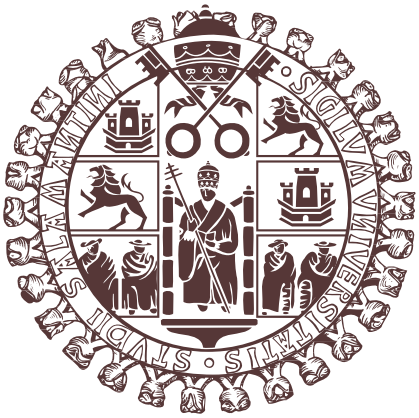
BIPLOT LOGÍSTICO PARA DATOS NOMINALES Y ORDINALES

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

Julio César Hernández Sánchez

Salamanca

2016



Universidad de Salamanca

Departamento de Estadística

JOSE LUIS VICENTE VILLARDÓN

Profesor titular del Departamento de Estadística de la Universidad de Salamanca

CERTIFICA:

Que D. Julio César Hernández Sánchez, Licenciado en Matemáticas, ha realizado en el Departamento de Estadística de la Universidad de Salamanca, bajo su dirección, el trabajo que para optar al Grado de Doctor, presenta con el título: ***Biplot logístico para datos nominales y ordinales***; y para que conste, firma el presente certificado en Salamanca, a 26 de Noviembre de 2015.

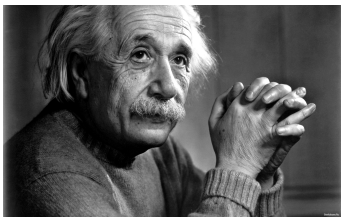
Fdo: José Luis Vicente Villardón

A la memoria de mi amigo
Antonio Jamardo Villamarín,
fallecido en el accidente del Alvia 04155
en Santiago de Compostela
el 24 de Julio de 2013.
Tú nos enseñaste a disfrutar de
cada momento de la vida y a sonreir.

I didn't fail the test. I just found 100 ways to do it wrong.

Benjamin Franklin

Agradecimientos



Try not to become a man of success but a man of value.

Albert Einstein

Mi agradecimiento más profundo y sincero a la persona que ha creído en mí y que me ayudo de tal forma que jamás le podré devolver tal gratitud, dedicación, profesionalidad y sabiduría, que es José Luis Vicente-Villardón, mi director de tésis y director del Departamento de Estadística de la Universidad de Salamanca. Contigo he aprendido muchas cosas de la profesión de estadístico y he compartido muchas charlas que nos han acercado hasta la amistad, que es el mejor final de esta aventura.

A la Doctora Mª Purificación Galindo Villardón, que fué la que me animó, empujó y alentó a comenzar con unos pasos que culminan en este estudio y a la cuál siempre estaré agradecido.

También quiero tener unas palabras de gratitud y agradecimiento a todos los profesores del Departamento de Estadística de la Universidad de Salamanca, por su cercanía, preocupación, disposición y sus valiosos consejos a lo largo de este trabajo.

A mi hermano Tomy, con el que me he reído tanto y he disfrutado tanto, que siempre llevo conmigo. Eres un ejemplo para mí y te admiro por tu inteligencia, por tu valía, por lo bien que nos haces sentir a los que estamos a tu lado y por tantos detalles del día a día que nos das y que nos encantan y nos hacen tener mucha más ilusión en todas las cosas.

A mi tío Antonio, por la fuerza que demuestra cada día y en la cuál debemos mirarnos ante las adversidades de la vida, una fuerza que me ha llevado a insistir y persistir en esta investigación, y que se lo debo en parte a él.

A mi tía Cruz y a mi tío Miguel, mis padrinos, que siempre se acuerdan de mí y los siento cerca a pesar de las distancias.

A mi tío Rafa a quien admiro muchísimo y con quien tan buenos momentos he vivido, y a mi tío Jesús que me ayudó en momentos decisivos y al que siempre estaré agradecido.

A mi abuela Teresa, a la que quiero y admiro por su grandísima lucha en la vida y por haberme tratado tan bien y haberme hecho sentir tan bien en su compañía y en la de Mari Tere durante tantos años.

A mis abuelos paternos, Felicidad, cuyo nombre indica la bondad, generosidad y alegría indescriptible de la persona que las lleva, e Hilario, un referente absoluto en la familia y un hombre adelantado al tiempo en el que vivió y al que le debemos gran parte de lo que tenemos.

A mi amigo Miguel, que siempre está ahí y que para mí es una de las personas más íntegras que conozco en todos los sentidos. Tu ayuda en los momentos difíciles ha sido fundamental y disfrutar de tus conocimientos, inteligencia y valía es algo que siempre recuerdo y recordaré.

No puedo olvidarme de tantos amigos que han estado cerca de mí en esta aventura y en otras muchas que nos han acercado y que no podré olvidar, Loli, Pablo, Carmen, Amor y Pedro. Os quiero un montón.

También me quiero acordar de Miguel Angel Gimeno y de todo el equipo de Medialab de la Universidad de Salamanca, por su apoyo en este proyecto, su

amabilidad e implicación.

Tengo un especial recuerdo de mi amigo Antonio Valadés, una persona entrañable y buena con la que he pasado menos tiempo del que me hubiera gustado charlando en la naturaleza y disfrutando de momentos inolvidables.

Y como no, a Nuria, que tan pendiente ha estado durante este trayecto investigador y muchos otros y que ha hecho que la vida de nuestra familia haya sido muy divertida.

A mi cuñado Julio, que ha estado siempre pendiente de si necesitaba algo con el inglés y de mejorar la redacción de los artículos. Muchas gracias por tu ayuda y por tu cercanía.

Quiero tener unas palabras de cariño para mis cuñados y cuñadas, Victoria, Chus, Llorente y Estefanía. Habéis creado un entorno con el que nunca hubiera soñado y vuestra amistad me ha hecho crecer como persona y aprender a convivir mejor, apreciando unos valores fundamentales para caminar a lo largo de los años.

Y a mis suegros Graci y Fernando, que son unas personas buenísimas, generosas, cariñosas, atentas y con los que siempre me he sentido como uno más de la familia.

También me acuerdo de mis sobrinas Adriana y Mencía, y de mi ahijado Alvar, a los que un día contaré esta historia y explicaré por qué, como dice Alvar, "parrino" siempre está estudiando.

Y qué decir de mis padres Tomás y Flor. Para vosotros no tengo palabras, ni hojas, ni libros en los que escribir cuánto os tengo que agradecer por todo el amor y el cariño que me habeis dado, por tantas oportunidades brindadas, por tanto sacrificio para poder ofrecérmelas, por tanta confianza en mí, por tanta compañía en las vivencias de tantos años, por darme tanto, y por inculcarme valores como el esfuerzo, el respeto y la justicia ante los demás.

Nada de esto hubiera sido posible sin el apoyo incondicional de la persona a la que más quiero en este mundo, que es mi mujer. Isa, mil gracias por tu amor, por tu cariño, por tu sinceridad y apoyo en todo momento, por tu eterna sonrisa que

hace que vivir contigo sea maravilloso cada día y por creer tanto en mí. Eres la luz de mi vida.

Y por supuesto, a Julia y Teresa, mis queridas hijas. Cuánto tiempo os debo, cuántas horas de jugar con vosotras sacrificadas por este proyecto, cuántos "ahora no hijas, que tengo que trabajar en la tesis...". Espero que me perdoneis y que algún día entendáis el por qué y el para qué de todo esto. Sois lo mejor de mi vida.

Índice general

Capítulo 1

Introducción

"True science teaches, above all, to doubt and to be ignorant."

Miguel de Unamuno

Hay numerosas técnicas adecuadas para trabajar con datos nominales, algunas de las cuales analizan el problema que supone este tipo de datos desde el punto de vista del Análisis Factorial (FA), cuyo objetivo es obtener factores latentes que expliquen la correlación entre las variables. Otras inciden en algunos tipos de aproximaciones no paramétricas para explorar las similaridades entre los individuos (Análisis de Coordenadas Principales (PCoA) o Escalamiento Multidimensional (MS)), pero existen pocas técnicas exploratorias generales que permitan la representación simultánea de individuos y variables,

excepto el Análisis de Correspondencias Múltiple (MCA)[1], basado en la distancia chi-cuadrado, que no siempre es la más adecuada para describir similaridades entre individuos y correlaciones entre variables.

La interpretación del MCA es más compleja que el Análisis de Correspondencias y se han propuesto algunos métodos para simplificar el problema que se presenta al trabajar con datos categóricos, como por ejemplo utilizar una métrica distinta que la proporcionada por la χ^2, como la distancia de Hellinger(ver Escofier [1978] y Rao [1995]). El Análisis de Correspondencias Conjunto (JCA), que ajusta los bloques fuera de la diagonal de la matriz de Burt [Greenacre, 1993], puede considerarse como una aproximación diferente de este enfoque.

En la actualidad hay una continua investigación en este tipo de métodos y de metodologías [Greenacre, 2012], las cuales tratan de avanzar en el desarrollo de nuevas propuestas que resuelvan situaciones con diferentes conjuntos y tipos de datos (ver Cecere y col. [2013], Zerrin y Greenacre [2011] y Greenacre y Groenen [2013]), incluso para las herramientas más extendidas y adecuadas en el tratamiento de datos categóricos, como el Análisis de Correspondencias (CA), en el que todavía es necesario mejorar la representación gráfica de los casos y las variables para facilitar la interpretación de los resultados [Greenacre, 2013] y proporcionar nuevas formas de visualizar las interacciones entre los puntos fila y columna [Gower y col., 2010].

El CA se ha comparado con otros métodos con el objetivo de representar tablas de contingencia (Cuadras y Greenacre [2012]), desarrollando una versión paramétrica que engloba todas ellas (Cuadras y Cuadras [2006] y Cuadras y Cuadras

[1]Este procedimiento se usa para detectar estructuras en un conjunto de datos categóricos, representando los datos como puntos en un espacio euclídeo de baja dimensión, el cuál es suficientemente conocido. Pueden consultarse más detalles en Benzécri [1973] y Greenacre y Blasius [2006]. También se la conoce a esta técnica como Análisis de homogeneidad [Gifi, 1990], la cuál busca para un conjunto de variables categóricas unos valores de escala óptimos.

[2011]). Publicaciones recientes describen el CA junto con técnicas relacionadas [Greenacre, 2008] y con los métodos biplot (Greenacre [2010] y Gower y col. [2011]).

Los métodos biplot [Gabriel, 1971] estudian una representación gráfica conjunta de I filas y J columnas de una matriz de datos X en un espacio de dimensión reducida, en el cuál las filas representan a los individuos, objetos o muestras, y las columnas a las variables medidas sobre ellos, y se están convirtiendo en técnicas muy populares para el análisis multivariante de datos.

Los métodos biplot clásicos están íntimamente relacionados con el Análisis de Componentes Principales (PCA) y con el FA, que son dos métodos muy conocidos, los cuales están todavía evolucionando [Browne y McNicholas, 2013], aún cuando han sido y son utilizados durante más de cien años, y de hecho existen representaciones gráficas del PCA y del FA que se usan para obtener combinaciones lineales que maximizan la variabilidad total. Recientes estudios, como los de Nieto y col. [2014], proponen metodologías gráficas basadas en intervalos de confianza bootstrap para los parámetros definidos por los marcadores característicos del biplot, cuyo objetivo es el de profundizar en el estudio de medidas de bondad de ajuste y de las relaciones entre variables, asi como su calidad de representación.

Las técnicas de representación gráfica, como son los biplots, que se utilizan para visualizar el contenido de las matrices de datos o de modelos asociados a dichos datos son populares en la literatura científica, como puede apreciarse en Scrucca [2013], e incluso este tipo de análisis y herramientas están muy extendidas para estudiar datos de ensayos multi-ambiente [Frutos y col., 2013].

En la práctica, el ajuste de un biplot se produce o bien mediante la Descomposición en Valores Singulares (SVD) de una matriz de datos o bien llevando a cabo lo que se conoce como un procedimiento de regresiones alternadas [Gabriel y Zamir, 1979]. Esta aproximación se trata esencialmente de un algoritmo de mínimos cuadrados alternados, equivalente a un algoritmo-EM cuando puede considerarse que los datos siguen una distribución normal. Jongman y col. [1987] ajustan el biplot mediante un procedimiento que alterna una regresión y una calibración, lo

cuál es fundamentalmente un método de regresiones alternadas.

Si trabajamos con conjuntos de datos cuyas distribuciones pertenecen a la familia exponencial, Gabriel [1998] describe la "regresión bilineal" como un método para estimar los parámetros del biplot, pero el procedimiento no se ha llegado a implementar y las propiedades geométricas de las representaciones a las que dá lugar nunca se han estudiado. De Leeuw [2006] propone un PCA para datos binarios, basado en un proceso alternado en el que cada iteración se lleva a cabo utilizando lo que él llama mayorización iterativa y Lee y col. [2010] lo generalizan para matrices con gran cantidad de datos faltantes, pero ninguno de ellos describe la representación biplot para dicho tipo de datos. Vicente-Villardón y col. [2006] proponen una representación basada en respuestas logísticas llamándolo "Biplot Logístico", que es lineal, estudian la geometría de este tipo de biplots y utilizan un procedimiento de estimación que es ligeramente distinto del método de Gabriel. Una versión heurística del mismo para grandes matrices de datos en el cuál las puntuaciones de los individuos se calculan en un procedimiento externo se describe en Demey y col. [2008]. Dicho método se llama "Biplot Logístico Externo". Los biplots logísticos para datos binarios se han aplicado satisfactoriamente en diversos conjuntos de datos, como por ejemplo Gallego-Álvarez y Vicente-Villardón [2012], Vicente-Galindo y col. [2011] o Demey y col. [2008].

Existen varios métodos posibles para realizar la estimación de los parámetros de un biplot:

- Regresiones Alternadas Generalizadas e Interpolaciones. (Máxima Verosimilitud Conjunta, Gabriel [1998],Vicente-Villardón y col. [2006]).
- Máxima verosimilitud marginal (Similar a la Teoría de la Respuesta al Item, Baker [1992],Bock y Aitkin [1981],Chalmers [2012]).
- Biplots Logísticos Externos: Aproximación heurística para grandes matrices de datos. (Ajustes logísticos sobre las coordenadas principales, Demey y col. [2008]).

En el contexto de los biplots logísticos binarios los dos primeros procedimientos son particularmente útiles cuando el número de individuos es mayor que el número de variables, siendo el segundo de ellos más estable para situaciones con un gran número de individuos. El tercer método es más útil cuando el número de ítems o variables es mayor que el de objetos, aunque también se puede aplicar en cualquier caso. En este trabajo se ha elegido una versión del segundo método. La estimación de los parámetros relativos a las variables se han calculado utilizando un algoritmo desarrollado para un uso general, mediante regresiones logísticas estándar realizadas utilizando las puntuaciones proporcionadas por diversos métodos, como mirt o el Análisis de Coordenadas Principales.

Cuando el conjunto de datos contiene variables con más de dos categorias, los biplots lineales e incluso los biplots logísticos binarios no son adecuados y técnicas como el MCA, el Análisis de Rasgos Latentes (LTA) o la Teoría de Respuesta al Ítem (IRT) serían más convenientes para el tratamiento de este tipo de variables. Recientemente Hernández Sanchez y Vicente-Villardón [2013] han desarrollado lo que llaman "Biplot Logístico Nominal (NLB)" como un procedimiento que por un lado reduce la dimensión del espacio de partida, explicando la correlación existente entre variables nominales, y por otro se utiliza como una técnica exploratoria. Los biplots logísticos nominales representan las filas de la matriz de datos como puntos en una representación correspondiente a un espacio de dimensión reducida(generalmente 2 ó 3) y las variables como regiones de predicción(polígonos convexos), de la misma forma que se hace en Gower y Hand [1996] para el MCA.

La principal ventaja del NLB es que la interpretación del biplot se hace en términos de distancias, de tal forma que para cada individuo la categoría que se predice en una variable es la más cercana a él en el biplot. De esta forma, este tipo de biplots extienden tanto al Análisis de Correspondencias Múltiples como al Análisis de Rasgos Latentes, en el sentido de que provee una representación gráfica para el LTA similar a la que se obtiene en MCA.

En el caso del Análisis de Correspondencias Múltiples los puntos que denotan

cada categoria de una variable son los primeros que se calculan, y con ellos las regiones de predicción se obtienen como regiones de un diagrama de Voronoi utilizando una transformación adecuada considerando dichos puntos como los puntos clave para construir el diagrama. En este estudio se va a proponer que las regiones de predicción se obtengan primero mediante una regresión logística nominal que define teselaciones en el espacio. El problema es entonces encontrar el diagrama de Voronoi más cercano a la "teselación logística" y calcular un conjunto de generadores para tal diagrama, que serian los puntos categoría de cada variable. De esta forma establecemos un diálogo entre la Estadística y la Geometría Computacional, que en este ámbito es un aspecto novedoso. La ventaja de proceder así es que la interpretación del biplot es en términos de distancias como hemos comentado.

Cuando los datos contienen variables ordinales, los biplots lineales, binarios o los logísticos nominales tampoco son adecuados, situación en la cuál, el Análisis de Componentes Principales Categórico (CATPCA) ó la IRT para variables ordinales serían propuestas más válidas. Lo que haremos es extender el concepto de biplot a aquellas situaciones en las que aparezcan este tipo de datos, resultando un método que llamaremos "Biplot Logístico Ordinal (OLB)". Las puntuaciones de las filas se calculan teniendo en cuenta el supuesto de que tengan superficies de respuesta logística ordinales sobre las dimensiones consideradas y los parámetros columna producen superficies de respuesta logística que, proyectadas sobre el espacio reducido por las puntuaciones de las filas definen un biplot lineal. Se utilizará un modelo de odds proporcionales, obteniendo así un modelo multidimensional conocido como modelo de respuesta graduada en la literatura del IRT. Estudiaremos la geometría de tales representaciones e implementaremos algoritmos computacionales para la estimación de los parámetros y de las direcciones de la predicción. El OLB extiende tanto CATPCA como IRT puesto que ofrece una representación gráfica para IRT parecida al biplot correspondiente al CATPCA.

Esta tésis se estructura en una serie de capítulos que describimos brevemente. En el segundo se plantean los objetivos del trabajo tanto generales como específi-

cos. En el tercero se plasman las principales características de los biplots lineales clásicos, centrándose el cuarto en una revisión sintética de los biplots de variables binarias. El capítulo 5 extiende el concepto de biplot logístico al caso de variables nominales, describiendo la metodología de construcción de este tipo de biplots y algunas aplicaciones con conjuntos de datos reales, aprovechando uno de los ejemplos para contrastar esta técnica con el MCA y analizar sus diferencias y particularidades. El sexto apartado se centra en el caso de variables ordinales y de cómo construir los biplots asociados. El capítulo 7 se centra en la estimación de los parámetros de los diferentes modelos planteados y en el tratamiento de la separación en regresión logística. Los capítulos 8 y 9 describen dos de los paquetes de R que se han implementado para la construcción de Biplots Logísticos Nominales y Ordinales respectivamente, ilustrando su funcionamiento mediante la utilización con conjuntos de datos reales. El capítulo 10 conjuga las características de los biplots analizados con el objetivo de crear un Biplot Logístico para datos categóricos, completando así uno de los principales objetivos de esta investigación. Además, se describe el último de los paquetes de R desarrollados al efecto. Por último, los capítulos 11 y 12 concluyen este estudio con una discusión de los resultados y unas conclusiones finales.

Capítulo 2

Objetivos de la investigación

I really believed that I was on the right track, but that did not mean that I would necessarily reach my goal.

– *Andrew Wiles*

El principal objetivo es el estudio de los métodos biplot cuando el conjunto de datos está formado por variables categóricas, y determinar sus características y aplicaciones.

Los objetivos específicos que se postulan son los siguientes:

- Descripción de las características principales de los biplots lineales y logísticos, como punto de partida para desarrollar el caso nominal, detallando en especial el caso de variables binarias.
- Descripción y adaptación del algoritmo EM a nuestro plantemiento como herramienta necesaria en la construcción de los algoritmos para el cálculo de los parámetros del modelo.

- Extender el biplot logistico binario al caso de variables nominales, estudiando el modelo y sus principales características geométricas. Llamaremos a esta técnica novedosa "Biplot Logístico Nominal (NLB)".

- Investigar cómo la rama conocida como Geometria Computacional puede ayudar a resolver problemas estadísticos relacionados con los biplots, interconectando ambas disciplinas científicas en el análisis multivariante y utilizando los resultados para construir representaciones gráficas más sencillas de interpretar.

- Implementación tanto de los algoritmos para la estimación de los parámetros del modelo, los cuales deben ser capaces de funcionar incluso en el caso del problema de separación en regresión logistica, como de una herramienta que permita la construcción de este tipo de biplots y esté disponible al público en general.

- Presentar una visión diferente en la interpretación gráfica de los mapas, aportando un punto de vista también distinto en la forma de leer los biplots respecto del Análisis de Correspondencias Múltiples, que se basa en perfiles fila y columna, lo cuál introduce dificultades en dichas interpretaciones. Lo que se intentará buscar son interpretaciones basadas en distancias y no en proyecciones.

- Ilustrar con un ejemplo estudiado convenientemente en la literatura una comparativa del resultado de la aplicación de algunas tecnicas multivariantes respecto al NLB.

- Mostrar las principales diferencias entre el NLB y el MCA a través del análisis de un estudio de algunas características demográficas y relativas al mercado de trabajo de las personas que han leído una tésis doctoral en la región de Castilla-León (España), utilizando para ello la "Encuesta sobre recursos humanos en ciencia y tecnología" que elabora el Instituto Nacional de

Estadística.

- Extender la construcción de biplots al caso de variables ordinales, mediante la utilización del modelo de odds proporcionales, obteniendo asi un modelo multidimensional, conocido como modelo de respuesta graduada en la literatura de la Teoria de la Respuesta al Ítem. El método resultante se llamará "Biplot Logístico Ordinal (OLB)".

- Estudiar la geometria del "Biplot Logístico Ordinal", así como proporcionar una herramienta software que permita su cálculo mediante los algoritmos necesarios para ello, que deberán estimar los parámetros del modelo y determinar las direcciones de predicción en cada caso.

- Aplicar la técnica del OLB al estudio de la satisfacción sobre el empleo de las personas que tienen el título de doctorado en España, mediante los datos que proporciona la encuesta citada anteriormente.

- Adaptar los algoritmos diseñados en los casos anteriores para poder trabajar en el mismo conjunto de datos con variables nominales y ordinales a la vez, de tal forma que quedarían caracterizados los biplots para cualquier tipo de variable categórica, y que esta sea una herramienta software pública.

Capítulo 3

Biplots Clásicos Lineales y la Descomposición en Valores Singulares

Statistics is the grammar of science.
– Karl Pearson

La investigación sobre conjuntos de datos en la que se trata de comprender la existencia de patrones y relaciones entre individuos y variables o entre estas últimas se ha visto fortalecida por el empeño de la comunidad científica en desarrollar los métodos gráficos. La representación gráfica de dichos conjuntos de datos de la forma más exacta posible ha sido el objeto de numerosos trabajos en el pasado y en la actualidad. Puesto que los humanos sólo pueden visualizar objetos como mucho en tres dimensiones, el foco de interés se centra en una representación de los datos en una, dos o tres dimensiones.

Gabriel [1971] propuso una representación conjunta de invididuos y variables, a la cuál llamó biplot, en la que cada fila y cada columna de la matriz de datos se plasmaban en el plano como vectores que parten del origen. Estos vectores son tales que si efectuamos el producto escalar de cada vector que representa una fila por cada vector que representa una columna el resultado es una aproximación al valor correspondiente a esa fila-columna en la matriz de datos original.

La principal debilidad de los biplots tradicionales es la dificultad de visualizar los productos escalares, y por tanto las aproximaciones de los elementos de la matriz de datos en el biplot. Para solventar esta cuestión Gower y Hand [1996] propusieron que en el biplot las variables se representaran mediante ejes calibrados de tal forma que las mediciones aproximadas de un elemento de la muestra se pudieran leer fuera de los ejes como en los diagramas de dispersión. Estos ejes se construyen prolongando los vectores de las variables en ambas direcciones. Los biplots construidos de esta forma pueden verse como versiones multivariantes de los diagramas de dispersión ordinarios [Gower y Hand, 1996]. Por otra parte, las observaciones o filas serían representadas mediante puntos en el gráfico.

Cualquier técnica de reducción de la dimensión se utiliza normalmente para representar los elementos de la muestra como puntos, algunos de los cuales son métodos muy conocidos como el PCA (Pearson [1901]; Hotelling [1933]) o el Análisis Canónico (CVA) [Hotelling, 1935]. La situación o posicionamiento de los ejes depende del método elegido para situar los puntos o individuos. El PCA biplot[1] tiene ejes lineales para situar los puntos que han sido calculados mediante PCA [Gower y Hand, 1996], el biplot de regresión [Gower y Hand, 1996] produce ejes lineales aproximados para cualquier ordenación de los elementos de la muestra, y se podrían citar otros tipos de biplots con características particulares.

Autores como Le Roux y Gardner [2005] muestran y citan numerosos ejemplos del uso de este tipo de biplots lineales en diversas ramas del conocimiento, como la arqueología, la agricultura, la educación, la gestión financiera, la mineralogía, la

[1]Lo describiremos brevemente en la sección 3.1.

cefalometría [Naidoo y col., 2006] o la química [Alves y col., 2005].

3.1. Biplots lineales basados en regresiones alternadas

Sea $\mathbf{X}_{I\times J}$ una matriz con información sobre J variables(continuas) medidas sobre I individuos. Un biplot S-dimensional es una representación gráfica de la matriz $\mathbf{X}$ mediante marcadores(puntos o vectores) $\mathbf{a}_1,\ldots,\mathbf{a}_I$ para sus filas y marcadores $\mathbf{b}_1,\ldots,\mathbf{b}_J$ para sus columnas, de tal forma que el producto $\mathbf{a}_i'\mathbf{b}_j$ aproxime el elemento x_{ij} lo mejor posible. Ordenando dichos marcadores como vectores fila en dos matrices que llamaremos $\mathbf{A}$ y $\mathbf{B}$, la aproximación de $\mathbf{X}$ se puede escribir como $\mathbf{X}\approx\mathbf{AB}'$. Aunque el biplot clásico es bien conocido, vamos a describirlo brevemente, en términos de las regresiones alternadas, lo cuál está en concordancia con nuestras propuestas detalladas más adelante.

El camino más usual de obtener un biplot es mediante la descomposición en valores singulares de una matriz. Sea $R = rango(\mathbf{X})$, entonces existe una factorización en la siguiente forma:

$$\mathbf{X} = \mathbf{U}\Lambda\mathbf{V}' = \sum_{r=1}^{R}\lambda_r\mathbf{u}_r\mathbf{v}_r', \tag{3.1}$$

donde $\mathbf{U}$ es una matriz unitaria de dimensión $I\times R$, $\mathbf{\Lambda}$ es la matriz diagonal $R\times R$ con números reales no negativos en la diagonal, y $\mathbf{V}$ una matriz unitaria de $J\times R$. Tal factorización se conoce con el nombre de Descomposición en Valores Singulares(SVD) de $\mathbf{X}$. Los valores de la diagonal de $\mathbf{\Lambda}$, λ_r, se llaman valores singulares de $\mathbf{X}$, y están ordenados de forma decreciente, y las columnas $\mathbf{u}_r$ y $\mathbf{v}_r$ de $\mathbf{U}$ y $\mathbf{V}$ son los vectores singulares por la izquierda y por la derecha respectivamente. Las *Descomposiciones en Valores Singulares* están muy relacionadas con las *Descomposiciones en Valores Propios*, las columnas de $\mathbf{U}$ son los vectores propios de $\mathbf{XX}'$, las columnas de $\mathbf{V}$ los vectores propios de $\mathbf{X}'\mathbf{X}$ y los elementos de la diagonal

de $\mathbf{\Lambda}$ son las raices cuadradas de los valores propios no nulos de ambas matrices (que son los mismos).

Es conocido que la mejor aproximación de rango S para $\mathbf{X}$ viene dada por sus primeros S vectores y valores propios:

$$\mathbf{X} \cong \sum_{s=1}^{S} \lambda_s \mathbf{u}_s \mathbf{v}_s' = \mathbf{U}_{(S)} \Lambda_{(S)} \mathbf{V}_{(S)}'. \tag{3.2}$$

Partiendo de la SVD es sencillo obtener una factorización en la forma Biplot con las restricciones deseadas tomando:

$$\mathbf{A} = \mathbf{U}_{(S)} \Lambda_{(S)}^{\gamma}, \quad \mathbf{B} = \mathbf{V}_{(S)} \Lambda_{(S)}^{(1-\gamma)}, \tag{3.3}$$

con $0 \leq \gamma \leq 1$, como coordenadas de las filas y columnas respectivamente. A esta configuración o estructura nos referiremos en lo sucesivo con el nombre de Biplot de Componentes Principales, PCA-Biplot o Biplot Clásico. Por ejemplo, para $\gamma = 1$, $\mathbf{A}$ son las coordenadas de los individuos sobre las componentes principales y $\mathbf{B}$ son los vectores propios de la matriz de covarianzas.

Existe otro procedimiento para obtener biplots mediante regresiones alternadas. Si consideramos los marcadores fila $\mathbf{A}$ como fijos, los marcadores columna se pueden calcular con regresiones:

$$\mathbf{B}' = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{X}. \tag{3.4}$$

De la misma forma, fijando $\mathbf{B}$, $\mathbf{A}$ se obtienen como:

$$\mathbf{A}' = (\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'\mathbf{X}'. \tag{3.5}$$

Alternando los pasos (3.4) y (3.5) el producto converge al mismo subespacio generado por la SVD. El algoritmo puede ser completado con un paso de ortogonalización para asegurar la unicidad de la solución. Las regresiones en (3.4) y (3.5) se pueden separar para cada fila y columna de la matriz inicial de datos. Este proceso simétrico se utiliza usualmente para ajustar modelos bilineales(o bi-aditivos) que asignan la misma importancia a las filas y a las columnas. Para una matriz que

contiene información de individuos y variables, normalmente las filas y columnas no juegan un rol simétrico, a pesar de lo cuál el algoritmo es válido igualmente y se interpreta como un proceso en dos etapas en el que se alternan una regresión y una interpolación o calibración. La etapa en la que se ejecuta la regresión en esencia lo que hace es ajustar para cada columna(variable) una regresión lineal de forma separada, y en la etapa de interpolación, utilizando los marcadores columna como referencia, se interpola cada individuo. La geometría de esta etapa de interpolación está descrita y puede consultarse en Gower y Hand [1996].

Podemos encontrarnos, por tanto, con el modelo en dimensión reducida(S-dimensional) en ocasiones presentado como:

$$\mathbf{X} = \mathbf{1}_I \mathbf{b}_0' + \mathbf{A}\mathbf{B}' + \mathbf{E} \tag{3.6}$$

donde $\mathbf{b}_0'$ es un vector de constantes, normalmente el vector columna de medias($\mathbf{b}_0 = \bar{\mathbf{x}}$), $\mathbf{A}$ y $\mathbf{B}$ son matrices de rangos S con I y J columnas respectivamente, y $\mathbf{E}$ es una matriz $I \times J$ de errores o residuos. La aproximación en rango reducido de la matriz de datos centrada (con los valores esperados) puede escribirse como:

$$\tilde{\mathbf{X}} = E\left[\mathbf{X} - \mathbf{1}_I \mathbf{b}_0'\right] = \mathbf{A}\mathbf{B}' \tag{3.7}$$

ó

$$E\left[\mathbf{X}\right] = \mathbf{1}_I \mathbf{b}_0' + \mathbf{A}\mathbf{B}', \tag{3.8}$$

que se obtiene normalmente, como hemos comentado, de la Descomposición en Valores Singulares(SVD). Por este motivo está íntimamente relacionada con sus Componentes Principales, y se llama Biplot [Gabriel, 1971] porque se puede utilizar simultáneamente para dibujar los individuos y las variables utilizando las filas de $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_I)'$ y $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_J)'$ como marcadores, de tal forma que el producto escalar $\mathbf{a}_i'\mathbf{b}_j$ aproxima al elemento $\tilde{x}_{ij}$ de la mejor manera posible en términos de cercanía.

Si consideramos los marcadores fila $\mathbf{A}$ como fijos y la matriz de datos previamente centrada, los marcadores columna se pueden calcular mediante una regresión sin término independiente:

$$\mathbf{B}' = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'(\mathbf{X} - \mathbf{1}_I\bar{\mathbf{x}}'). \tag{3.9}$$

Del mismo modo, fijando $\mathbf{B}$, $\mathbf{A}$ se obtiene:

$$\mathbf{A}' = (\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'(\mathbf{X} - \mathbf{1}_I\bar{\mathbf{x}}')'. \tag{3.10}$$

Alternando los pasos (3.9) y (3.10), el citado producto converge, como hemos comentado, al mismo subespacio generado por la (SVD) de la matriz de datos centrada.

Este procedimiento es de alguna manera lo que se conoce como Algoritmo-EM, en el cuál el paso de la regresión se correspondería con la etapa de maximización y el paso de interpolar con la etapa de calcular la esperanza. Una extensión para matrices de frecuencia puede encontrarse en Gabriel y col. [1998]. En resumen, los valores esperados de la matriz de datos original se obtienen en el biplot utilizando un producto escalar sencillo, es decir, proyectando el punto $\mathbf{a}_i$ sobre la dirección definida por $\mathbf{b}_j$. Este es el motivo por el que los marcadores fila se representan normalmente como puntos y los marcadores columna como vectores(o también llamados ejes biplot [Gower y Hand, 1996]).

En cuanto a las medidas de la calidad de representación y bondad de ajuste del biplot de componentes principales, pueden consultarse las referencias [Gabriel, 2002], [Galindo, 1986], [Vicente-Villardón, 1992], [Gower y Hand, 1996], [la Grange y col., 2009] y [Brand, 2003] para una descripción completa de las mismas.

3.2. Geometría de los biplots de regresión

Como hemos comentado, el método que posiciona los puntos en el biplot condiciona el posicionamiento de los ejes. En esencia los biplots de regresión se corres-

ponden con el PCA-Biplot para puntos calculados mediante PCoA basado en las disimilaridades utilizando la distancia Pitagórica.

Gower y Hand [1996] describen la geometría de los biplots para subespacios lineales, pero son Vicente-Villardón y col. [2006] los que desarrollan y presentan la geometría de los biplots de regresión.

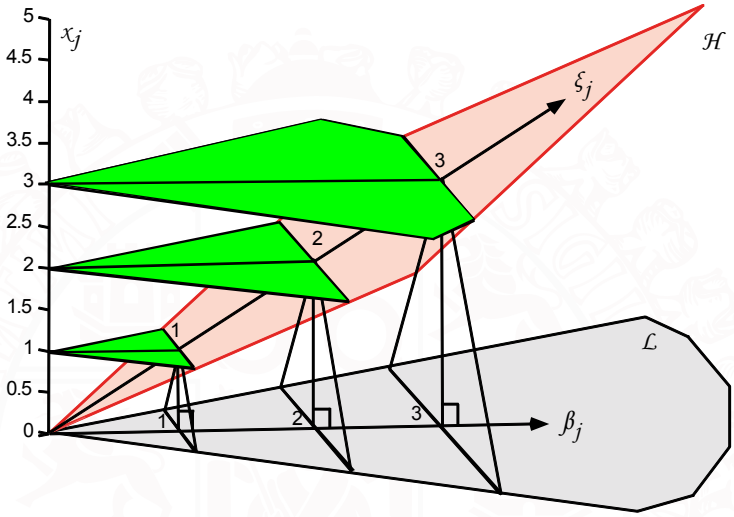


Figura 3.1: Geometría del Biplot Clásico ajustado con modelos de regresión lineal. Tomado de [Vicente-Villardón y col., 2006].

La figura 3.1 ilustra de forma resumida dicha geometría. Se trata de encontrar la dirección $\boldsymbol{\beta}_j$ en el espacio $\boldsymbol{\mathcal{L}}$ generado por las dos columnas de $\mathbf{A}$, de tal forma que las proyecciones de los marcadores de $\mathbf{A}$ sobre esa dirección predigan de la mejor manera posible los valores de la variable j. Es decir, para la j-ésima columna de $\mathbf{X}$, $\boldsymbol{x}_j \approx \mathbf{A}\boldsymbol{\beta}_j$. Dicha dirección queda determinada por los marcadores de la j-ésima columna.

Si llamamos $\boldsymbol{\mathcal{H}}$ al plano de regresión ajustado para la variable j, los puntos de dicho plano que predicen un valor fijo de dicha variable $\boldsymbol{x}_j$ vienen dados por la línea recta resultante de la intersección entre el plano $\boldsymbol{\mathcal{H}}$ y el plano paralelo a

$\boldsymbol{\mathcal{L}}$ que pasa por ese valor fijo. De esta forma, diferentes valores van asociados a lineas paralelas distintas en $\boldsymbol{\mathcal{H}}$. Considerando la recta $\boldsymbol{\xi}_j$ que es ortogonal a todas ellas, esta será el eje de referencia que se utilizará para la predicción. A su vez, los puntos de $\boldsymbol{\mathcal{L}}$ que predicen distintos valores están también en líneas rectas y la proyección de $\boldsymbol{\xi}_j$ sobre $\boldsymbol{\mathcal{L}}$ es perpendicular a estas rectas, que es justamente $\boldsymbol{\beta}_j$.

Los ejes del biplot se pueden completar con escalas que ayuden a predecir los valores de cada individuo de la matriz de datos. Para encontrar el punto en la dirección del biplot que predice un valor fijo μ de la variable observada cuando se proyecta el punto correspondiente a un individuo, tenemos que buscar aquel punto (x, y) que descansa en el eje del biplot, es decir, aquel que verifica

$$y = \frac{b_{j2}}{b_{j1}} x$$

y

$$\mu = b_{j0} + b_{j1}x + b_{j2}y$$

Resolviendo para x y y, obtenemos

$$x = (\mu - b_{j0}) \frac{b_{j1}}{b_{j1}^2 + b_{j2}^2}$$

y

$$y = (\mu - b_{j0}) \frac{b_{j2}}{b_{j1}^2 + b_{j2}^2}$$

ó

$$(x, y) = (\mu - b_{j0}) \frac{\mathbf{b}_j}{\mathbf{b}_j' \mathbf{b}_j}$$

Por tanto, el marcador unitario para la j-ésima variable se calcula dividiendo las coordenadas de su correspondiente marcador por su longitud al cuadrado, y de esta forma, etiquetando diversos puntos para valores específicos de μ se obtiene una escala de referencia. Si los datos están centrados, entonces $b_{j0} = 0$ y las etiquetas se pueden determinar sumando la media al valor de μ $(\mu + \bar{x}_j)$. La representación resultante será como la que se aprecia en la figura 3.2.

La bondad de ajuste se mide mediante los coeficientes de determinación R_j^2 calculados para las regresiones, los cuales son interpretados como medidas de la "calidad de representación" de las variables.

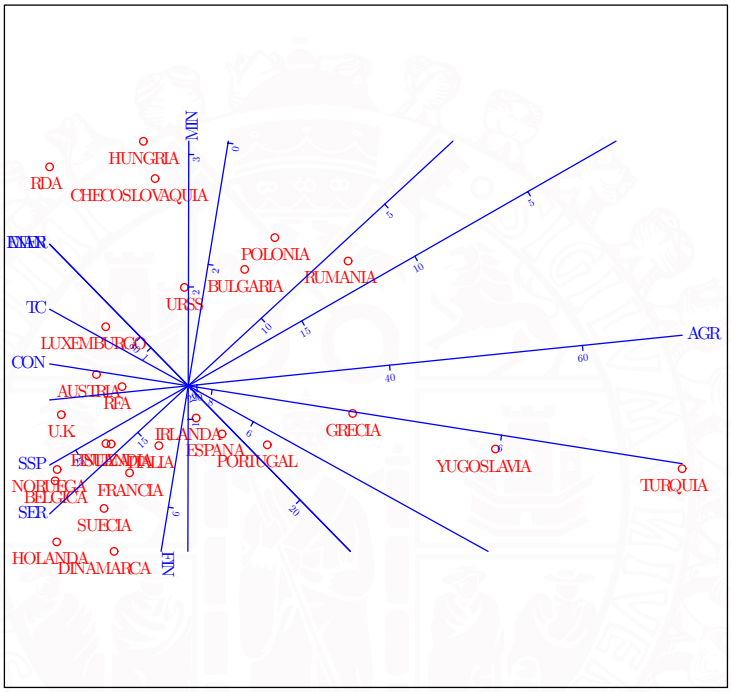


Figura 3.2: PCA-Biplot[a] con escalas para las variables.

[a]Imagen generada con el paquete de R PCABiplot [Vicente-Villardón, 2014]

Proyectando los marcadores fila sobre los ejes del biplot para cada una de las variables proporciona las predicciones en la representación(figura 3.3).

En cuanto a los ejes del biplot es interesante comentar que en 1996, Gower y Hand definen en el contexto de los biplots de componentes principales los Biplots de

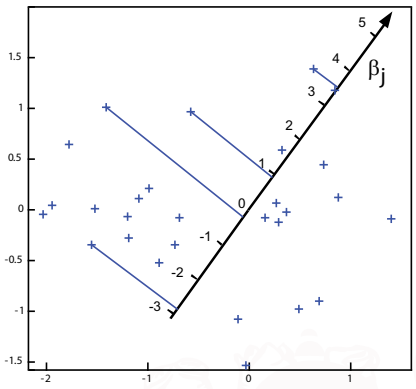
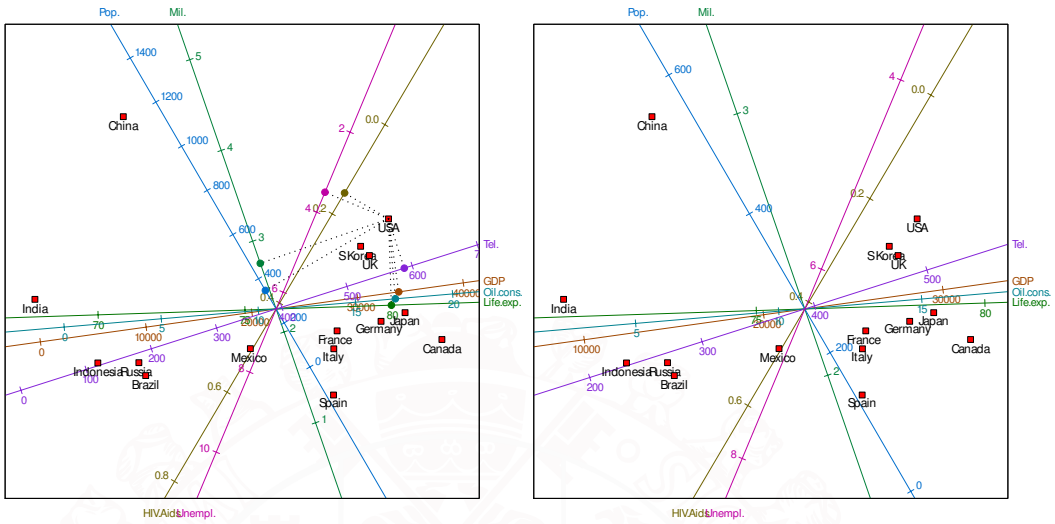


Figura 3.3: Proyecciones de los individuos sobre un hipotético eje biplot para predecir los valores de los mismos en la variable de estudio

Interpolación[2] y Predicción[3]. Con los de interpolación es posible superponer nuevos individuos proyectándolos sobre el subespacio de la representación por medio de la suma de vectores en el gráfico, mientras que con los de predicción es posible inferir valores de las variables originales dado un punto sobre la representación en dimensión reducida (Figura 3.4). La razón por la que es necesario este matiz es porque dado un conjunto único de ejes no ortogonales para interpolar y predecir este arroja representaciones inconsistentes para un mismo individuo (ver figura 3.6, Gower y col. [2011], Capítulo 2). La interpolación del punto dado por las coordenadas mostrado en la figura 3.6(a) (que es el resultado de la predicción del cuadrado azul sobre el conjunto de ejes no ortogonales) se lleva a cabo completando

[2]Por Interpolación se entiende el proceso destinado a encontrar la posición de un elemento de la muestra en el espacio del biplot(en $\mathcal{L}$), dados los valores originales de dicho elemento medidos sobre las variables de estudio. Se lleva a cabo relacionando los valores anteriores con el conjunto de ejes del biplot, que se llaman ejes de interpolación

[3]Es el procedimiento de inferir los valores que toma un elemento de la muestra en las variables consideradas teniendo en cuenta la posición de dicho punto en el biplot.



(a) (b)

Figura 3.4: PCA Biplot predictivo[a](a) e interpolativo(b) de un conjunto de datos de países centrados y escalados, mostrando para el primero de ellos a USA proyectado en todos los ejes biplot para el caso predictivo.

[a]Para representar este biplot se ha utilizado el paquete de R BiplotGUI(ver la Grange y col. [2009]). La situación de los ejes puede depender de cómo van a utilizarse. Los ejes de tipo predictivo se posicionan y se calibran de tal forma que la proyección ortogonal de un punto sobre ellos predice lo mejor posible el valor del individuo sobre la variable en cuestión. Sin embargo, los ejes de tipo interpolativo se sitúan y escalan para que un nuevo individuo pueda ser añadido a la configuración existente(figura 3.5). Pueden consultarse estos conceptos con más detalle en Gower y col. [2011], así como una descripción resumida y detallada de los cálculos necesarios para etiquetar los ejes según la técnica utilizada en la Grange y col. [2009].

el paralelogramo, obteniendose así otro punto distinto, mostrado en la figura 3.6(b). Repitiendo este proceso se obtienen la representaciones de las figuras 3.6(c) y (d). Es claro pues, que esa inconsistencia está patente, puesto que un mismo punto de

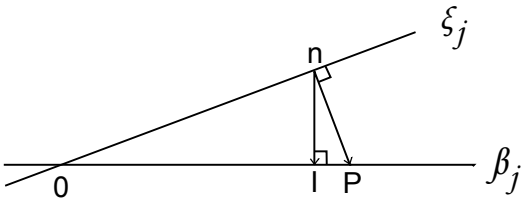


Figura 3.5: Relación de los marcadores en los ejes de predicción e interpolación.

la muestra no se representa igual dado ese conjunto de ejes no ortogonales.

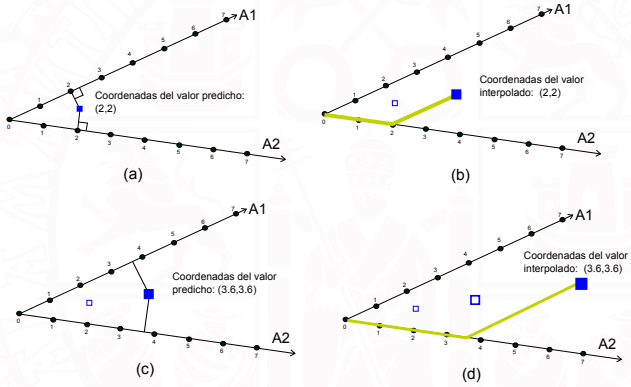


Figura 3.6: Inconsistencia de la predicción y la interpolación para un conjunto de ejes del biplot no ortogonales.

Los ejes que se utilizan en el proceso de interpolación son los mismos(tienen la misma dirección) que para la predicción [Gower y Hand, 1996], pero están calibrados o etiquetados de diferente forma, puesto que los de interpolación están relacionados de forma inversa con los de predicción, como veremos. Las coordenadas de una unidad en el eje k-ésimo para interpolar vienen dadas por $\mathbf{e}_k'\mathbf{V}_{(S)}$, que es la k-ésima columna de $\mathbf{V}_{(S)}$. Además, no sólo las escalas para predecir e interpolar son diferentes, sino que se utilizan también de forma distinta.

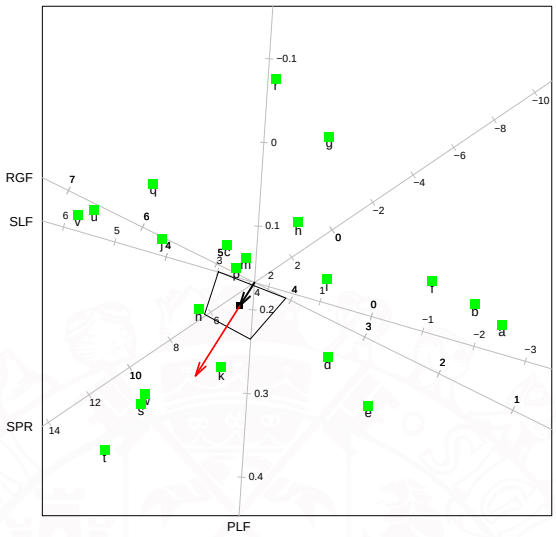


Figura 3.7: Interpolación mediante el vector suma. Los vértices del polígono de 4 lados proporcionan los valores de las 4 variables que se van a interpolar. El extremo de la flecha roja es cuatro veces la longitud de la flecha negra e indica la posición del punto interpolado

Si consideramos un punto correspondiente a un individuo de coordenadas $\mathbf{x}'$, entonces

$$\mathbf{x}'\mathbf{V}_{(S)} = \sum_{k=1}^{J} x_k(\mathbf{e}_k'\mathbf{V}_{(S)}) = J \times \left(\frac{1}{J}\sum_{k=1}^{J} x_k(\mathbf{e}_k'\mathbf{V}_{(S)})\right) = J \times centroide \qquad (3.11)$$

con lo cual interpolar un punto equivale a la suma de vectores de los puntos correspondientes a los marcadores $x_1, x_2, \ldots, x_J$. La suma se obtiene de un forma sencilla símplemente, en lugar de construir paralelogramos, encontrando el centroide de los puntos dados por dichos marcadores y multiplicar por J ese vector extendiéndolo así desde el origen de coordenadas(figura 3.7).

Es sencillo comprobar que la interpretación de los Biplot Clásicos en términos de producto escalar está relacionada con los Biplot de Predicción, y que en definitiva, el proceso de predicción se puede ver desde un punto de vista del análisis de regresión multivariante.

Puesto que el espacio reducido del biplot, $\mathcal{L}$ está contenido en $\mathbb{R}^J$, cualquier punto de dicho espacio también está en $\mathbb{R}^J$ y por tanto se puede expresar tanto en una base de $\mathbb{R}^J$ como en una base de $\mathcal{L}$.

Consideremos pues un punto $\mathbf{z}$ del espacio S-dimensional del PCA-Biplot, $\mathcal{L} = \mathcal{V}(\mathbf{V}_{(S)})$, donde las coordenadas de $\mathbf{z}$ están referidas a las columnas de $\mathbf{V}_{(S)}$.

El proceso de predicción consiste en encontrar las coordenadas de $\mathbf{z}$ respecto a la base del espacio de medida J-dimensional $\mathbb{R}^J$. Llamemos $\mathbf{x}$ al vector de coordenadas de $\mathbf{z}$ en una base de $\mathbb{R}^J$. Puesto que $\mathbf{x}$ pertenece al espacio reducido del biplot, $\mathbf{x}$ se proyecta sobre sí mismo cuando se proyecta ortogonalmente sobre el espacio del biplot, es decir,

$$\mathbf{x}'\mathbf{V}_{(S)}\mathbf{V}'_{(S)} = \mathbf{x}'.$$

Dado que las coordenadas del vector $\mathbf{x}'$ respecto a la base de $\mathcal{L}$, dada por las columnas de $\mathbf{V}_{(S)}$, se calculan mediante $\mathbf{z}' = \mathbf{x}'\mathbf{V}_{(S)}$, entonces obtenemos que $\mathbf{x}' = \mathbf{z}'\mathbf{V}'_{(S)}$, es decir, la interpolación del punto.

Las coordenadas de las proyecciones de los individuos en el subespacio S-dimensional $\mathcal{L}$ vienen dadas por las filas de $\mathbf{Z}$, siendo $\mathbf{Z} = \mathbf{X}\mathbf{V}_{(S)}$. Esta es la definición de predicción, en definitiva, es decir, predecir $\mathbf{X}$ mediante $\mathbf{Z}$. Por otra parte, la predicción de un vector correspondiente a los valores de una variable mediante otro vector es precisamente el objetivo del análisis de regresión multivariante. La diferencia entre el PCA y dicho análisis es que en éste algunas variables se definen como variables predictoras o otras como variables respuesta, mientras que en el PCA las variables no tienen roles distintos. Cuando el proceso de predicción se lleva a cabo mediante regresión multivariante, las componentes principales juegan el papel de variables independientes y las variables de $\mathbf{X}$ son las dependientes o variables respuesta. Por tanto, el modelo de regresión multivariante, en

notación matricial, viene dado por

$$\mathbf{X} = \mathbf{ZB} + \mathbf{E} \tag{3.12}$$

donde $\mathbf{B}$ es la matriz de parámetros y $\mathbf{E}$ es una matriz de errores aleatorios. Este modelo lineal aproxima la matriz $\mathbf{X}$ mediante $\widehat{\mathbf{X}} = \mathbf{Z}\widehat{\mathbf{B}}$, donde $\widehat{\mathbf{B}}$ es el estimador obtenido por mínimos cuadrados de $\mathbf{B}$, es decir, $\widehat{\mathbf{B}} = (\mathbf{Z}'\mathbf{Z})^I\mathbf{Z}'\mathbf{X}$, con $(\mathbf{Z}'\mathbf{Z})^I$ una inversa condicional arbitraria de $\mathbf{Z}'\mathbf{Z}$.

Como $\mathbf{XV}_{(S)} = \mathbf{U}_{(S)}\mathbf{\Lambda}_{(S)}$ es claro que el rango de la matriz resultante de ambos productos es S, y por tanto la matriz $\mathbf{Z} = \mathbf{XV}_{(S)}$ es una matriz de rango completo, satisfaciendo así una propiedad e hipótesis importante del análisis de regresión. Esto implica pues que la matriz $\mathbf{Z}'\mathbf{Z}$ es no singular, y entonces el estimador de $\mathbf{B}$ es

$$\begin{aligned}\widehat{\mathbf{B}} &= (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X} = (\mathbf{V}'_{(S)}\mathbf{X}'\mathbf{X}\mathbf{V}_{(S)})^{-1}\mathbf{V}'_{(S)}\mathbf{X}'\mathbf{X} \\ &= (\mathbf{V}'_{(S)}\mathbf{V}_p^2\mathbf{\Lambda}^2_{(S)}\mathbf{V}'_p\mathbf{V}_{(S)})^{-1}\mathbf{V}'_{(S)}\mathbf{V}_p^2\mathbf{\Lambda}^2_{(S)}\mathbf{V}'_p \\ &= (\mathbf{\Lambda}^2_{(S)})^{-1}\mathbf{\Lambda}^2_{(S)}\mathbf{V}'_{(S)} = \mathbf{V}'_{(S)}\end{aligned} \tag{3.13}$$

de tal forma que

$$\widehat{\mathbf{X}} = \mathbf{XV}_{(S)}\mathbf{V}'_{(S)}$$

De esta forma se comprueba que la aproximación de $\mathbf{X}$ que se obtiene como resultado de la regresión multivariante de $\mathbf{X}$ sobre $\mathbf{Z}$ es la misma que la obtenida mediante el PCA-Biplot de predicción.

Capítulo 4

Biplot Logístico de Variables Binarias

Individuals vary, but percentages remain constant.
So says the statistician.

– *Arthur Conan Doyle*

Tanto el biplot clásico lineal como el PCA son técnicas que se aplican si suponemos que la respuesta a lo largo de las dimensiones es lineal. Por este motivo, si disponemos de observaciones de caracteres cualitativos correspondientes a variables binarias, que toman el valor 0 si la característica está ausente y el valor 1 si está presente, los procedimientos anteriores no son válidos para realizar un análisis de esta situación.

Cuando el conjunto de datos corresponde a la medición de variables binarias sobre un colectivo, técnicas como el MCA podrían considerarse como una forma

particular de ajuste biplot, en las cuales las distancias entre los individuos y las categorías determinan las regiones de predicción [Gower y Hand, 1996]. Otras opciones podrían ser la utilización de modelos de regresiones generalizadas alternadas para estimar la respuesta binaria, estimando cada columna de $\mathbf{X}$ suponiendo la independencia entre individuos y de los parámetros de cada una de las variables; o bien la utilización de la regresión bilineal generalizada para estimar todas las columnas de $\mathbf{X}$; o la estimación conjunta simultánea de todas las filas y columnas de $\mathbf{X}$ (van Eeuwijk [1995a,b]; Blázquez [1998];Gabriel [1998];Vicente-Villardón y col. [2006]).

Existen numerosos estudios sobre tablas de 2 vías resultantes de la clasificación cruzada de dos variables, por ejemplo, investigaciones llevadas a cabo por van Eeuwijk [1995b] o Falguerolles y Francis [1994] se aplican al campo de la agricultura experimental, con atención a la interacción de genotipos y ambientes, y están concebidas en la filosofía de los modelos de asociación RC ("Row and Columns", filas y columnas) de Goodman [1991], es decir, modelos que describen la asociación entre las filas y columnas de una tabla de 2 vías. Para datos continuos esto equivaldría a una variable(por ejemplo, el rendimiento) medida para diversos genotipos(variedades de cultivos) en varios ambientes(diferentes localizaciones). En una situación como esta los papeles de las filas y las columnas son de alguna forma simétricas y la variable respuesta es de tipo numérico. En un contexto general la respuesta podría ser, por ejemplo, recuentos con una distribución de Poisson.

Nuestras investigaciones están más relacionadas con modelos de rasgos latentes o, para datos continuos, con componentes principales o modelos factoriales. Trataremos con matrices de individuos a los que se les miden una serie de variables, en los que los papeles de las filas y columnas no son simétricos. Uno de los propósitos que se persiguen es la reducción de la dimensión con el objetivo de interpretar la ordenación o clasificación de los individuos y variables responsables de la misma. Desde otro punto de vista, el propósito podría verse como la explicación de la relación entre las variables observadas en términos de un número reducido de factores

latentes comunes y la investigación de las puntuaciones de los individuos en tales factores. Los roles de todas las variables son simétricos, en el sentido de que no hay distinción alguna entre variables respuesta y variables de clasificación. En ambos casos, la representación conjunta de invididuos y variables es útil para entender la estructura de los datos, y además puede ayudar a descubrir patrones ocultos en los mismos e incluso ayudar a contrastar hipótesis de la misma forma que se hace en el escalado multidimensional. Dicha representación(biplot) es también una herramienta para averiguar las variables asociadas a las diferencias entre los individuos. Al respecto de estas consideraciones, nuestra propuesta está más cerca del punto de vista de la cuantificación asociada al CA que es propia del equipo Gifi[1].

En ambos planteamientos, las tablas de dos vías y las matrices de individuos por variables, la descomposición en valores singulares está presente, pero existen algunas diferencias en la interpretación de los parámetros. En el primero de ellos el producto escalar de los marcadores fila y columna se interpreta en términos de interacción(o asociación para datos categóricos), mientras que en el segundo se interpreta como el valor esperado(o probabilidad esperada) que un individuo toma sobre una variable. En el capítulo 5, cuando desarrollemos el biplot de variables nominales, convertiremos la esperanza del logaritmo de los odds y las probabilidades esperadas en categorías esperadas utilizando para ello los puntos categoría más cercanos. Incluso se podrían utilizar otro tipo de aproximaciones y posibilidades, pero hemos elegido aquella que nos parece más sencilla de interpretar.

En el caso de roles no simétricos de filas y columnas, los algoritmos alternados necesitan una serie de adaptaciones porque el procedimiento utilizado para las

[1]De Leeuw es el pionero del equipo Albert Gifi que escribió “Nonlinear Multivariate Analysis”. En el libro “Multidimensional Scaling”, Volumen 1, Cox y Cox [2000] escriben que “Albert Gifi” es el nombre de pila de los miembros actuales y anteriores del departamento de Teoría de Datos de la Universidad de Leiden, los cuales desarrollaron un sistema de análisis multivariante no lineal que generaliza varias técnicas, como el PCA y el Análisis de Correlación Canónica (CCA).

tablas de 2 vías puede no funcionar para una matriz de datos. Por ejemplo, en el caso binario que veremos a continuación, la puntuación para un individuo que tiene presencias o ausencias en todas las variables no se puede calcular. De este modo cambiaremos la regresión para las filas por una interpolación utilizado esperanzas, y máxima verosimilitud marginal para los parámetros de las columnas como en la IRT.

Vicente-Villardón y col. [2006] proponen, como describiremos en las siguientes secciones, el ajuste de un biplot logístico lineal para datos binarios, en el cual la respuesta a lo largo de las dimensiones retenidas es logística, que está basado en regresiones o interpolaciones alternadas. El método que resulta de estas investigaciones difiere de las propuestas de van Eeuwijk [1995a,b], Gabriel [1998] o Falguerolles [1998], puesto que el principal objetivo es analizar la matriz de datos de variables medidas sobre un conjunto de invididuos y no modelar una tabla de dos vías.

4.1. Formulación

Consideremos $\mathbf{X}_{I\times J}$ como una matriz de datos en la que las filas corresponden a I individuos y las columnas a J variables binarias. Sea $\pi_{ij} = E(x_{ij})$ la probabilidad esperada de que la variable j esté presente en el individuo i, y x_{ij} el valor observado, o bien 0 ó 1, conformando así una matriz de datos binaria. El biplot logistico S-dimensional en escala *logit* se expresa como sigue:

$$logit(\pi_{ij}) = \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = b_{j0} + \sum_{s=1}^{S} b_{js}a_{is} = b_{j0} + \mathbf{a}_i'\mathbf{b}_j, \tag{4.1}$$

donde a_{is} y b_{js}, $(i = 1,\ldots,I; j = 1,\ldots,J; s = 1,...,S)$, son los parámetros del modelo, utilizados como marcadores fila y columna respectivamente. Esta configuración es un modelo (bi)lineal generalizado que tiene la función *logit* como función de enlace. En términos de probabilidades ,en lugar de *logits*:

$$\pi_{ij} = \frac{e^{b_{j0}+\sum_k b_{jk}a_{ik}}}{1+e^{b_{j0}+\sum_k b_{jk}a_{ik}}} = \frac{1}{1+e^{-(b_{j0}+\sum_k b_{jk}a_{ik})}} \tag{4.2}$$

En forma matricial:

$$logit(\mathbf{\Pi}) = \mathbf{1}_I \mathbf{b}_0' + \mathbf{A}\mathbf{B}', \tag{4.3}$$

con $\mathbf{\Pi}$ la matriz de probabilidades esperadas, $\mathbf{1}_I$ es un vector de unos y $\mathbf{b}_0 = (b_{j0})$ es un vector que contiene los términos independientes, que han sido añadidos porque no es posible centrar la matriz de datos de la misma forma que se hacía en el caso de los biplots lineales o clásicos. Dichos términos son desplazamientos de los centroides, de la misma forma que lo es el primer eje de ordenación en el Análisis de Correspondencias. Este modelo es un modelo de rasgo latente para datos binarios, siendo las coordenadas de las filas las puntuaciones de los individuos sobre dicha respuesta latente.

4.2. Estimación del modelo logístico binario

El modelo presentado en (4.1), como hemos dicho, es un modelo de rasgo latente similar a los modelos propios de la teoría de respuesta al ítem, en el cuál los ejes principales se consideran como variables latentes que explican la asociación entre las variables observadas.

Supondremos que los individuos contestan de forma independiente a las variables, y que las variables son independientes de valores predefinidos de las dimensiones latentes. De esta forma, la función de verosimilitud es:

$$\text{Prob}(x_{ij} | \mathbf{b}_0, \mathbf{A}, \mathbf{B}) = \prod_{i=1}^{I} \prod_{j=1}^{J} \pi_{ij}^{x_{ij}} (1 - \pi_{ij})^{1 - x_{ij}} \tag{4.4}$$

Tomando el logaritmo de dicha función obtenemos:

$$L = \log \text{Prob}(x_{ij} | \mathbf{b}_0, \mathbf{A}, \mathbf{B}) = \sum_{i=1}^{I} \sum_{j=1}^{J} [x_{ij} \log(\pi_{ij}) + (1 - x_{ij}) \log(1 - \pi_{ij})] \tag{4.5}$$

El cálculo de los estimadores está supeditado a operar la función L, derivándola con respecto a los parámetros e igualando a cero cada derivada, lo cual obliga

a resolver un sistema de $3J + 2I$ ecuaciones. Mediante del Método de Newton-Raphson se podrían obtener las soluciones, pero presenta problemas cuando el número de individuos o variables es alto.

Vicente-Villardón y col. [2006] proponen un esquema de estimación iterativo en el que se alterna la actualización de las matrices **A** y **B** en cada paso del algoritmo hasta que se alcanza un nivel de precisión predefinido. En cada iteración la función L se puede separar en una parte para cada fila o cada columna de la matriz de datos, maximizando entonces cada una por separado. Este proceso convergerá a un máximo local, y puede considerarse una generalización del método de regresión/interpolación de los biplots clásicos, puesto que si los datos siguen una distribución normal multivariante y se utiliza la función de enlace identidad en lugar de la función logit, la solución coincide con la del biplot clásico.

Para describir el proceso con más detalle, si consideramos **A** fijo, (4.5) se puede separar en J partes, una para cada variable:

$$L = \sum_{J=1}^{J} L_j = \sum_{J=1}^{J} \left(\sum_{i=1}^{I} [x_{ij} log(\pi_{ij}) + (1 - x_{ij}) log(1 - \pi_{ij})] \right). \tag{4.6}$$

Maximizar cada L_j es equivalente a llevar a cabo una regresión logística utilizando la columna j-ésima de **X** como la variable respuesta y las columnas de **A** como los regresores. Esta etapa sería la correspondiente a la regresión. De la misma forma, la función de probabilidad se puede separar en varios sumandos, uno para cada fila de la matriz de datos

$$L = \sum_{i=1}^{I} L_i = \sum_{i=1}^{I} \sum_{j=1}^{J} [x_{ij} log(\pi_{ij}) + (1 - x_{ij}) log(1 - \pi_{ij})]$$

El vector gradiente, que denotamos por $\mathbf{g} = (g_1, \ldots, g_S)$ tiene por componentes $g_s = \frac{\partial L_i}{\partial a_{is}} = \sum_{j=1}^{J} b_{js}(x_{ij} - \pi_{ij})$, lo cual puede comprobarse derivando respecto de a_{is}, con $s = 1, \ldots, S$, de tal forma que:

$$\frac{\partial L_i}{\partial a_{is}} = \sum_{j=1}^{J} x_{ij} \frac{1}{\pi_{ij}} \frac{\partial \pi_{ij}}{\partial a_{is}} + \sum_{j=1}^{J} (1 - x_{ij}) \frac{1}{(1 - \pi_{ij})} \frac{\partial (1 - \pi_{ij})}{\partial a_{is}}$$

con

$$\frac{\partial \pi_{ij}}{\partial a_{is}} = b_{js}\pi_{ij}(1-\pi_{ij}) \qquad \frac{\partial (1-\pi_{ij})}{\partial a_{is}} = -b_{js}\pi_{ij}(1-\pi_{ij})$$

y operando convenientemente se obtiene la expresión simplificada de los componentes del gradiente.

La matriz Hessiana $\mathbf{H}$ tiene por componentes:

$$h_{ss} = \frac{\partial^2 L_i}{\partial a_{is}^2} = -\sum_{j=1}^{J} b_{js}^2 \pi_{ij}(1-\pi_{ij}) \qquad h_{ss'} = \frac{\partial^2 L_i}{\partial a_{is}\partial a_{is'}} = -\sum_{j=1}^{J} b_{js}b_{js'}\pi_{ij}(1-\pi_{ij})$$

con lo que el método iterativo de Newton-Raphson para encontrar las soluciones del sistema establece:

1. Asignar valores iniciales a los marcadores fila, en $t = 0$, $[a_{i1}, \ldots, a_{iS}]_0^T$
2. Actualizar el vector $[a_{i1}, \ldots, a_{iS}]_{t+1}^T$ con

$$\begin{bmatrix} a_{i1} \\ \vdots \\ a_{iS} \end{bmatrix}_{t+1} = \begin{bmatrix} a_{i1} \\ \vdots \\ a_{iS} \end{bmatrix}_{t} + \mathbf{H}_g^{-1}$$

   estimando π_{ij} con los valores de los parámetros en el periodo t.
3. Incrementar el contador $t = t + 1$
4. Si la variación del vector $[a_{i1}, \ldots, a_{iS}]_0^{t+1}$ es pequeña el proceso termina, y si no volvemos al paso 2.

Cuando los vectores respuesta son dispersos o cuando son todo ceros o todo unos este procedimiento presenta algunos problemas, que se pueden resolver mediante correcciones de las probabilidades esperadas. No obstante el método funciona en la mayoría de los casos.

Por tanto el algorimo propuesto por Vicente-Villardón y col. [2006] es el siguiente:

Paso 1 Elegir valores iniciales para los parámetros $\mathbf{A}$. Pueden ser por ejemplo el resultado de hacer un PCA a $\mathbf{X}$

Paso 2 Ortonormalizar la matriz **A** para evitar indeterminaciones

Paso 3 **(Etapa de regresión)** Calcular para cada columna $\mathbf{x}_j$ de **X** mediante una regresión logística estándar los parámetros $b_{j0}, \ldots, b_{jS}$

Paso 4 **(Etapa de interpolación)** Interpolar cada individuo de forma separada, es decir, calcular $a_{i0}, \ldots, a_{iS}$ mediante el método de Newton-Raphson citado anteriormente.

Paso 5 Si los cambios en la función de verosimilitud son pequeños parar el proceso y si no continuar.

4.3. Geometría de los biplots logísticos binarios

La descripción de la geometría la haremos para una solucion bidimensional, de tal forma que si suponemos conocidos los marcadores de las filas **A** y ajustamos el modelo descrito por las ecuaciones 4.2 obtenemos superficies de respuesta similares a las de la figura 4.1. Es palpable que aunque las superficies de respuesta no son lineales, la intersección de las mismas con planos perpendiculares al eje de probabilidad son líneas rectas. Por tanto, los puntos en el plano de representación del biplot que predicen las diferentes probabilidades están situados en líneas rectas paralelas sobre el mismo. Los cálculos necesarios para obtener marcadores en el eje del biplot son sencillos. Para encontar el marcador para una probabilidad establecida π lo que hacemos es observar el punto (x, y) que predice π y que está en la dirección $\mathbf{b}_j$, es decir, sobre la línea que une los puntos (0, 0) y (b_{j1}, b_{j2}), o sea

$$y = \frac{b_{j2}}{b_{j1}} x$$

y

$$logit(\pi) = b_{j0} + b_{j1}x + b_{j2}y$$

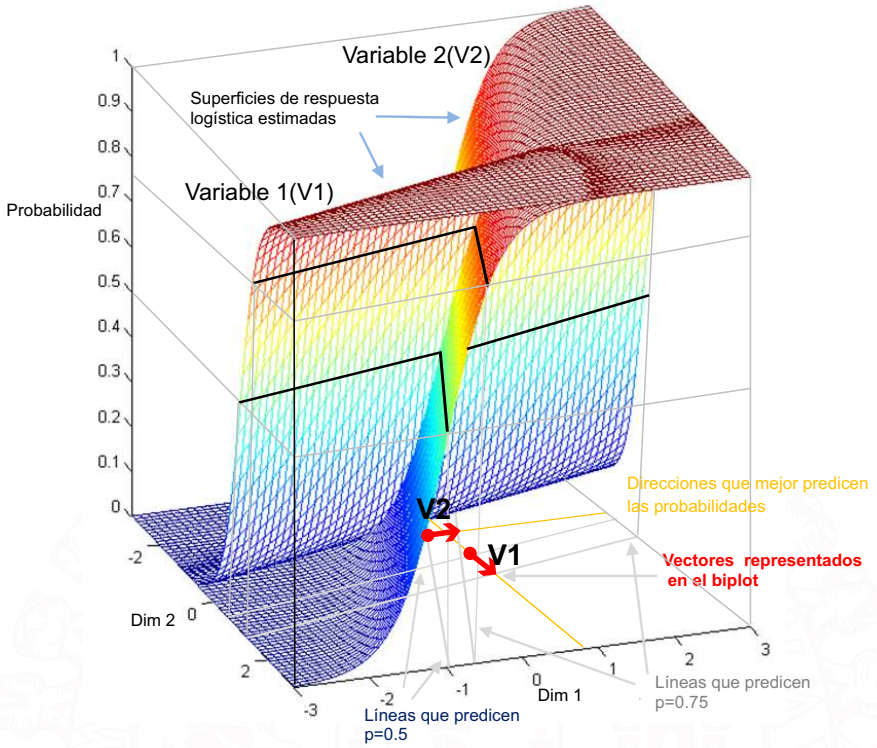


Figura 4.1: Geometría del Biplot Logístico Binario.

Obtenemos de esta forma

$$x = \frac{(logit(\pi) - b_{j0})b_{j1}}{b_{j1}^2 + b_{j2}^2}$$

y

$$y = \frac{(logit(\pi) - b_{j0})b_{j2}}{b_{j1}^2 + b_{j2}^2}$$

Se pueden etiquetar diversos puntos para valores concretos de π obteniendo así una escala con referencias(figura 4.2). Desde un punto de vista práctico el valor más interesante es $\pi = 0{,}5$ puesto que la línea que pasa por ese punto y es perpendicular a la dirección dada por el eje divide la representación en dos regiones, una que predice las *presencias* y otra que predice las *ausencias*. Dibujando pues ese punto y una flecha que apunte a la dirección en la que crecen las probabilidades debería ser suficiente para la mayoría de las aplicaciones(figura 4.3).

Una de las principales diferencias de la geometría de estos biplots respecto de los lineales es que las distancias entre las marcas de escala en el eje del biplot no se corresponden con el espaciamiento de las probabilidades, precisamente por la naturaleza de las superficies de respuesta ajustadas (figura 4.4).

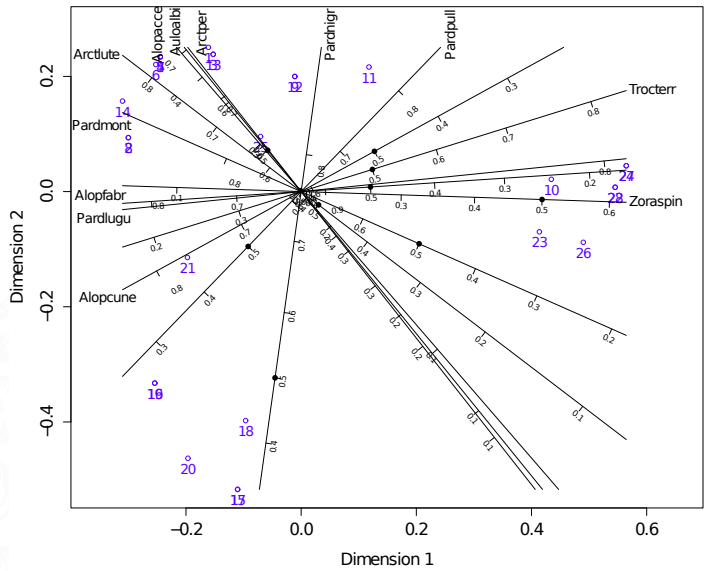


Figura 4.2: Biplot Logístico Binario con escalas de probabilidad para las variables.

4.4. Interpretación del biplot logístico de variables binarias

Las proyecciones en el biplot logístico binario se realizan de la misma forma que en los biplot clásicos lineales, es decir, se proyectan los marcadores de cada fila $\mathbf{a}_i = (a_{i1}, a_{i2})$ sobre las direcciones de cada uno de los $j (j = 1, \ldots, J)$ marcadores

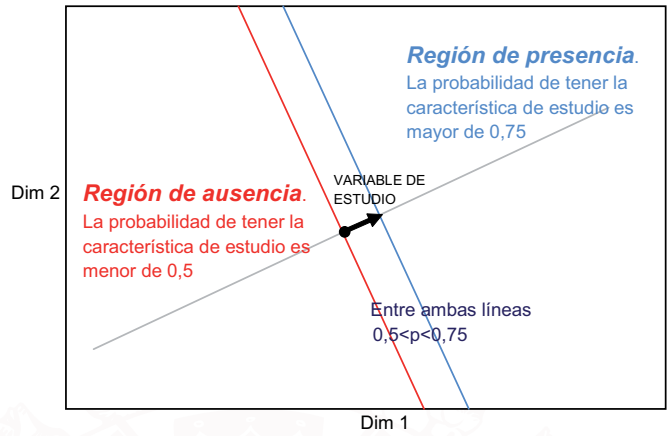
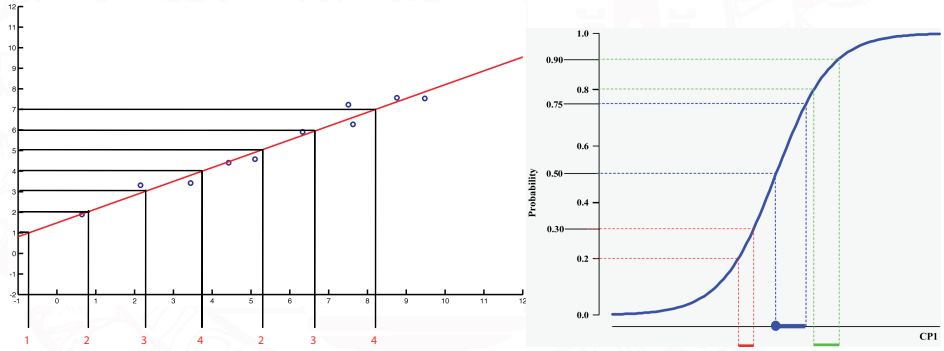


Figura 4.3: Regiones del Biplot Logístico Binario en el análisis de una variable.

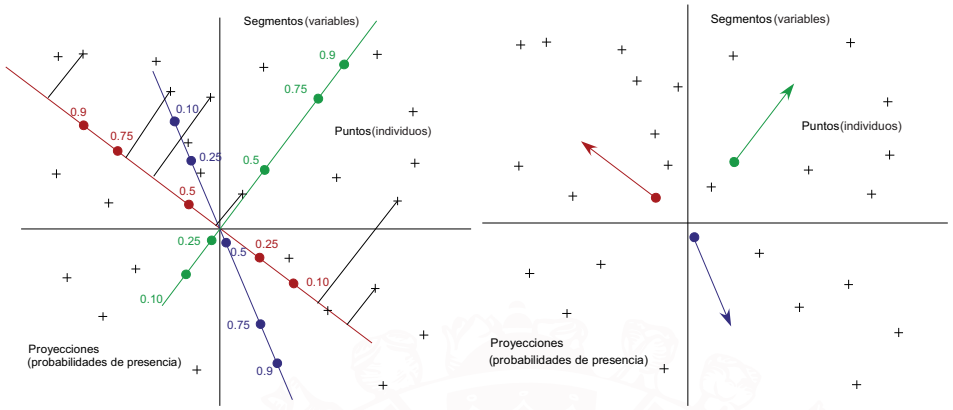


(a) (b)

Figura 4.4: Proyección de la distancia en el biplot entre conjuntos de probabilidades predichas en: (a) Biplot Clásico Lineal,(b) Biplot Logístico Binario

columna $\mathbf{b}_j = (b_{j1}, b_{j2})$(figura 4.5(a)).

Con el objetivo de hacer lo más legible posible el gráfico y la representación



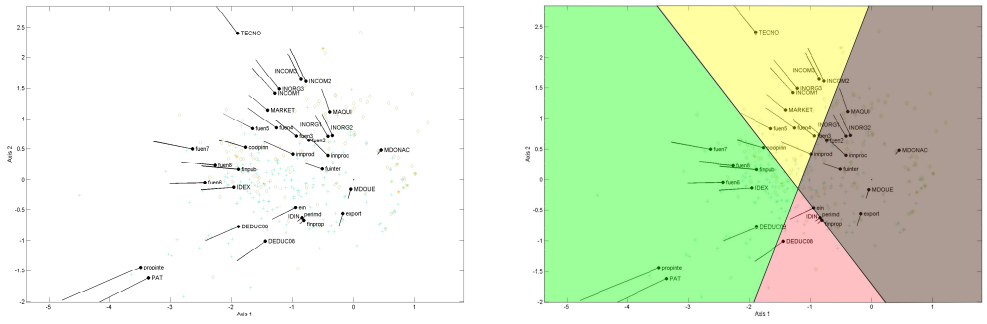
(a) (b)

Figura 4.5: (a) Situación de los ejes del biplot con escalas, los individuos y las proyecciones, y (b) simplificación de los elementos del biplot utilizando sólo el tramo de probabilidad de 0.5 a 0.75

biplot, lo que suele hacerse es considerar sólo la parte del eje biplot que comprende el tercer cuartil de probabilidad (figura 4.5(b)), con lo cuál, por ejemplo, en un caso real el biplot logístico quedaría como el que se muestra en la figura 4.6(a). Cuando se quieren analizar varias variables a la vez, lo que hay que tener en cuenta es la intersección de las regiones de presencia y ausencia de cada una de ellas, lo cual no es sencillo a medida que aumenta ese número(figura 4.6(b)).

Será necesario analizar la bondad de ajuste del modelo(ver capítulo 7.6), que se traduce en el porcentaje de clasificaciones correctas a nivel global y que determina de alguna forma si el mismo puede o no considerarse válido. Si el mismo es elevado significa que tanto las ausencias como las presencias se predicen bien en la mayoría de los casos.

Atendiendo a las regiones que mostrábamos en la figura 4.3 es posible ver qué características tienen los individuos mediante las intersecciones de dichas regiones.



(a) (b)

Figura 4.6: (a) Biplot Logístico Binario calculado con un conjunto de datos de empresas extraídos de la Encuesta de Innovación que realiza el INE.(b) Regiones de presencia y ausencia para las variables binarias IDIN(I+D interna) e innprod(innovación de producto)

Por tanto, cuanto más cortos sean los vectores mayor poder de discriminación tendrá la variable sobre los individuos para determinar si los mismos poseen esa característica o no y viceversa. El ángulo entre vectores indica el grado de asociación entre las variables que representan, de tal forma que ángulos agudos pequeños significan que las variables están muy relacionadas. Además, el ángulo que forman los vectores con las dimensiones latentes determina si están positiva o negativamente correlacionados con ellas, de forma que sea posible caracterizarlas con la información de las variables relevantes(figura 4.7).

Fundamentalmente para evaluar la calidad de representación del análisis habremos de fijarnos en varios valores diferentes:

1. Calidad de representación de los individuos: hay que apuntar que en realidad no se pueden calcular en este biplot logístico ya que no disponemos de las

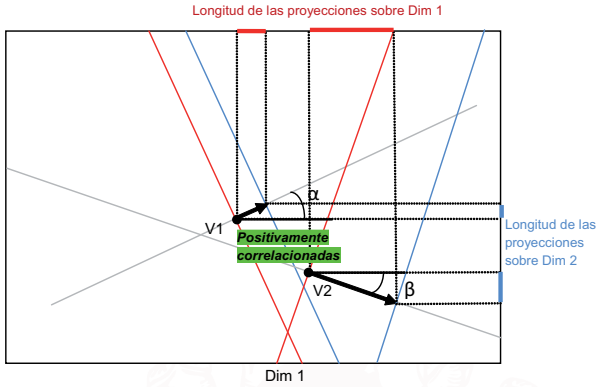


Figura 4.7: Relación de los vectores con las dimensiones latentes.

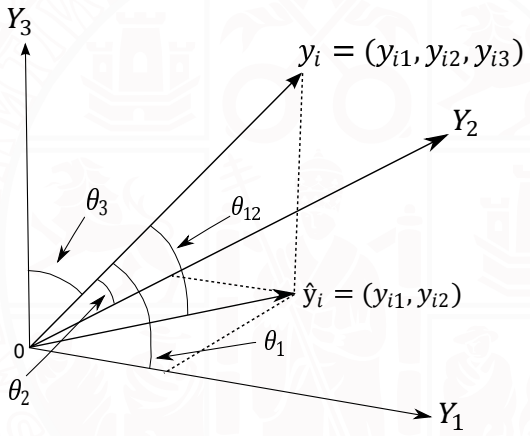


Figura 4.8: Interpretación geométrica de la calidad de representación del individuo i-ésimo.

coordenadas de los mismos en el espacio multidimensional. Solamente las calculamos cuando las estimamos a partir de un Análisis de Coordenadas Principales. En este caso, si la mayoría de la información de un individuo, medida mediante la variabilidad, se concentra en las primeras S dimensiones, diremos que dicho individuo está bien representado. Si consideramos una representación con los datos centrados en el origen, esa variabilidad viene dada

por la distancia al cuadrado entre la posición del punto en la representación y el origen del mismo. Por tanto la calidad de representación es el cociente entre la distancia en el espacio reducido y en el espacio original, es decir:

$$\mathrm{CR}_i^S = \left(\frac{\sum_{s=1}^{S} y_{is}^2}{\sum_{j=1}^{J-1} y_{ij}^2} \right) \times 100\,\%$$

con y_{ij} la coordenada principal del individuo i-ésimo sobre la j-ésima dimensión, o interpretándolo geométricamente(figura 4.8):

$$\cos^2(\theta) = \left(\frac{d^2(\hat{y}_i - 0)}{d^2(y - 0)} \right)$$

siendo $\cos(\theta_{12}) = \cos(\theta_1) + \cos(\theta_2)$ y $\cos(\theta_1) + \cos(\theta_2) + \cos(\theta_3) = 1$, expresión de la que se deriva la importancia relativa de cada individuo. Es decir, el coseno elevado al cuadrado del ángulo formado entre el individuo y su proyección en el eje correspondiente nos indicará el porcentaje de variabilidad del individuo capturado por dicho eje.

2. De igual forma, el valor absoluto del coseno elevado al cuadrado del ángulo de cada variable con otras o con cada eje biplot nos indicará con qué variable o eje está más relacionada.

3. En cuanto a la calidad de representación de las variables hemos de considerar tres índices:

   (a) p-valor del ajuste de la regresión logística para cada variable, con el objetivo de analizar si son significativas.

   (b) R^2 de Nagelkerke: para conocer la bondad del ajuste (capacidad explicativa del modelo) de la regresión logística de cada variable.

   (c) Porcentaje de individuos clasificados correctamente según el modelo: se utiliza el valor 0,5 como punto de corte para la probabilidad esperada ($< 0{,}5$ es ausente y $> 0{,}5$ presente).

Los Biplots Logísticos Binarios se pueden calcular mediante el software informático MULTBIPLOT [Vicente-Villardón, 2010], e interesantes estudios de los mismos se han llevado a cabo, como los de Demey y col. [2008] en el campo de la genética, el cuál es una la solución al problema de estimación cuando el número de variables es muy elevado. En este caso, se estiman las coordenadas de los individuos mediante Coordenadas Principales y se proyectan las variables con regresiones logísticas. Una ventaja de hacerlo de esta forma es que se pueden calcular las calidades de representación de cada individuo ya que tenemos las coordenadas en el espacio multidimensional y no solamente en dimensión reducida como ocurre con el algoritmo alternado.

Capítulo 5

Biplot Logístico Nominal

I paint objects as I think them, not as I see them.
– Pablo Picasso

En este capítulo se desarrollan las fases necesarias para construir este tipo de biplots, que son novedosos y una de las principales aportaciones de esta investigación. Primeramente se detalla el planteamiento matemático, describiendo posteriormente la geometría del modelo, la cual resulta, para cada variable, en la construcción de una serie de regiones de predicción mediante un algoritmo general. Veremos cómo el ajuste de un modelo de rasgos latentes para variables nominales, equivalente a ajustar una regresión logística multinomial para cada variable, proporciona teselaciones del espacio de dimensión reducida en el que construímos el biplot, es decir, el espacio de la representación quedará dividido en tantas regiones convexas como categorías de la variable (salvo casos

degenerados). A partir de esta teselación, basada en las probabilidades esperadas del modelo logístico, buscaremos el diagrama de Voronoi más cercano y un conjunto de generadores del mismo. De esta forma convertiremos el problema de la predicción de las categorías en un criterio simple, consistente en la búsqueda del generador más cercano. La estimación del modelo se tratará a continuación de una forma resumida y por último se mostrarán algunas aplicaciones de este método con datos reales y una comparativa del mismo respecto a la técnica más extendida para trabajar con variables nominales, que es el MCA.

5.1. Metodología de construcción de biplots logísticos para datos nominales

5.1.1. Presentación del modelo

Sea $\mathbf{X}_{I\times J}$ una matriz de datos que contiene los valores de J variables nominales, cada una con K_j $(j = 1, \ldots, J)$ categorías, para I individuos, y sea $\mathbf{G}_{I\times L}$ la matriz indicadora correspondiente con $L = \sum_j K_j$ columnas. La última(o la primera) categoría de cada variable se usará como categoría base o de referencia. Denotamos por $\pi_{ij(k)}$ la probabilidad esperada de que la categoría k de la variable j esté presente en el individuo i. Un modelo logístico multinomial de respuesta latente con S rasgos latentes establece que las probabilidades se obtienen de la siguiente forma:

$$\pi_{ij(k)} = \frac{e^{b_{j(k)0}+\sum_{s=1}^{S} b_{j(k)s}a_{is}}}{\sum_{l=1}^{K_j} e^{b_{j(l)0}+\sum_{s=1}^{S} b_{j(l)s}a_{is}}}, (k = 1, \ldots, K_j). \tag{5.1}$$

Utilizando la última categoría como la base para hacer el modelo identificable, el parámetro para esa categoría siempre será 0, es decir, $b_{j(K_j)0} = b_{j(K_j)s} = 0$,

$(j = 1, \ldots, J; \quad s = 1, \ldots, S)$. El modelo entonces puede reescribirse como:

$$\pi_{ij(k)} = \frac{e^{b_{j(k)0} + \sum_{s=1}^{S} b_{j(k)s} a_{is}}}{1 + \sum_{l=1}^{K_j - 1} e^{b_{j(l)0} + \sum_{s=1}^{S} b_{j(l)s} a_{is}}}, (k = 1, \ldots, K_j - 1). \tag{5.2}$$

Con este matiz de la restricción, asumimos que el logaritmo de los odds para cada respuesta (en relación con la última categoría) sigue un modelo lineal:

$$\log\left(\frac{\pi_{ij(k)}}{\pi_{ij(K_j)}}\right) = b_{j(k)0} + \sum_{s=1}^{S} b_{j(k)s} a_{is} = b_{j(k)0} + \mathbf{a}_i' \mathbf{b}_{j(k)},$$

donde a_{is} y $b_{j(k)s}$ $(i = 1, \ldots, I; \quad j = 1, \ldots, J; \quad k = 1, \ldots, K_j - 1; \quad s = 1, \ldots, S)$ son los parámetros del modelo. En forma matricial:

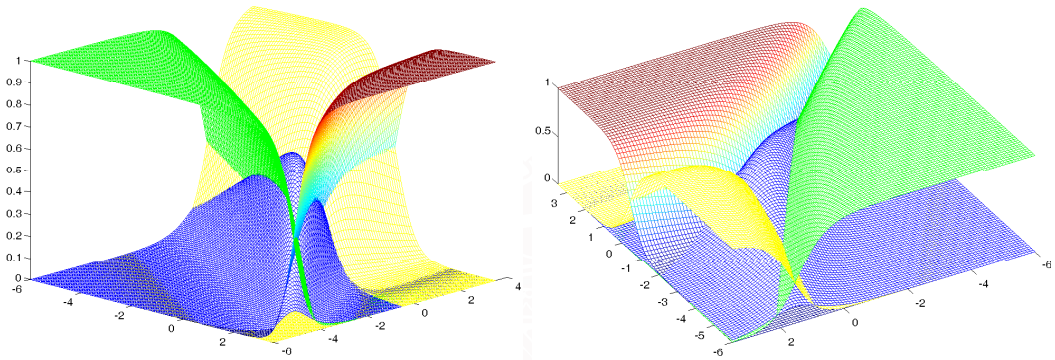
$$\mathbf{O} = \mathbf{1}_I \mathbf{b}_0' + \mathbf{A}\mathbf{B}', \tag{5.3}$$

define un biplot para los odds, siendo $\mathbf{O}_{I \times (L-J)}$ la matriz que contiene el logaritmo de los odds esperados. Aunque el biplot para los odds podría ser útil, sería más interpretable en términos de probabilidades de predicción y categorías. Este biplot lo llamaremos "Biplot Logístico Nominal", y está relacionado con los modelos nominales latentes de la misma forma que los biplots clásicos lineales lo están con el análisis de componentes principales o el análisis factorial, o los biplots logísticos binarios se relacionan con la teoría de respuesta al ítem o el análisis de respuesta latente para datos binarios.

5.1.2. Geometría

Vamos a considerar una situación en la cual las coordenadas de las filas están definidas por las dos primeras columnas de $\mathbf{A}$ y llamemos $\mathcal{L}$ al espacio generado por dichas columnas. Las ecuaciones 5.2 definen un conjunto de superficies de respuesta de probabilidad (una para cada categoría y cada variable, Figura 5.1) que no son sigmoides, como en el caso binario. Esto significa que el conjunto de puntos que predicen las diferentes probabilidades (curvas de nivel) ya no se sitúan

a lo largo de líneas rectas paralelas. La Figura 5.3(a) muestra las curvas de nivel, para la probabilidad 0.5, para una hipotética variable con cuatro categorías.



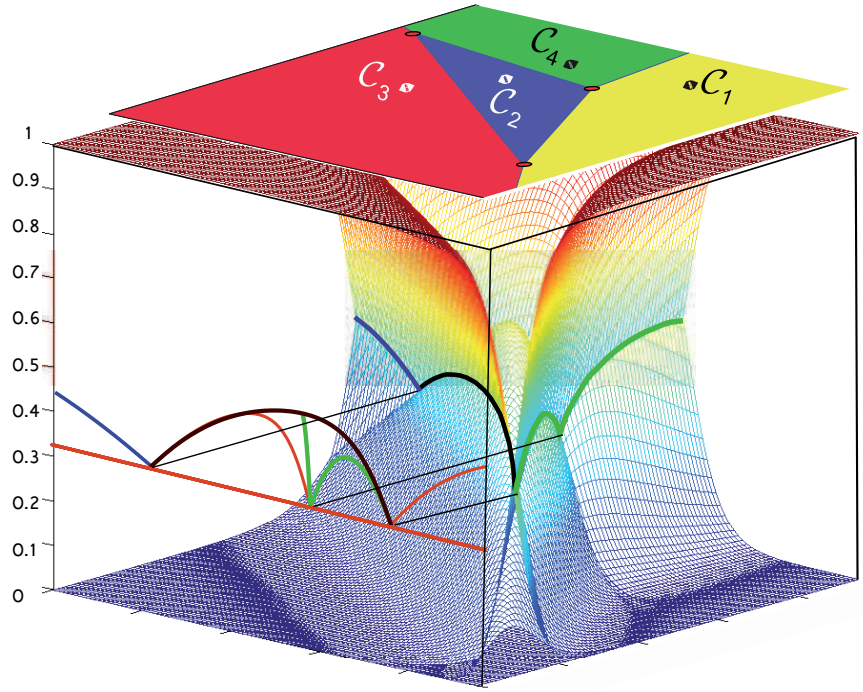
(a) (b)

Figura 5.1: Superficies de respuesta del modelo logístico nominal (a,b) para una variable con 4 categorías y dos variables explicativas.

Por tanto, las predicciones en el biplot logístico nominal no pueden hacerse de la misma forma que en el caso de los biplots lineales. Las superficies de respuesta definen regiones de predicción en lugar de direcciones para cada categoría. Mostraremos que en este caso las probabilidades predichas, para cada variable, definen un conjunto de polígonos convexos que pueden ser interpretados como regiones de predicción, de forma similar a como hacían Gower y Hand [1996]. Para cada variable habrá tantas regiones como categorías tenga y cada una estará formada por los puntos en los que la probabilidad esperada para dicha categoría sea mayor que la probabilidad asignada al resto de categorías de la variable.

En la figura 5.2 se aprecia cómo cada superficie tiene una intersección con el resto, las cuales, vistas de perfil aparecen como si fueran parábolas invertidas, y que sin embargo, la planta de la representación es una teselación, que aparece superpuesta en la parte superior.

Llamaremos $\mathcal{R}_k$ a la región asociada a la categoría k de una variable j, que



Los colores de las curvas las indentifican en 2D en el plano frontal y en 3D como corte de las superficies de regresión logística. El plano superpuesto a la representación correponde a la vista de planta del corte de las superficies.

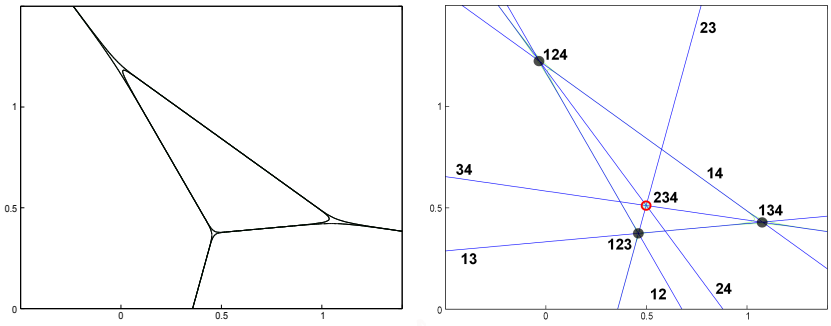
Figura 5.2: Interpretación en 3D de las líneas de la teselación.

puede definirse como:

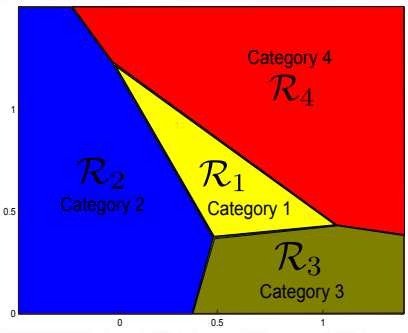
$$\mathcal{R}_k = \left\{ \mathbf{a}_h = (a_{h1}, a_{h2}) \in \mathcal{L} / \pi_{hj(k)} \geq \pi_{hj(m)}, \forall m \neq k; k, m = (1, \ldots, K_j) \right\}.$$

Las regiones de predicción para la variable hipotética citada anteriormente se muestran en la Figura 5.3(c). Es inmediato comprobar que están muy relacionadas con las curvas de nivel.

Hay que apuntar que pueden presentarse casos en los que alguna de las categorías nunca se predice, puesto que su probabilidad es inferior a la del resto de categorías de la variable, las cuales se llamarán **categorías ocultas** y que deben ser tenidas en cuenta para construir la representación final.



(a) (b)



(c)

Figura 5.3: Geometría para una solución bidimensional y una hipotética variable con 4 categorías:(a) Curvas de nivel de las superficies de respuesta para $p = 0{,}5$, (b) Líneas de igual probabilidad para cada par de categorías y puntos de intersección(candidatos para ser lados y vértices de la teselación), y (c) teselación del plano definida por las regiones de predicción.

5.1.3. Obtención de las "regiones de predicción"

En este apartado describiremos un procedimiento para obtener las regiones de predicción. El conjunto de polígonos convexos que predicen cada categoría conforman una teselación del plano. Cada celda de la teselación está delimitada por un

conjunto de líneas rectas que se correponden con puntos que tienen igual probabilidad para dos de las categorías de la variable(serán los lados).

Consideremos cada variable j $(j = 1, \ldots, J)$ de forma separada. Vamos a comprobar que cada par de superficies de respuesta definidas por (5.1) se cortan en una línea recta que proyectada en el espacio de los predictores es el conjunto de puntos en los que la probabilidad de ambas categorías es la misma. Estas líneas a priori son las candidatas a ser los lados de los polígonos convexos que definirán las regiones de predicción. Es decir, buscamos el conjunto de puntos $\mathcal{E}_{kl}$ en $\mathcal{L}$ tal que el par de categorías k y l $(k, l = 1, \ldots, K_j)$, tienen la misma probabilidad esperada, es decir, $\pi_{ij(k)} = \pi_{ij(l)}$. Por tanto $\mathcal{E}_{kl}$ es el conjunto de puntos que verifican:

$$\frac{e^{b_{j(k)0}+\sum_{s=1}^{2} b_{j(k)s}a_s}}{\sum_{m=1}^{K_j} e^{b_{j(m)0}+\sum_{s=1}^{2} b_{j(m)s}a_s}} = \frac{e^{b_{j(l)0}+\sum_{s=1}^{2} b_{j(l)s}a_s}}{\sum_{m=1}^{K_j} e^{b_{j(m)0}+\sum_{s=1}^{2} b_{j(m)s}a_s}}. \tag{5.4}$$

Entonces

$$b_{j(k)0} + \sum_{s=1}^{2} b_{j(k)s}a_s = b_{j(l)0} + \sum_{s=1}^{2} b_{j(l)s}a_s$$

ó

$$(b_{j(k)1} - b_{j(l)1})a_1 + (b_{j(k)2} - b_{j(l)2})a_2 = (b_{j(l)0} - b_{j(k)0}).$$

La anterior ecuación puede escribirse también:

$$a_2 = \frac{(b_{j(l)0} - b_{j(k)0})}{(b_{j(k)2} - b_{j(l)2})} - \frac{(b_{j(k)1} - b_{j(l)1})}{(b_{j(k)2} - b_{j(l)2})}a_1,$$

donde a_1 y a_2 son coordenadas genéricas en las dimensiones de $\mathcal{L}$, ecuación que confirma claramente que es una recta en este espacio[1]. Cada variable j tiene $\binom{K_j}{2}$ de tales líneas, como se muestra en el ejemplo de la figura 5.3(b).

[1]Si en lugar de considerar una solución bidimensional hubiéramos elegido 3 o más dimensiones latentes, en la representación final siempre nos restringiríamos a un plano conformado por dos de ellas, con lo cual el planteamiento seria el mismo que estamos describiendo una vez elegidas ambas.

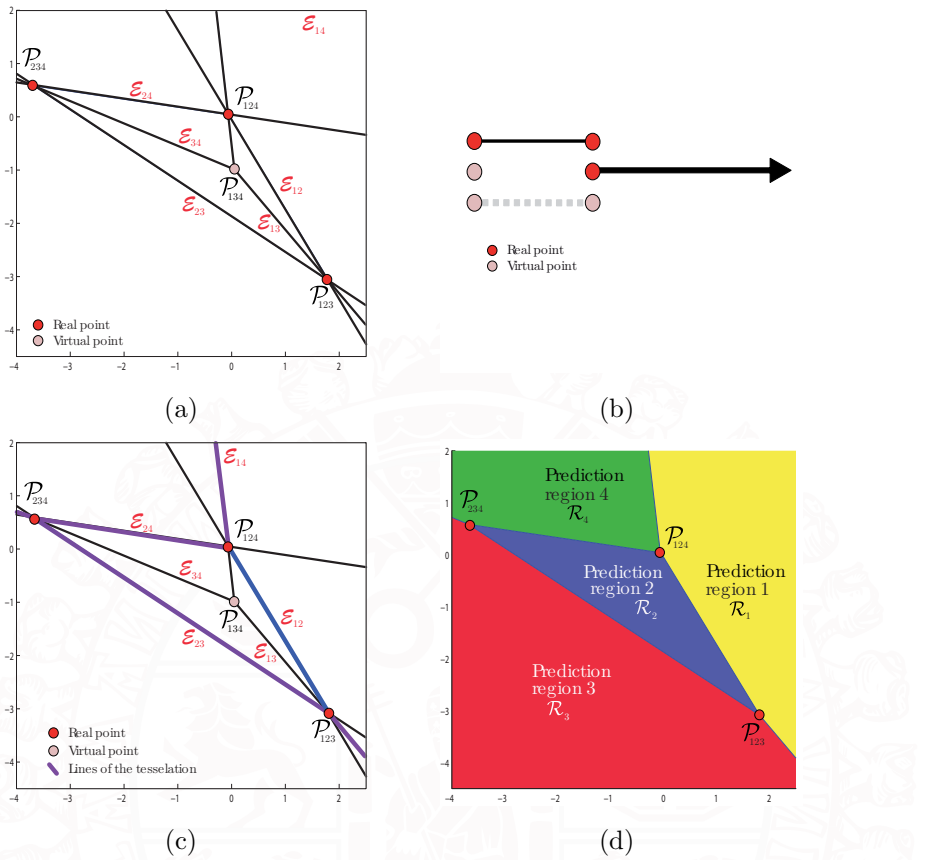


Figura 5.4: (a) Puntos reales y virtuales calculados como resultado de la comparación de $\binom{K_j}{2}$ líneas de equiprobabilidad para una variable con 4 categorías, (b) Definición de unión para construir la teselación. Rojo: punto real; Gris: punto virtual, (c) Aplicación a un caso real de la definición de unión de dos puntos candidatos dada por (b), y (d) teselación del plano definida por las regiones de predicción.

Excepto los casos degenerados, cualesquiera dos líneas con un índice en común, $\mathcal{E}_{kl}$ y $\mathcal{E}_{km}$, se cortan en un punto que denotamos $\mathcal{P}_{klm}$. Por tanto, habrá $\binom{K_j}{3}$

posibles puntos candidatos como vértices de la teselación. Un punto $\mathcal{P}_{klm}$ es un vértice de la teselación si no existe un $t \notin \{k, l, m\}$ tal que $\pi_{(\mathcal{P}_{klm})t} > \pi_{(\mathcal{P}_{klm})r}$ para $r \in \{k, l, m\}$, donde $\pi_{(\mathcal{P}_{klm})t}$ es la probabilidad esperada de la categoría t en el punto $\mathcal{P}_{klm}$, es decir, la probabilidad esperada para cada una de las categorías involucradas es la mayor posible. Si un punto es un vértice de la teselación lo llamaremos **punto real**, en otro caso será un **punto virtual**. Los casos degenerados podrían presentar líneas paralelas, aunque es extremadamente poco probable que ocurra, al igual que es muy extraño que 4 o más rectas sean coincidentes en el mismo punto.

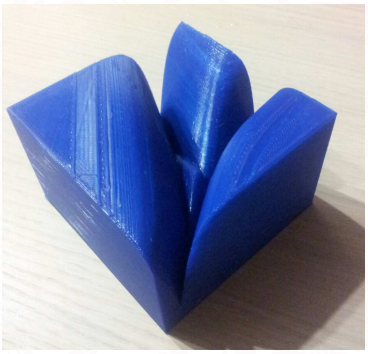


Figura 5.5: Impresión 3d por cortesía de Medialab USAL(Departamento de la Universidad de Salamanca concebido como un espacio de encuentro físico y virtual creado por el Servicio de Innovación y Producción Digital.) del conjunto de curvas para la variable con 4 categorías que estamos manejando en los ejemplos anteriores.

Las regiones de predicción $\mathcal{R}_k$ están delimitadas por todas las líneas $\mathcal{E}_{kl}$ con el índice k y sus vértices son todos los puntos $\mathcal{P}_{klm}$ con el índice k que sean puntos reales. Diremos que una categoría está oculta cuando su índice no está presente en

ninguno de los puntos reales, de tal forma que la región de esta categoría oculta no será mostrada en la representación. Definamos ahora el significado de la **unión** de dos puntos $\mathcal{P}_{klm}$ y $\mathcal{P}_{kln}$ como sigue (ver Figura 5.4(b)):

1. Dos puntos reales deberían ser unidos si tienen dos índices en común, conformando la línea $\mathcal{E}_{kl}$.

2. Dos puntos virtuales no se unen entre sí nunca.

3. Un punto virtual y un punto real se unen a lo largo de la línea $\mathcal{E}_{kl}$, comenzando desde el punto real y en dirección contraria al punto virtual.

Podemos ahora adaptar fácilmente el algoritmo descrito por Gower y Hand [1996] para construir la teselación generada por las superficies de probabilidad:

1. Calcular las coordenadas de los $\binom{K_j}{3}$ puntos $\mathcal{P}_{klm}$.

2. Decidir si el punto es real o virtual.

3. Unir todos los pares de puntos que comparten dos subíndices, interpretando el término "unión" como se describió anteriormente.

Por tanto, resumiendo gráficamente el proceso descrito en un caso sencillo daría como resultado la imagen de la figura 5.4.

El procedimiento es claramente diferente al descrito por Gower y Hand [1996] en dos aspectos fundamentales: estos autores comienzan con un conjunto de puntos $\mathcal{C}_k, k = (1, \ldots, K_j)$ que ellos llaman "Puntos Categoría (Category Level Points) (CLP)", obtenidos a partir del Análisis de Correspondencias Múltiples con algunas modificaciones, y con ellos construyen una teselación utilizando distancias; nosotros no tenemos esos puntos y utilizamos probabilidades en lugar de distancias.

Existe una configuración basada en distancias que se llama Diagrama de Voronoi y que es una estructura muy popular en el campo o disciplina de la Geometría Computacional; en este tipo de diagrama el espacio se divide en un conjunto de polígonos o regiones $\mathcal{R}_k$ de tal forma que los puntos de una región cualquiera están

más cerca de $\mathcal{C}_k$ que de cualquier otro “punto categoría”. La principal ventaja de conectar ambas áreas científicas (Geometría Computacional y Estadística Multivariante) es que si somos capaces de encontrar esos puntos proporcionaríamos una interpretación muy simple de la representación de los marcadores fila y columna de la matriz de datos, en el sentido de que la categoría predicha para cada punto es la correspondiente a su “punto categoría” más cercano. Además, si representamos los puntos en lugar de las “regiones” el gráfico resultante será mucho más claro y limpio, lo cual facilitará su lectura enormemente.

En el caso que nos ocupa tenemos las regiones, pero no los puntos, y aunque desde un punto de vista formal el problema está resuelto, puesto que tenemos una representación simultánea de individuos y variables, resultaría más conveniente poder calcular los CLP para interpretar el biplot en términos de distancias. Llamemos a este conjunto de puntos $\mathcal{C}_{j(k)}, j = (1, \ldots, J), k = (1, \ldots, K_j)$. Esta sería una contribución fundamental de nuestra investigación porque la interpretación de las distancias entre puntos fila y columna es sencilla y no es una propiedad intrínseca de la mayoría de técnicas multivariantes, como el MCA, excepto el Unfolding, que está diseñado y pensado para un propósito completamente diferente.

Con este planteamiento se plantean tres problemas:

1. ¿Es nuestra teselación un diagrama de Voronoi?

2. Si no lo es, ¿existe alguna forma de aproximarla por el diagrama de Voronoi más cercano o parecido?

3. Dada una teselación de Voronoi, ¿es posible obtener un conjunto de generadores para ella?

En el próximo apartado describiremos un procedimiento para obtener los generadores dada una teselación.

5.1.4. Cálculo de los generadores de la teselación

La construcción de diagramas de Voronoi que aproximen teselaciones de polígonos convexos del plano ha sido un campo de continuo estudio e investigación, como muestran los trabajos de Suzuki y Iri [1986], Alonso Ferrero [2011] ó Banerjee y col. [2012].

Suzuki y Iri [1986] estudian el problema de calcular los generadores de un diagrama de Voronoi dado y formulan el problema de obtener un diagrama de Voronoi que aproxime a una teselación dada del plano, de forma que plantean una función objetivo como la discrepancia entre ambas configuraciones. Dicha función no es convexa en general ni diferenciable, lo cual dificulta mucho el cálculo de su mínimo. Además, presentan un algoritmo para obtener mínimos locales, con el que se obtienen buenos resultados, en cuyas etapas hay que calcular el gradiente de dicha función, que es un proceso complejo. El problema de comprobar si una teselación de polígonos convexos es un diagrama de Voronoi y obtener sus centros ha sido estudiado también por varios autores como Evans y Jones [1987], Hartvigsen [1992],Trinchet-Almaguer y Pérez-Roses [2007] ó Aloupis y col. [2013]. Evans y Jones establecen un conjunto de ecuaciones de pendiente y de distancia que la teselación debe cumplir para ser un diagrama de Voronoi, de tal forma que resolviendo el sistema de ecuaciones lineales es posible obtener el conjunto de centros (Figura 5.6). Veamosló con un poco más de detalle.

Consideremos primeramente el siguiente resultado (omitimos el índice j de la variable para simplificar la notación) en el plano, que es el espacio que nos interesa si tenemos en mente el objetivo final de la técnica multivariante que estamos proponiendo: Una teselación de K polígonos o regiones convexas $\mathcal{R}_k, k = 1, \ldots, K$ es un diagrama de Voronoi con centros $\mathcal{C}_k = (x_k, y_k), k = (1, \ldots, K)$ si y sólo si $\mathcal{R}_k = \{(x, y) : (x - x_k)^2 + (y - y_k)^2 \leq (x - x_l)^2 + (y - y_l)^2, \forall l \neq k\}$, es decir, cada polígono de la teselación es el conjunto de puntos que están más cerca de su centro que de cualquier centro de otro polígono (figura 5.7).

Si tenemos en cuenta dos polígonos adyacentes, $\mathcal{R}_l$ y $\mathcal{R}_m$, cuyo lado común

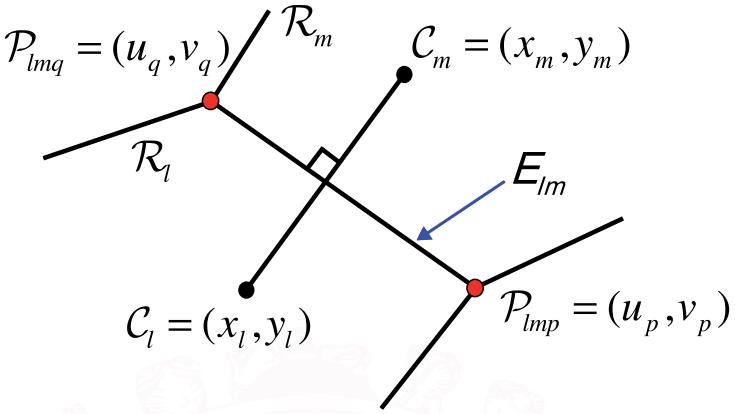


Figura 5.6: Centros $\mathcal{C}_l = (x_l, y_l)$ y $\mathcal{C}_m = (x_m, y_m)$, que son equidistantes del lado que comparten E_{lm} (5.5) y ambos se sitúan en la línea perpendicular a E_{lm} (5.6)

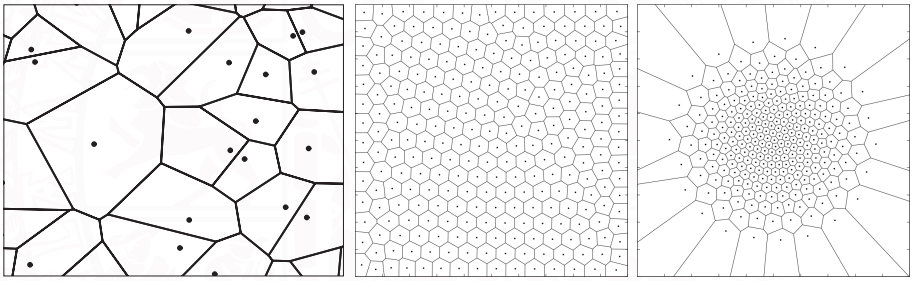


Figura 5.7: Teselaciones de Dirichlet[a] en $\mathbb{R}^2$

[a]En general las teselaciones o diagramas de Voronoi se refieren a configuraciones en un espacio genérico $\mathcal{S}$, que en el caso particular de tratarse del plano $\mathbb{R}^2$ o una parte de él se conocen como "Teselaciones de Dirichlet". Tienen diversas propiedades como que cada celda es un polígono convexo, cada vértice es el circuncentro de los centros de las celdas que comparten dicho vértice, o que la media del número de segmentos por celda no podrá ser mayor que 6. Pueden consultarse más propiedades en Okabe y col. [2000]. Hay que comentar que en las teselaciones comunes cada vértice es la intersección de exactamente tres segmentos, aunque no siempre es el caso, por ejemplo, con teselaciones que forman un retículo rectangular. Si a un vértice llegan 4 ó más segmentos se dice que es *degenerado*.

es E_{lm} de ecuación $y = s_i x + b_i$, y que contiene a los vértices (u_p, v_p) y (u_q, v_q), sean $\mathcal{C}_l = (x_l, y_l)$ y $\mathcal{C}_m = (x_m, y_m)$ los centros de Voronoi de las regiones (nuestros "puntos categoría"). Las ecuaciones de pendiente y de distancia son:

$$\frac{(y_l - y_m)}{(x_l - x_m)} = \frac{-1}{s_i} \tag{5.5}$$

$$-s_i x_l + y_l - b_i = -s_i x_m + y_m - b_i, \tag{5.6}$$

donde $s_i = \frac{(v_p - v_q)}{(u_p - u_q)}$ y $b_i = s_i u_p - v_p$.

Estas ecuaciones, si tenemos k lados y n polígonos forman un sistema lineal de ecuaciones con $2k$ ecuaciones y $2n$ incógnitas, que se puede resolver por mínimos cuadrados. En forma matricial el sistema queda:

$$\begin{cases} \mathbf{Bx} = \mathbf{0} \\ \mathbf{Ax} = \mathbf{b}, \end{cases}$$

con $\mathbf{x} = [x_1, y_1, \ldots, x_n, y_n]'$, $\mathbf{b} = -2[b_1, \ldots, b_k]'$. Las matrices $\mathbf{A}$ y $\mathbf{B}$ son dispersas, pero esto no es un inconveniente porque el número de categorías es normalmente pequeño. Los cálculos para obtener la solución se basan en tres algoritmos que pueden proporcionar diferentes centros en el caso de que los polígonos de la teselación no sean de Voronoi. Estos tres métodos son:

Algoritmo 1: Minimizar las condiciones de distancia y pendiente, es decir, buscar el $Min \left\| \begin{bmatrix} \mathbf{A} \\ \mathbf{B} \end{bmatrix} \mathbf{x} - \begin{bmatrix} \mathbf{b} \\ \mathbf{0} \end{bmatrix} \right\|^2$, siendo $\|.\|^2$ la norma euclídea.

Algoritmo 2: Minimizar $\|\mathbf{Bx}\|^2$, sujeto a $\mathbf{Ax} = \mathbf{b}$.

Algoritmo 3: Minimizar $\|\mathbf{Ax} - \mathbf{b}\|^2$, sujeto a $\mathbf{Bx} = \mathbf{0}$.

En la práctica, el principal problema con el sistema lineal de ecuaciones es la inestabilidad de los algoritmos debido al bajo condicionamiento de las matrices. Además, la propiedad de que el segmento que une los centros de dos polígonos o celdas adyacentes está cortado perpendicularmente por el segmento de la teselación que comparten las fronteras de dichas celdas, es de hecho la única condición sobre la que se proponen diversos métodos que estudian el problema de la inversión, como los de Hartvigsen [1992], que propone un algoritmo de orden polinomial

para invertir diagramas de Voronoi en $\mathbb{R}^d$ que tengan vértices de cualquier grado, Aurenhammer [1987], ó Adamatzky [1993].

Schoenberg y col. [2003], al igual que ya habían intentado otros autores como Okabe y col. [2000], trataron de encontrar los centros en cada celda teniendo en cuenta los ángulos que los vértices de la celda formaban con vértices adyacentes. En estas propuestas la idea fundamental era utilizar tanto la propiedad del bisector perpendicular citada anteriormente como la inspección y propiedades de los ángulos entre los segmentos, de forma que se mejorara la estabilidad de la solución final, utilizando en la propuesta de algoritmos el siguiente resultado (figura 5.8):

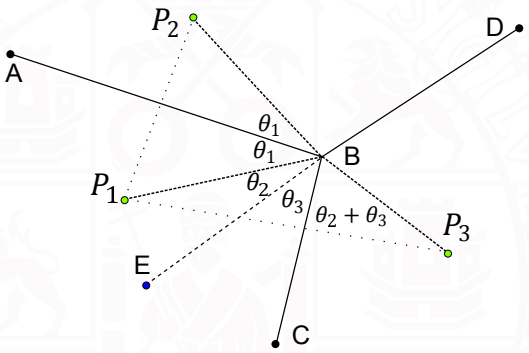


Figura 5.8: Propiedad de ángulos en los vértices no degenerados de las teselaciones de Dirichlet.

Teorema 5.1.1. *Sea T una celda de una teselación de Dirichlet en el plano $\mathbb{R}^2$. Sea P_1 el centro de la celda T. Supongamos que T tiene un vértice no degenerado B, y sean $\overline{AB}$ y $\overline{BC}$ dos segmentos de T. Sea $\overline{BD}$ el tercer segmento en la teselación, que tiene a B por uno de sus extremos, y sea E cualquier punto del interior de T, tal que E, B y D sean colineales. Entonces se verifica que:*

$$\angle ABP_1 = \angle EBC \qquad \text{y} \qquad \angle AEB = \angle P_1BC$$

En definitiva, la investigación parecía determinar la conveniencia de proponer métodos que calculen los centros de forma local más que global, puesto que los

métodos globales pueden arrastrar en la inversión los errores en la determinación de dichos centros. Schoenberg y col. [2003] proponen varios algoritmos, estudiando la estabilidad de los mismos, y resultando que el algoritmo C$'^2$ parece funcionar mejor que el de Adamatzky [1993] en situaciones en las que el número de celdas es muy elevado, y siendo además bastante rápido, con una gran precisión y con un tiempo de ejecución de orden $O(n)$.

Evans y Jones [1987] propusieron una medida de bondad de ajuste, es decir, una medida de cómo de cerca está una teselación de un diagrama de Voronoi real. Dicha medida computa la desviación local de los polígonos de Voronoi, midiendo la distancia de cada vértice (u_p, v_p) al centro del círculo (u_p^*, v_p^*) que pasa por los centros estimados de Voronoi de los polígonos apropiados (figura 5.9). El vector $[(u_p^* - u_p), (v_p^* - v_p)]$ se llama vector local de pérdida de ajuste. La longitud promedio

[2] **Algoritmo C$'$**:

Paso 1. Encontrar las celdas $C_1, C_2, \ldots, C_n$.

Paso 2. Para cada celda C_i:

(a) Encontrar todos los vértices no degenerados de C_i;

(b) Encontrar la pendiente del rayo asociado con cada uno de los vértices.

(c) Encontrar las intersecciones $S_{k,l}$ de cada par (k, l) de los rayos en la celda C_i;

(d) Para cada par (k, l) de rayos de la celda C_i, estimar la estabilidad de su intersección perturbando las pendientes de cada uno de los rayos mediante una pequeña cantidad en todas las direcciones y viendo cuánto cambian los puntos de intersección; guardar el valor $\delta_{k,l}$ que se define como la suma de los tamaños de dichos cambios.

(e) Calcular una media ponderada de los puntos de intersección dando a $S_{k,l}$ el peso $\frac{(\delta_{k,l})^{-1}}{\sum_{k',l'} \delta_{k',l'}^{-1}}$

Paso 3. Para cada celda C_i:

(a) Para cada segmento T de la celda C_i, encontrar la otra celda C_j que comparte este segmento; obtener la estimación del centro en la celda C_j calculado en el paso 2 y encontrar la imagen en espejo de este centro al otro lado del segmento T.

(b) Calcular la media de los resultados del paso 3(a) junto con las estimaciones del paso 2 para obtener un estimador refinado del centro en la celda C_i.

de dichos vectores es una medida del ajuste total, que para hacerla que no dependa de las escalas se divide la suma por el valor medio de la longitud de los bordes ($\overline{E}$) obteniendo así una medida de ajuste normalizada, τ:

$$\tau = \frac{\sum_{p=1}^{N}[(u_p - u_p^*)^2 + (v_p - v_p^*)^2]^{\frac{1}{2}}}{N \cdot \overline{E}}$$

donde N es el número de vértices interiores.

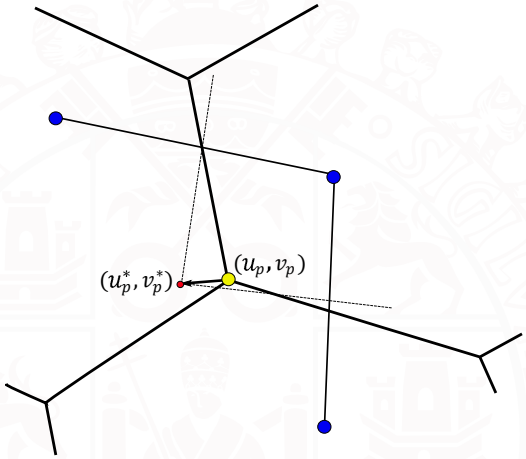


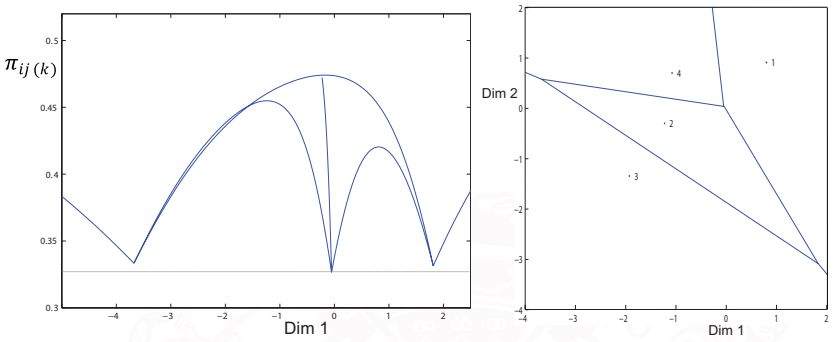
Figura 5.9: Vector local de pérdida de ajuste. Los centros estimados de la teselación se sitúan en un círculo cuyo centro es el extremo del vector. Si la teselación fuera un diagrama de Voronoi dicho vector tendría longitud cero.

Dichos autores estudian las características de este indicador con varias simulaciones estudiando su distribución.

Es necesario recalcar que en la mayoría de los casos en los que se va a utilizar esta técnica del biplot logístico nominal, las variables tienen un número bajo de categorías y los métodos de Evans y Jones [1987] tienen buenos resultados y son satisfactorios.

Para el ejemplo dado en la Figura 5.1 mostramos el resultado de la inversión

de la teselación obtenida de la respuesta logística en las figuras 5.10 y 5.11; para este caso la teselación practicamente es un diagrama de Voronoi.



(a) (b)

Figura 5.10: (a) Vista frontal de la intersección de las 4 curvas de respuesta obtenidas de la regresión logística nominal, y (b) teselación generada así como posición de los "puntos categoría" como resultado del algoritmo de inversión descrito

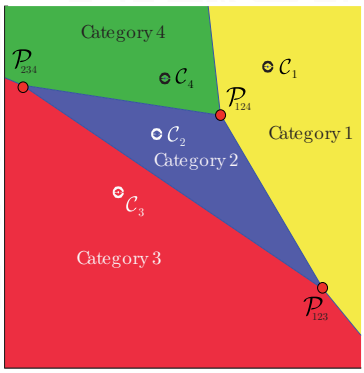


Figura 5.11: Posiciones de los CLPs ("puntos categoría") como aplicación a un caso real de una variable con 4 opciones de respuesta

Esquemáticamente, la metodología se puede comprimir en una imagen que resume un caso sencillo, que es con el que estamos ilustrando cada etapa constructiva del biplot nominal, que es la que se muestra en la figura 5.12.

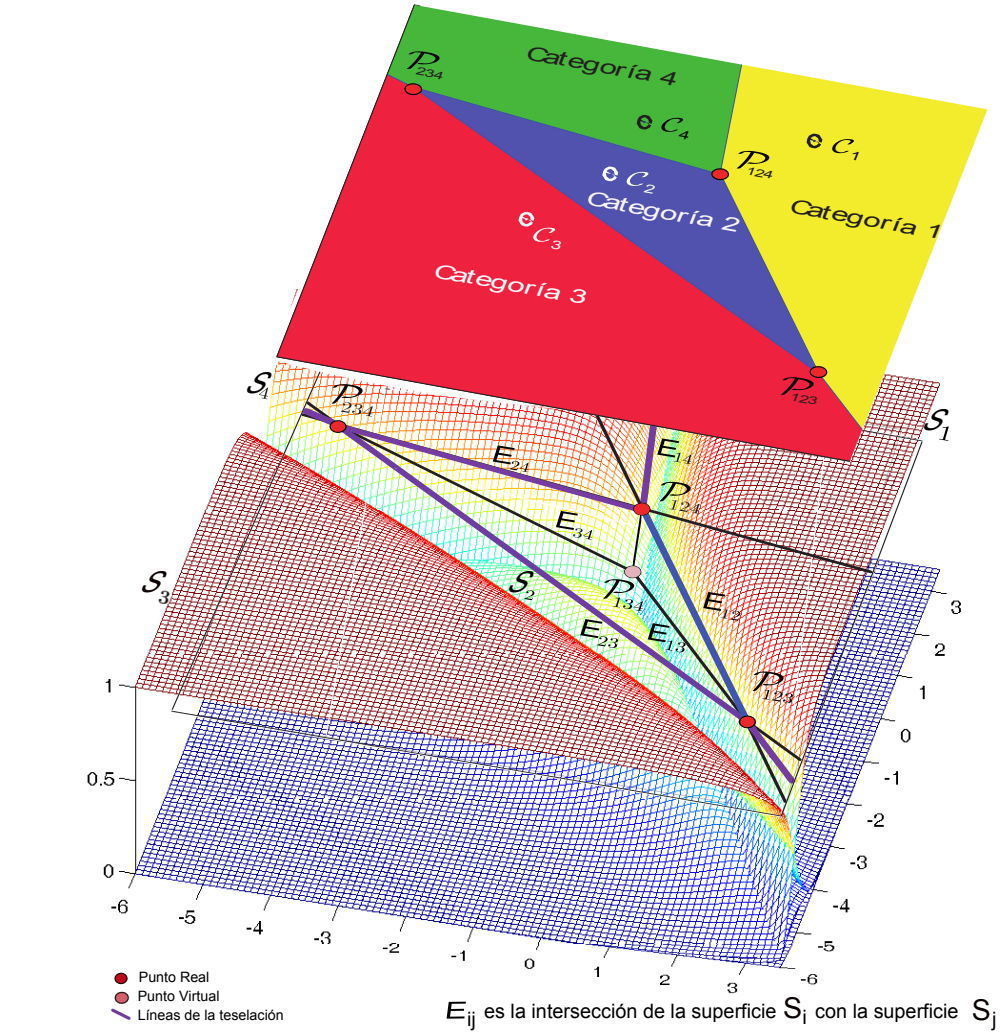


Figura 5.12: Geometría en 3D del biplot logístico nominal

5.1.5. Estimación de los parámetros del modelo

Como hemos comprobado, el caso nominal no comparte las propiedades geométricas del caso de variables binarias, sin embargo, el algoritmo alternado descrito en Vicente-Villardón y col. [2006] se puede extender fácilmente reemplazando las regresiones logísticas binarias por regresiones logísticas multinomiales. El problema de este planteamiento es que los parámetros de los individuos no pueden ser estimados cuando estos individuos presentan en todas las variables o todo ceros

o todos unos, en el caso binario, o cuando todas las respuestas son la categoria base en el caso nominal. Por este motivo utilizaremos un procedimiento que es similar al método de regresiones alternadas, excepto que la etapa de interpolación es eliminada considerando los parámetros de las filas como incidentales. La técnica asume que las puntuaciones para los individuos tienen efectos aleatorios extraídos de alguna distribución. El procedimiento de estimación que implementaremos será un algoritmo EM que utiliza la cuadratura de Gauss-Hermite para aproximar las integrales, considerando las puntuaciones de los individuos como datos faltantes. Pueden encontrarse más detalles de procedimientos similares en Bock y Aitkin [1981] ó Chalmers [2012].

La función de verosimilitud en este caso es:

$$M\left(\mathbf{G}|\,\mathbf{b}_0,\mathbf{A},\mathbf{B}\right)=\prod_{i=1}^{I}\prod_{j=1}^{J}\prod_{k=1}^{K_j}\pi_{ij(k)}^{g_{ij(k)}},$$

donde $g_{ij(k)}=1$ si el individuo i elige la categoría k de la variable j y $g_{ij(k)}=0$ en otro caso. El logaritmo de la verosimilitud se expresa:

$$L\left(\mathbf{G}|\,\mathbf{b}_0,\mathbf{A},\mathbf{B}\right)=\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K_j}g_{ij(k)}\log\left(\pi_{ij(k)}\right). \tag{5.7}$$

Como en el caso binario, si los parámetros $\mathbf{A}$ para los individuos se conocieran, el logaritmo de la verosimilitud se podría separar en J partes, una para cada variable, de tal forma que:

$$L\left(\mathbf{G}|\,\mathbf{b}_0,\mathbf{B}\right)=\sum_{j=1}^{J}L_j(\mathbf{G}|\,\mathbf{b}_{j0},\mathbf{B}_j)=\sum_{j=1}^{J}\left(\sum_{i=1}^{I}\sum_{k=1}^{K_j}g_{ij(k)}\log\left(\pi_{ij(k)}\right)\right), \tag{5.8}$$

donde $\mathbf{b}_{j0}$ y $\mathbf{B}_j$ son las submatrices de parámetros para la variable j-ésima. También apuntábamos que maximizar el logaritmo de la verosimilitud es equivalente a maximizar cada parte, y que maximizar cada L_j es equivalente a llevar a cabo una regresión logística multinomial utilizando la columna j-ésima columna de $\mathbf{X}$ como variable respuesta y las columnas de $\mathbf{A}$ como variables explicativas.

Llegados a este punto hay que tener en cuenta un aspecto crucial de la regresión logística multinomial, que es el "*problema de separación en regresión logística*", que trataremos convenientemente en el capítulo 7.4, el cual nos lleva a que en lugar de maximizar $L_j(\mathbf{G}|\,\mathbf{b}_{j0},\mathbf{B}_j)$ maximizaremos:

$$L_j(\mathbf{G}|\,\mathbf{b}_{j0},\mathbf{B}_j) - \lambda\left(\|\mathbf{b}_{j0}\|^2 + \|\mathbf{B}_j\|^2\right). \tag{5.9}$$

para obtener estimadores consistentes mediante la regresión ridge.

De la misma forma decíamos que si los parámetros para las variables fuesen conocidos se podria separar el logaritmo de la verosimilitud en I partes, una para cada individuo, de forma que:

$$L\left(\mathbf{G}|\,\mathbf{A}\right) = \sum_{i=1}^{I} L_i(\mathbf{G}|\,\mathbf{a}_i) = \sum_{i=1}^{I}\left(\sum_{j=1}^{J}\sum_{k=1}^{K_j} g_{ij(k)}\log\left(\pi_{ij(k)}\right)\right).$$

Podríamos utilizar entonces Newton-Raphson con una penalización para maximizar cada parte, pero lo que haremos es trabajar con estimadores a posteriori esperados para los marcadores de los individuos. Este proceso se describe convenientemente en el capítulo 7 que está dedicado exclusivamente a la estimación de los modelos propuestos.

5.2. Algunas aplicaciones sobre datos reales

En este apartado pasamos a utilizar la metodología propuesta para la técnica del Biplot Logístico Nominal con varios conjuntos de datos reales, con el objetivo de analizar su funcionamiento y sacar conclusiones de las aportaciones del método.

5.2.1. Las granjas de la isla de Terschelling

Vamos a considerar el conjunto de datos mostrado en la tabla 5.1, que ha sido tomado de Gower y Hand [1996], y muestra las observaciones de 4 variables observadas en 20 granjas o explotaciones en la isla holandesa de Terschelling. La

descripción de las variables y su significado se detallarán en la sección 8.3, puesto que está incluído en un paquete de R a disposición del público, de forma que dicha matriz de datos se llama Env.

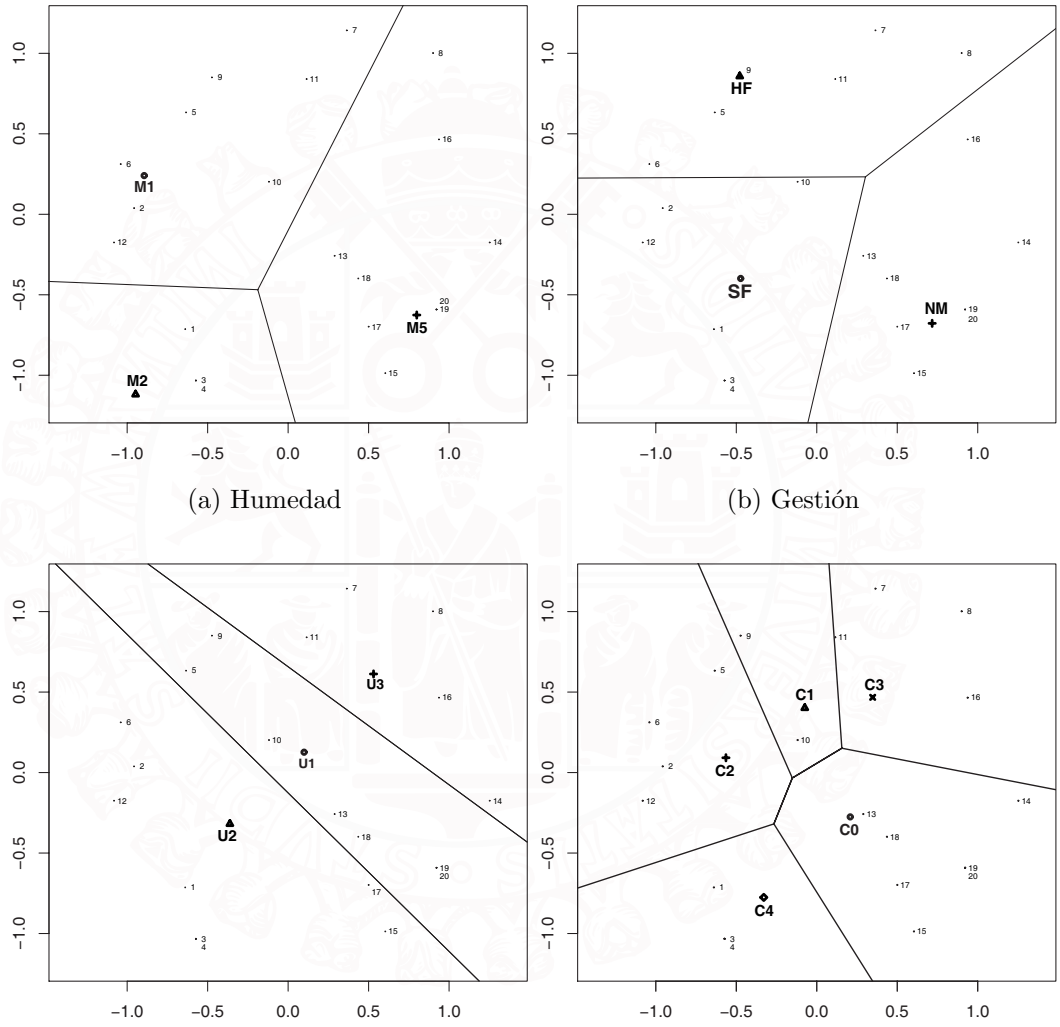
Cuadro 5.1: Datos observados para 4 variables en 20 granjas de la isla de Terschelling

Número de granja	Clase de humedad	Tipo de gestión de los pastos	Uso de los pastos	Clase de estiercol
1	1	SF	2	4
2	1	BF	2	2
3	2	SF	2	4
4	2	SF	2	4
5	1	HF	1	2
6	1	HF	2	2
7	1	HF	3	3
8	5	HF	3	3
9	4	HF	1	1
10	2	BF	1	1
11	1	BF	3	1
12	4	SF	2	2
13	5	SF	2	3
14	5	NM	3	0
15	5	NM	2	0
16	5	SF	3	3
17	2	NM	1	0
18	1	NM	1	0
19	5	NM	1	0
20	5	NM	1	0

Este conjunto de datos ha sido estudiado por Gower y Hand [1996], por lo que

con el objetivo de comparar el análisis que allí se hacía con este nuevo método vamos a considerar una solución en dos dimensiones.

Las regiones de predicción obtenidas con el algoritmo propuesto junto con los CLP asociados a dichas regiones se muestran en la figura 5.13.



(a) Humedad (b) Gestión

(c) Uso de los pastos (d) Estiercol

Figura 5.13: Regiones de predicción y puntos categoría para cada una de las 4 variables, obtenidos con NLB

Los cuatro gráficos se pueden superponer, aunque la imagen resultante sería

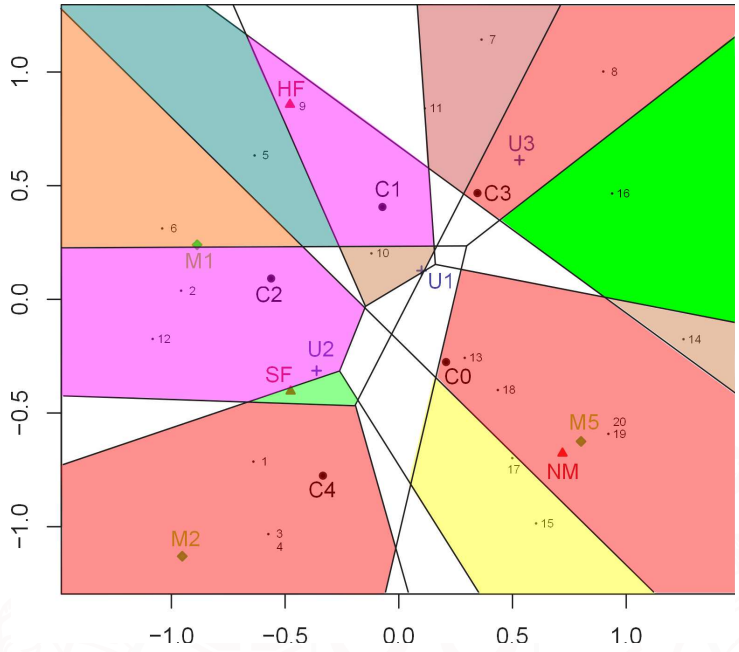


Figura 5.14: Superposición de las 4 teselaciones.

casi ilegible (Figura 5.14), asi que si tenemos más variables la interpretación sería muy complicada. El procedimiento que hemos propuesto en el que se obtienen los CLP para cada variable nos permite eliminar del gráfico las teselaciones y obtener una representación mucho más simple y sencilla, como se muestra por ejemplo en la figura 5.15.

Podemos observar que las explotaciones que tienen una "gestión natural"(NM) están en áreas con gran humedad(M5), sin fertilizantes(CO) y con producción(U1). Las granjas con una "gestión científica"(SF) están en la región con humedad M1 y M2, altos valores de fertilizantes(C4) y una utilización de los pastos intermedia(U2). Las granjas del tipo(HF) están asociadas a lugares secos(M1), usos bajos de fertilizantes(C1) y una tendencia hacia U3. Las granjas etiquetadas como BF están ocultas en el modelo de predicción porque la probabilidad de esta categoría no es nunca superior a las restantes de esta variable.

Como hemos comentado, para analizar diferencias entre el método propuesto

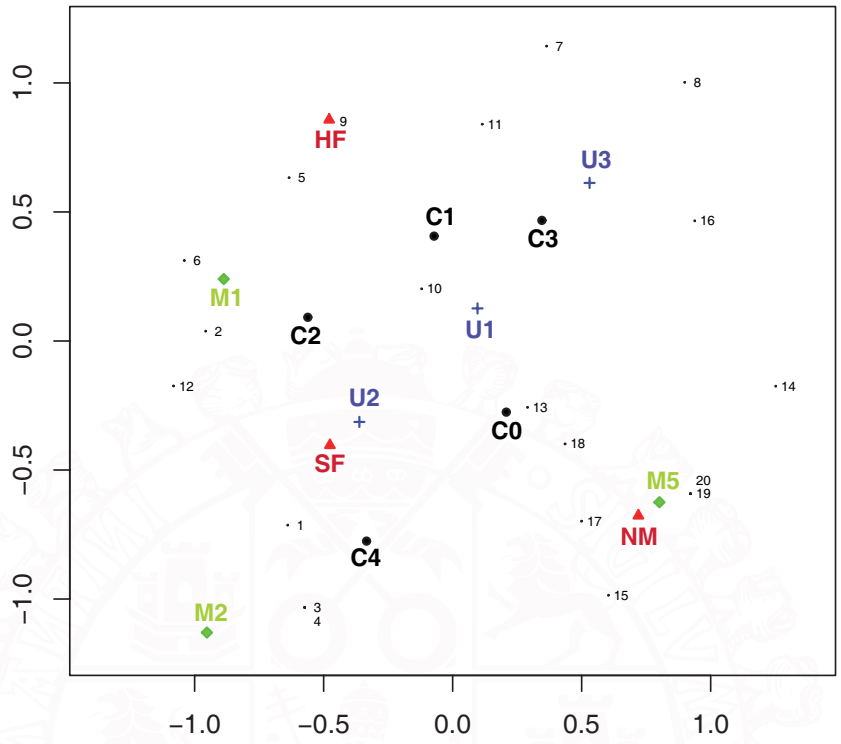


Figura 5.15: NLB bidimensional de las variables categóricas mostradas en la Tabla 5.1.

y el MCA, desde el punto de vista de Gower y Hand [1996], y teniendo diversas alternativas como métodos de estimación según hemos descrito, hemos estimado los parámetros del modelo utilizando el algoritmo EM con las modificaciones pertinentes para tener en cuenta el problema de la separación y el paquete mirt [Chalmers, 2012] con una regresión logística multinomial adicional. Las regiones de predicción obtenidas mediante nuestro método producen 14 clasificaciones incorrectas, mientras que MCA obtiene 21 incorrectas y mirt 31(ver tabla 5.2). Dicha tabla muestra los valores verdaderos y las categorías predichas para toda la matriz de datos. No hay categorías ocultas para la variable "Estiercol", pero para la "Humedad" y la "Gestión", las categorías M4 y BF, respectivamente son ocultas. Este último valor BF está presente en las explotaciones 2, 10 y 11 y ninguno de los métodos es capaz de predecirla correctamente.

Si analizamos las regiones de predicción de forma combinada para todas las variables con el método de estimación detallado, podemos ver en la figura 5.14 que se presentan 28 regiones convexas separadas. Excepto la región que contiene a las granjas 13, 18, 19 y 20, la mayoría de ellas son pequeñas y tienen pocos puntos dentro, lo que pone de manifiesto la riqueza de esta técnica para interpretar las datos. En el estudio descrito por Gower y Hand [1996], hay 16 regiones diferentes para el MCA, pero solo tres de ellas estaban claramente pobladas, así que en este caso obtenemos una clasificación más fina de las granjas.

5.2.2. Biplot Logístico Nominal(NLB) Vs Análisis de Correspondencias Múltiples(MCA)

5.2.2.1. La técnica del Análisis de Correspodencias

El MCA se puede obtener o deducir por diferentes caminos o vías. Un conjunto de problemas bastante extenso se centra en cuantificar los niveles de las categorías, es decir, asignarles puntuaciones numéricas. Este sería el enfoque de Guttman [1941], que también es conocido como análisis de homogeneidad [Gifi, 1990]. En este apartado vamos a exponer los planteamientos de esta herramienta desde el punto de vista de los biplots de Gower [Gower y Hand, 1996].

El Análisis de Correspondencias [Benzecri, 1973] estudia la asociación en una tabla de contingencia de dos vías $\mathbf{X}_{p\times q}$, analizando las desviaciones respecto de la independencia. Los elementos de $\mathbf{X}$ pueden contener cualquier valor no negativo, aunque realmente sólo los totales por fila y por columna deben ser positivos. Sean $\mathbf{R}$ y $\mathbf{C}$ las matrices diagonales con los totales por filas y columnas de la matriz $\mathbf{X}$, es decir, $\mathbf{1}'\mathbf{X}$ y $\mathbf{X1}$. La suma total de los valores de $\mathbf{X}$ es $n = \mathbf{1}'\mathbf{X1} = \mathbf{1}'\mathbf{R1} = \mathbf{1}'\mathbf{C1}$, de tal forma que las frecuencias esperadas en el modelo de independencia vienen dadas por $\mathbf{E} = \mathbf{R11}'\mathbf{C}/n$. El procedimiento del CA puede presentarse y expresarse de muchas formas similares (ver Greenacre [1984], Greenacre [1993], Greenacre [2004]), aunque nos centraremos en la aproximación de las desviaciones $\mathbf{X} - \mathbf{E}$

respecto del modelo de independencia. Una posibilidad sencilla es basar los biplots directamente en la descomposición en valores singulares de dichas desviaciones, pero se ha comprobado que ponderando, la expresión $\mathbf{R}^{-\frac{1}{2}}(\mathbf{X}-\mathbf{E})\mathbf{C}^{-\frac{1}{2}}$ tiene un interés mayor que las simples diferencias, puesto que llamando a los totales de las filas y columnas respectivamente $x_{i.}$ y $x_{.j}$ los elementos de $\mathbf{R}^{-\frac{1}{2}}(\mathbf{X}-\mathbf{E})\mathbf{C}^{-\frac{1}{2}}$ se escriben:

$$\frac{x_{ij}-\frac{x_{i.}x_{.j}}{n}}{\sqrt{x_{i.}x_{.j}}}=\frac{1}{\sqrt{n}}\left(\frac{x_{ij}-\frac{x_{i.}x_{.j}}{n}}{\sqrt{\frac{x_{i.}x_{.j}}{n}}}\right) \tag{5.10}$$

Por otra parte, en una tabla de contingencia, la hipótesis de que el modelo de independencia, $\mathbf{E}=\mathbf{R}\mathbf{1}\mathbf{1}'\mathbf{C}/n$, describe correctamente las frecuencias observadas se puede contrastar con el conocido estadístico Chi-cuadrado de Pearson (que sigue una distribución $\chi^2_{(p-1)\times(q-1)}$), expresado como:

$$\chi^2=\sum\frac{(observado-esperado)^2}{esperado}=\sum_i\sum_j\frac{(x_{ij}-\frac{x_{i.}x_{.j}}{n})^2}{\frac{x_{i.}x_{.j}}{n}} \tag{5.11}$$

que, elemento a elemento, y comparando con la expresión 5.10 ,tenemos que, multiplicando por $n^{\frac{1}{2}}$ los elementos de $\mathbf{R}^{-\frac{1}{2}}(\mathbf{X}-\mathbf{E})\mathbf{C}^{-\frac{1}{2}}$ obtenemos las raíces de lo que contribuye cada sumando al estadístico χ^2 para la tabla de contingencia. Podríamos, por tanto, tratar de minimizar $||\mathbf{R}^{-\frac{1}{2}}(\mathbf{X}-\mathbf{E})\mathbf{C}^{-\frac{1}{2}}-\hat{\mathbf{X}}||^2$ utilizando la SVD:

$$\mathbf{R}^{-\frac{1}{2}}(\mathbf{X}-\mathbf{E})\mathbf{C}^{-\frac{1}{2}}=\mathbf{U}\mathbf{\Sigma}\mathbf{V}' \tag{5.12}$$

De esta forma tenemos

$$\hat{\mathbf{X}}=\mathbf{U}\mathbf{\Sigma}\mathbf{J}\mathbf{V}'$$

con $\mathbf{J}$ una matriz diagonal con unos en sus primeras r posiciones. Por tanto, para una aproximación r-dimensional de $\hat{\mathbf{X}}$, podríamos dibujar las primeras r columnas de $\mathbf{U}\mathbf{\Sigma}^{\frac{1}{2}}$ y $\mathbf{V}\mathbf{\Sigma}^{\frac{1}{2}}$, aunque existen otras alternativas igualmente válidas.

Una variante del CA, que se usa frecuentemente, se expresa en términos de la distacia chi-cuadrado (entre filas o entre columnas), definida como:

$$d^2_{ii'}=\sum_{j=1}^{q}\frac{1}{x.j}\left(\frac{x_{ij}}{x_{i.}}-\frac{x_{i'j}}{x_{i'.}}\right)^2 \tag{5.13}$$

para las filas i-ésima e i'-ésima, o en forma matricial:

$$d^2_{ii'} = \left(\frac{\mathbf{x}_i}{x_{i.}} - \frac{\mathbf{x}'_i}{x_{i'.}}\right)\mathbf{C}^{-1}\left(\frac{\mathbf{x}_i}{x_{i.}} - \frac{\mathbf{x}'_i}{x_{i'.}}\right) \tag{5.14}$$

que se conoce como "Distancia de Mahalanobis" en la métrica de los totales por columna, entre puntos cuyas coordenadas son las proporciones por filas. Se puede demostrar que la fusión de dos filas o columnas no afecta a la distancia, propiedad que se conoce con el nombre de "Principio de equivalencia distribucional".

Las distancias chi-cuadrado entre filas dadas por 5.14 pueden interpretarse como distancias euclídeas ponderadas entre las filas de $\mathbf{X}$ y puede comprobarse fácilmente que se pueden calcular como distancias euclídeas ordinarias entre pares de filas de la matriz $\mathbf{R}^{-1}\mathbf{X}\mathbf{C}^{-\frac{1}{2}}$. Puesto que las traslaciones no afectan a la distancia, podemos efectuar la traslación $\frac{\mathbf{11'C}^{\frac{1}{2}}}{n}$ de forma que las distancias chi-cuadrado se calcularían como

$$\mathbf{R}^{-1}\mathbf{X}\mathbf{C}^{-\frac{1}{2}} - \frac{\mathbf{11'C}^{\frac{1}{2}}}{n} = \mathbf{R}^{-1}\left(\mathbf{X} - \frac{\mathbf{R11'C}}{n}\right)\mathbf{C}^{-\frac{1}{2}} = \mathbf{R}^{-1}(\mathbf{X}-\mathbf{E})\mathbf{C}^{-\frac{1}{2}} \tag{5.15}$$

y podríamos considerar una solución en la que $\hat{\mathbf{X}}$ aproxima los puntos que generan las distancias chi-cuadrado por filas:

$$||\mathbf{R}^{\frac{1}{2}}\{\mathbf{R}^{-1}(\mathbf{X}-\mathbf{E})\mathbf{C}^{-\frac{1}{2}} - \hat{\mathbf{X}}\}||^2 = ||\mathbf{U}\mathbf{\Sigma}\mathbf{V}' - \mathbf{R}^{\frac{1}{2}}\hat{\mathbf{X}}||^2 \tag{5.16}$$

la cual no está asociada a ningún tipo de ponderación por columnas. Además, la ponderación de las filas en 5.16 otorga pesos más bajos a las categorías que son menos frecuentes.

La aproximación $\hat{\mathbf{X}}$ se obtiene pues del producto $\mathbf{R}^{-\frac{1}{2}}\mathbf{U}\mathbf{\Sigma}\mathbf{V}'$ y representaríamos las primeras r columnas de $\mathbf{R}^{-\frac{1}{2}}\mathbf{U}\mathbf{\Sigma}$ para las filas y $\mathbf{V}$ para las columnas. Planteamientos similares pueden hacerse trabajando con las distancias por columnas. Frecuentemente se utilizan las dos distancias al mismo tiempo, dibujando las columnas de $\mathbf{R}^{-\frac{1}{2}}\mathbf{U}\mathbf{\Sigma}$ y $\mathbf{C}^{-\frac{1}{2}}\mathbf{V}\mathbf{\Sigma}$ a la vez como dos conjuntos de puntos. Esta visión proporciona aproximaciones a la distancia chi-cuadrado por filas y columnas, pero las relaciones entre dichos conjuntos no tienen una interpretación sencilla, y aún así se utiliza con mucha frecuencia.

Una vez introducidos los biplots para el Análisis de Correspondencias de una tabla de contingencia de 2 vías, vamos a extender la definición para el caso de tener más de dos variables categóricas, que es lo que se conoce como Análisis de Correspondencias Múltiples(MCA).

Partimos de una matriz $\mathbf{X}_{n\times p}$ cuyas columnas almacenan los niveles de p variables categóricas medidas sobre cada elemento de la muestra. Este tipo de conjuntos de datos se suele representar de forma que la información relativa a la variable k-ésima, que tiene L_k categorías, se guarda en una matriz $\mathbf{G}_k$ de dimensiones $n \times L_k$(matriz indicadora). La i-ésima fila de $\mathbf{G}_k$ tiene todo ceros excepto un uno en la columna correspondiente a la categoría que presenta el i-ésimo individuo. La suma de las columnas de $\mathbf{G}_k$ son las frecuencias de cada categoría de la variable en cuestión en los n individuos y lo denotamos por $\mathbf{L}_k$, que se considerará como matriz diagonal. Se verifica que $\mathbf{G}_k\mathbf{1} = \mathbf{1}$, $\mathbf{1}'\mathbf{G}_k = \mathbf{1}'\mathbf{L}_k$ y $\mathbf{1}'\mathbf{L}_k\mathbf{1} = n$. La matriz indicadora para el conjunto de datos completo, $\mathbf{G}$, se obtiene:

$$\mathbf{G}_{n\times L} = [\mathbf{G}_1, \mathbf{G}_2, \mathbf{G}_3, \ldots, \mathbf{G}_p]$$

con $L = L_1 + L_2 + \cdots + L_p$. Todas las filas de $\mathbf{G}$ suman p, y las sumas por columnas dan las frecuencias de todos los niveles de todas las categorías de las variables, que se pueden almacenar en una matriz diagonal $\mathbf{L}_{L\times L} = diag(\mathbf{L}_1, \mathbf{L}_2, \ldots, \mathbf{L}_p)$. Así, $\mathbf{G1} = p\mathbf{1}$, $\mathbf{1}'\mathbf{G} = \mathbf{1}'\mathbf{L}$ y $\mathbf{1}'\mathbf{L1} = np$. En este punto podemos considerar a $\mathbf{G}$ como si fuera una tabla de contingencia de dos vías y tratar las filas y columnas como si fueran variables, pudiendo aplicarse los razonamientos que detallamos para este caso.

Considerando la transformación $\mathbf{R}^{-\frac{1}{2}}\mathbf{X}\mathbf{C}^{-\frac{1}{2}}$ para el CA de una tabla de contingencia $\mathbf{X}$, si tenemos una matriz indicadora $\mathbf{G}$, en este caso sabemos que $\mathbf{R} = p\mathbf{I}$ y $\mathbf{C} = \mathbf{L}$, por tanto se define el MCA como el PCA de la matriz $\mathbf{X} = p^{-\frac{1}{2}}\mathbf{G}\mathbf{L}^{-\frac{1}{2}}$:

$$p^{-\frac{1}{2}}\mathbf{G}\mathbf{L}^{-\frac{1}{2}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}' \tag{5.17}$$

Aunque se pueden representar graficamente diversas variantes, la elección usual para el MCA es la asociada a la distancia chi-cuadrado, la cual establece la distancia

entre la i-ésima y la j-ésima fila por

$$d_{ij}^2 = \frac{1}{p}(\mathbf{g}_i - \mathbf{g}_j)\mathbf{L}^{-1}(\mathbf{g}_i - \mathbf{g}_j)' \tag{5.18}$$

siendo $\mathbf{g}_i$ la i-ésima fila de $\mathbf{G}$. De esta forma, la contribución a la distancia de la k-ésima variable para dichas filas es $\frac{1}{p}\left(\frac{1}{l_i} + \frac{1}{l_j}\right)$ y por tanto es evidente que el peso es inversamente proporcional a la frecuencia, de tal forma que respecto a los caractéres comunes, las categorías extrañas o poco frecuentes tienen pesos altos al calcular las distancias.

Hay que tener en cuenta que las componentes principales pasan por el centroide de los puntos que generan las distancias dadas por 5.18 (Principio de Huygens[3]) y suele asumirse que los datos están centrados al realizar un PCA.

La matriz $\mathbf{X}$ centrada viene dada por $\mathbf{X} - \frac{1}{n}\mathbf{1}\mathbf{1}'\mathbf{X}$ y teniendo en cuenta las propiedades y relaciones entre $\mathbf{G}$ y $\mathbf{L}$ es sencillo deducir que

$$\mathbf{X}(\mathbf{L}^{\frac{1}{2}}\mathbf{1}) = p^{-\frac{1}{2}}\mathbf{G}\mathbf{1} = p^{-\frac{1}{2}}\mathbf{1}$$
$$\mathbf{1}'\mathbf{X} = p^{-\frac{1}{2}}\mathbf{1}'\mathbf{G}\mathbf{L}^{-\frac{1}{2}} = p^{-\frac{1}{2}}\mathbf{1}'\mathbf{L}\mathbf{L}^{-\frac{1}{2}} = p^{-\frac{1}{2}}\mathbf{1}'\mathbf{L}^{\frac{1}{2}}$$

con lo cual

$$\mathbf{X}\frac{\mathbf{L}^{\frac{1}{2}}\mathbf{1}}{\sqrt{np}} = \frac{1}{\sqrt{n}}$$
$$\frac{\mathbf{1}'}{\sqrt{n}}\mathbf{X} = \frac{\mathbf{1}'\mathbf{L}^{\frac{1}{2}}}{\sqrt{np}}$$

Puesto que $\frac{\mathbf{L}^{\frac{1}{2}}\mathbf{1}}{\sqrt{np}}$ y $\frac{1}{\sqrt{n}}$ son vectores unitarios se deduce que estos son vectores singulares de $\mathbf{X}$ con valor propio 1. El teorema de Frobenius[4] establece que $\sigma = 1$

[3]La suma del cuadrado de las distancias desde cualquier punto $\mathbf{c} = (c_1, c_2, \ldots, c_p)$ al conjunto de puntos de los individuos es $||\mathbf{X} - \mathbf{1c}||$, lo cual nos lleva a la identidad

$$||\mathbf{X} - \mathbf{1c}|| = ||\mathbf{X} - \frac{1}{n}\mathbf{1}\mathbf{1}'\mathbf{X}|| + n||\frac{1}{n}\mathbf{1}'\mathbf{X} - \mathbf{c}||$$

que se hace mínimo cuando $\mathbf{c}$ es el centroide, es decir, $\mathbf{c} = \frac{1}{n}\mathbf{1}'\mathbf{X}$. Esta propiedad muestra que el subespacio de mejor ajuste en el sentido de mínimos cuadrados pasa a través del centroide de los datos.

[4]Puede consultarse su enunciado en Gower y Hand [1996],p261

es el mayor valor propio de $\mathbf{X}$, por lo que podemos escribir la SVD de dicha matriz como

$$\mathbf{X} = \frac{1}{n\sqrt{p}}\mathbf{1}\mathbf{1}'\mathbf{L}^{\frac{1}{2}} + \sum_{k=2}^{L}\sigma_k\mathbf{u}_k\mathbf{v}_k' \tag{5.19}$$

con lo cual, el primer término de 5.19 es la matriz de centrado, lo que nos lleva directamente a la conclusión de que no es necesario centrar $\mathbf{X}$ en este caso para hacer el PCA, aunque no hay que considerar el primer vector propio de $\mathbf{X}'\mathbf{X}$. Podemos pues escribir $\mathbf{X} = \frac{1}{n\sqrt{p}}\mathbf{1}\mathbf{1}'\mathbf{L}^{\frac{1}{2}} + \mathbf{U}\mathbf{\Sigma}'\mathbf{V}$

Luego los vectores $\mathbf{1}$ y $\mathbf{L}^{\frac{1}{2}}\mathbf{1}$, por las propiedades de ortogonalidad de los vectores singulares y de la SVD verifican que:

$$\mathbf{1}'\mathbf{U} = \mathbf{0}' \qquad \text{y} \qquad \mathbf{1}'\mathbf{L}^{\frac{1}{2}}\mathbf{V} = \mathbf{0}' \tag{5.20}$$

Cada variable dummie de la matriz indicadora sólo toma dos valores, o bien 0 o bien 1. Los CLPs, que tienen como coordenadas las dadas por las filas de la matriz $\mathbf{L}$, cuando se ponderan por la cantidad $\frac{1}{\sqrt{p}}\mathbf{L}^{-\frac{1}{2}}$, como marca la definición de $\mathbf{X}$, hace que se obtengan sus coordenadas relativas a los ejes principales, como las filas de

$$\mathbf{B} = \frac{1}{\sqrt{p}}\mathbf{L}^{-\frac{1}{2}}\mathbf{V} \tag{5.21}$$

Las primeras S columnas de $\mathbf{B}$ son las proyecciones de las L CLPs en el espacio S-dimensional dado por la aproximación PCA.

Los puntos que se dibujarán como resultado de las proyecciones de los individuos, que denotamos por $\mathbf{A}$ vienen dadas por:

$$\mathbf{A} = \mathbf{X}\mathbf{V} = \frac{1}{\sqrt{p}}\mathbf{G}\mathbf{L}^{-\frac{1}{2}}\mathbf{V} = \mathbf{G}\mathbf{B} \tag{5.22}$$

que significa que la posición de los individuos está centrada entre las CLPs que pertenecen a cada individuo. Si estuvieran divididas por p tendríamos los centroides, lo cual se solventa redefiniendo $\mathbf{B} = p^{\frac{1}{2}}\mathbf{L}^{-\frac{1}{2}}\mathbf{V}$ para que

$$\mathbf{A} = \frac{1}{\sqrt{p}}\mathbf{L}^{-\frac{1}{2}}\mathbf{V} = p^{-1}\mathbf{G}\mathbf{B}$$

, puesto que este reescalado no afecta a las distancias entre los individuos.

Este esquema de codificación, que utiliza variables binarias en las columnas con la restricción de que para cada variable sólo una de ellas tiene un valor uno, presenta el problema de que crea dimensiones artificiales adicionales, pues cada variable nominal se codifica mediante tantas variables binarias como categorías tenga. Como consecuencia, la inercia[5] del espacio solución se aumenta artificialmente y por tanto el porcentaje de inercia explicada por la primera dimensión se está subestimando. Pueden consultarse soluciones a este problema en Benzecri [1979] y en Greenacre [1993], así como un desarrollo más detallado y un análisis pormenorizado del estudio de esta técnica mediante la versión equivalente que utiliza la matriz de Burt en Gower y Hand [1996] y Gower y col. [2011].

5.2.2.2. La interpretación del Análisis de Correspodencias

Como en el CA, la interpretación de esta técnica se basa principalmente en las proximidades entre puntos en un espacio reducido (2 ó 3 dimensiones), las cuales tienen sentido entre puntos del mismo conjunto, es decir, se comparan filas con filas o columnas con columnas. Cuando dos individuos están cercanos en la representación significa que tenderán a seleccionar los mismos niveles de las variables nominales. Respecto a la proximidad entre variables, hay que distinguir el caso en el que hablamos de categorías de distintas variables, el cual haría referencia a que dichas categorías suelen aparecer juntas en las observaciones, y el caso de categorías de la misma variable, que como no pueden ocurrir a la vez para un individuo es necesario hacer otro tipo de interpretación. En este caso la proximidad entre ellas significa que los grupos de observaciones asociadas a estas 2 categorías son ellos mismos similares.

Vamos a presentar de una forma breve la formalización de una serie de indicadores necesarios para poder llevar a cabo la interpretación de los resultados de esta técnica. Siguiendo con la notación del apartado anterior, tenemos una matriz $\mathbf{X}_{n\times p}$ de n individuos a los que se les miden p variables categóricas. Denotamos

[5]Este concepto se presenta en la sección 5.2.2.2.

por $\mathcal{I} = \{1, 2, \ldots, n\}$ los índices de los individuos y por $\mathcal{Q} = \{1, 2, \ldots, p\}$ los de las variables, siendo L_q el número de categorías de la variable $v_q, q \in \mathcal{Q}$. Puesto que $L = \sum_{q \in \mathcal{Q}} L_q$, podemos renumerar dichas categorías desde 1 hasta L y denotar por $\mathcal{J} = \{1, 2, \ldots, L\}$ al conjunto de todas las categorías, siendo $\mathcal{J} = \mathcal{J}_1 \cup \mathcal{J}_2 \cup \ldots \mathcal{J}_p$ con $\mathcal{J}_q$ el conjunto de índices de $\mathcal{J}$ asociados a las categorías de la variable v_q. La matriz $\mathbf{G}_{n \times L}$, por tanto, es la que se muestra en la tabla 5.6, siendo para el individuo $i \in \mathcal{I}$ y la categoría $j \in \mathcal{J}$,

$$g_{ij} = \begin{cases} 1 & \text{si el individuo } i \text{ está en la categoría } j \\ 0 & \text{en otro caso} \end{cases} \tag{5.23}$$

A partir de la tabla 5.6 se pueden construir las tablas de frecuencias relativas $f_{i+} = \frac{g_{i+}}{np} = \frac{1}{n}$ y $f_{+j} = \frac{g_{+j}}{np}$ y las coordenadas de los individuos y de las categorías en los ejes principales, como hemos descrito anteriormente, que denotamos $\mathbf{A}_{(i\alpha)}$ y $\mathbf{B}_{(j\alpha)}$ respectivamente.

Se define la masa de cada punto como la frecuencia relativa de observaciones en la categoría correspondiente. Es una ponderación asignada con la finalidad de que cuando se extrae un eje, las categorías que presentan una mayor frecuencia se ven menos afectadas.

Denotamos a la *Nube de puntos de las categorías* como

$$N(\mathcal{J}) = \{(C_j, f_{+j}), j = 1, 2, \ldots, L\},$$

siendo C_j el punto columna asociado a la categoría j y f_{+j} el peso de C_j.

Introducimos en este punto el concepto genérico de *Inercia* como el estadístico que mide la dispersión de la nube de puntos. En esencia, la inercia es el promedio de las distancias de los puntos a su centro de gravedad, teniendo en cuenta las ponderaciones de los mismos según su masa.

La *Inercia de un punto columna* viene dada por:

$$I(C_j) = f_{+j} d_e^2(Y_j, O) \ ,$$

donde

$$d_e^2(Y_j, O) = \sum_{i=1}^{n} \left(\frac{f_{ij}}{f_{+j}\sqrt{f_{i+}}} - \sqrt{f_{i+}} \right)^2 .$$

Si llamamos $d_j = \frac{g_{+j}}{n}$ a la proporción de individuos que eligen la categoría j, se cumple que $d_e^2(Y_j, O) = \frac{1-d_j}{d_j}$, y por tanto la inercia del punto columna (categoría) C_j es

$$I(C_j) = \frac{1 - d_j}{p}$$

Se verifica que la *Inercia de la nube de categorías de una variable* v_q es:

$$I[N(\mathcal{J}_q)] = \sum_{j \in \mathcal{J}_q} I(C_j) = \sum_{j \in \mathcal{J}_q} \frac{1-d_j}{p} = \frac{1}{p} \sum_{j \in \mathcal{J}_q} (1 - d_j) = \frac{1}{p} \left(L_q - \sum_{j \in \mathcal{J}_q} d_j \right)$$
$$= \frac{1}{p} \left(L_q - \frac{1}{n} \sum_{j \in \mathcal{J}_q} g_{+j} \right) = \frac{L_q - 1}{p}$$

y que la *Inercia total de la nube de categorías* $N(\mathcal{J})$ es

$$I[N(\mathcal{J})] = \sum_{q \in \mathcal{Q}} I[N(\mathcal{J}_q)] = \sum_{q \in \mathcal{Q}} \frac{L_q - 1}{p} = \frac{L - p}{p}$$

De esta forma, la contribución de una categoria j de la nube $N(\mathcal{J})$ a la inercia total ($Ctr_{C_j} = \frac{1-d_j}{L-p}$) será mayor cuanto menor sea su frecuencia, motivo por el que es conveniente eliminar las categorías con frecuencias muy bajas o unirlas a categorías próximas. Además, como la contribución de una variable v_q a la inercia total $I[N(\mathcal{J})]$ ($Ctr_{v_q} = \frac{L_q - 1}{L-p}$) aumenta según su número de categorías L_q, cada variable debería tener el mismo, o un número parecido de categorías.

La *Inercia de un eje* α es

$$\mathbf{I}_\alpha = \sum_{j=1}^{L} f_{+j} (\mathbf{B}_{(j\alpha)})^2 = \sigma_\alpha$$

y la *Contribución de la categoría* j *a la inercia del eje* α es

$$ca_\alpha(j) = \frac{f_{+j} (\mathbf{B}_{(j\alpha)})^2}{\sigma_\alpha}$$

La *Contribución absoluta de una variable* v_q *a la inercia del eje* α es la suma de las contribuciones de sus categorías

$$ca_\alpha(\mathcal{J}_q) = \sum_{j \in \mathcal{J}_q} ca_\alpha(j) = \sum_{j \in \mathcal{J}_q} \frac{f_{+j}(\mathbf{B}_{(j\alpha)})^2}{\sigma_\alpha}$$

La *Contribución relativa de un eje* α *a la inercia de la categoría* j es

$$cr_\alpha(j) = \frac{(\mathbf{B}_{(j\alpha)})^2}{d_e^2(Y_j, O)} = \frac{d_j(\mathbf{B}_{(j\alpha)})^2}{1 - d_j}$$

y la *Contribución relativa de un eje* α *a la inercia de una variable* es

$$cr_\alpha(\mathcal{J}_q) = \frac{\sum_{j \in \mathcal{J}_q} f{+}j(\mathbf{B}_{(j\alpha)})^2}{I[N(\mathcal{J}_q)]} = \frac{p \sum_{j \in \mathcal{J}_q} f{+}j(\mathbf{B}_{(j\alpha)})^2}{(L_q - 1)}$$

Una vez que se han decidido el número de ejes que se retendrán, la interpretación de los resultados debe hacerse teniendo en cuenta las contribuciones relativas y absolutas. Estas últimas ayudan a dar una interpretación a los ejes principales. Primeramente se detectan las variables que más contribuyen a la construcción de los ejes.

La contribución absoluta media de las variables a la inercia del eje *alpha* es

$$\begin{aligned} \frac{1}{p} \sum_{q \in \mathcal{Q}} ca_\alpha(\mathcal{J}_q) &= \frac{1}{p\sigma_\alpha} \sum_{q \in \mathcal{Q}} \sum_{j \in \mathcal{J}_q} f_{+j}(\mathbf{B}_{(j\alpha)})^2 \\ &= \frac{1}{p\sigma_\alpha} \sum_{j \in \mathcal{J}} f_{+j}(\mathbf{B}_{(j\alpha)})^2 = \frac{1}{p} \end{aligned}$$

Se dice que una variable v_q aporta suficiente inercia al eje α si

$$ca_\alpha(\mathcal{J}_q) > \frac{1}{p}$$

Para cada variable significativa, lo que se debe hacer es observar sus categorías, de tal forma que cada una de ellas se considerará significativa si su contribución absoluta es mayor que la contribución absoluta media de las categorías de dicha variable.

Las contribuciones relativas indican cuál es la calidad de representación de las categorías por los ejes. Se suele considerar que una categoría está bien representada

en el espacio reducido si la suma de sus contribuciones relativas con los ejes es al menos de un 60 %. No es conveniente interpretar categorías que no estén bien representadas. Considerando un eje concreto, una categoría está bien representada en él si su contribución relativa es de al menos un 30 %, e igual que antes, sólo se interpreta una categoría en relación a un eje si está bien representada en él.

Un análisis análogo puede plantearse para los individuos, de forma que en la interpretación se deben tener en cuenta estos conceptos para leer correctamente los gráficos asociados a esta técnica.

5.2.2.3. NLB Vs MCA. Los doctorados en Castilla-León.

En este apartado utilizaremos un conjunto de datos que se llama PhD_nomCyL, que describiremos a continuación, y que está disponible en el paquete de R NominalLogisticBiplot, el cual se detallará posteriormente en la sección 8.

Antecedentes

Los recursos humanos son una parte esencial de la creación, comercialización y difusión de la innovación. Dentro de este gran colectivo, las personas que han obtenido el grado de doctor (nivel PhD) y que por tanto son doctores, no sólo tienen el nivel formativo más elevado, sino que están cualificados para llevar a cabo una investigación. Los antecedentes y patrones de comportamiento en cuanto a la mobilidad en el mercado de trabajo de este colectivo prácticamente son desconocidos. En este sentido, la Organización para la cooperación y el desarrollo (OECD) así como el departamento de estadística de la UNESCO, junto con la Oficina de Estadística de la Unión Europea (EUROSTAT), en 2004, pusieron de manifiesto su preocupación e intención sobre la posibilidad de estudiar estas deficiencias en cuanto a la disponibilidad de información sobre este tema, con el objetivo de desarrollar un conjunto de indicadores que fueran comparables a nivel internacional en este campo.

En 2008, un grupo de 26 países de todo el mundo, entre los que se encontraba España, considerando el año 2006 como referencia y siguiendo las directrices esta-

blecidas por los organismos citados anteriormente, comenzaron a realizar encuestas sobre este colectivo, cuyo objetivo era fundamentalmente clarificar la información sobre sus características como grupo. La mayoría de los países pioneros pertenecían a la Unión Europea, aunque también participaron miembros de la OECD como Estados Unidos y Australia. En el caso español, el Instituto Nacional de Estadística (INE) concentró todos los esfuerzos posibles parar llevar a cabo este proyecto con la intención de que la disponibilidad de información en esta parcela tuviera una continuidad en el tiempo. De esta forma, la conocida como "Encuesta sobre recursos humanos en ciencia y tecnología" fue incluída como parte del plan general de ciencia y tecnología llevado a cabo por EUROSTAT. Esta necesidad de información se plasma legislativamente en el Reglamento 753/2004 sobre Ciencia y Tecnología, el cual especifica la producción de estadisticas sobre recursos humanos en ciencia y tecnología.

Las encuestas sobre el Desarrollo Profesional de las Personas con un Doctorado (CDH) intentan medir algunos aspectos demográficos relacionados con el empleo, de tal forma que el nivel de investigación de este grupo, la actividad profesional que desarrollan, la mobilidad internacional y el salario e ingresos de este grupo se puedan cuantificar en España.

Este segmento de la fuerza laboral se considera crucial en la producción, la aplicación y distribución del conocimiento, y por tanto, clave para conseguir mejorar la competitividad de un país. Los estudios en este área están más que justificados debido a la enorme demanda de información por parte de usuarios en diferentes foros en 2003 y 2004.

Fuentes de información, población y muestra. Enfocando el estudio.

La encuesta recopila información sobre las personas que tienen un doctorado menores de 70 años que viven en España y que además obtuvieron este título en alguna universidad española pública o privada. El periodo de referencia para la estadistica es el año 2006, aunque algunas preguntas del cuestionario están relacionadas con diferentes periodos y momentos específicos del tiempo. Como marco

estadístico se utilizó un directorio de doctorados proporcionado por el Consejo de Universidades al INE, el cual incluía a todos aquellos que habian obtenido el título de doctor en alguna universidad española. La unidad de análisis es el individuo, residente en España, que ha obtenido el doctorado entre 1990 y 2006, y que sea menor de 70 años. Los doctores pertenecen al nivel 6 de la clasificación internacional de educación ISCED 97, la cual los define como el personal dedicado a programas de educación superior que conducen a una cualificación avanzada de investigación.

Estas estadísticas se basan en una herramienta de recogida comparable[6] y el cuestionario se divide en 6 módulos que tratan sobre diferentes aspectos de la carrera profesional de este grupo: formación del doctorado(EDU), ocupación de puestos en investigación previos(ECR), situación laboral(EMP), mobilidad internacional(MOB), experiencia profesional relacionada (CAR) y características personales(PER).

En cuanto al diseño de la muestra, se diseño una muestra representativa para cada comunidad autónoma al nivel NUTS-2[7], utilizando para ello un muestreo con probabilidades iguales. Los doctorados se agruparon de acuerdo a su lugar de residencia, y la selección se hizo de forma independiente para cada comunidad autonoma con un muestreo sistemático con arranque aleatorio.

Para obtener estimadores precisos a nivel nacional y regional se seleccionó una muestra de 17000 doctores. A nivel nacional, la tasa de respuesta fue de un 72 %, disponiendo así de 12193 cuestionarios válidos.

Las etiquetas de cada categoría y la pregunta en la que cada variable aparece en el cuestionario se puede ver en la tabla 5.3. Hemos considerado variables relacionadas con el estado civil(MS), el sector laboral de la organización para la que

[6]El cuestionario utilizado en el INE puede consultarse en el Apéndice J.

[7]La clasificación NUTS(Nomenclatura de Unidades Territoriales para las Estadísticas) es un sistema jerárquico que divide el territorio económico de la Unión Europea con diferentes propósitos. (ver *http : //epp.eurostat.ec.europa.eu/portal/page/nuts_nomenclature/introduction*)

trabajaba el doctorado a finales de 2006(SECT), el nivel mínimo requerido para el empleo que tenía(MIN), el nivel que se consideraba apropiado para el trabajo que desempeñaba(DES), el grado de relación entre el empleo principal y sus estudios de doctorado(PJREL), el campo científico-tecnológico que mejor se corresponde con su cualificación investigadora avanzada(FOSAT) y la principal fuente de financiación de los estudios de doctorado(SOF).

Inicialmente, algunas de las variables presentaban un gran número de categorías y las frecuencias en algunas de ellas eran muy bajas o cero, por lo que ha sido necesario agrupar aquellas que presentaban esta particularidad (por ejemplo, la fuente de financiación), de forma que no nos llevara a unas conclusiones distorsionadas. Cada una de ellas están codificados comenzando en 1 de forma ascendente en cada caso. Aunque el cuestionario presenta muchas variables binarias, que son en particular nominales, el objetivo no es el estudio de estas variables, sino de aquellas con un número mayor de categorías, puesto que el caso binario es conocido suficientemente. Además, debido a las particularidades del colectivo, en algunas preguntas que están codificadas como variables nominales no existe variabilidad alguna, puesto que un alto porcentaje de las respuestas se concentra en sólo una de las respuestas, por ejemplo, la nacionalidad, en la pregunta A.1.3. Por esta razón, aunque inicialmente se incluyeron algunas variables más, se han terminado eliminando puesto que las representaciones se reducían a un único punto.

Debido a la complejidad de los cálculos en la estimación con el algoritmo alternado, hemos enfocado el estudio a la comunidad autónoma de Castilla y León. Como hemos comentado, no existe problema alguno con la representatividad de la muestra puesto que se diseñó con este propósito. Hay 681 respuestas de doctores, que corresponden a personas que en 2006 tenían el título de doctor y que residían en esta región.

Si analizamos la matriz de correlaciones policóricas puede observarse que algunas de ellas tienen valores por encima de 0.7 y otras cercanas a 0.5, por lo que parece razonable y apropiado el uso de técnicas de reducción de la dimensionalidad

para ser capaces de interpretar la información de nuestro conjunto de datos.

Utilizando el algoritmo EM en la técnica NLB, si elegimos una solución bidimensional, teniendo en cuenta la geometría para este tipo de biplots, hemos obtenido las configuraciones de la figura 5.16. Para ello se ha utilizado el paquete de R llamado NominalLogisticBiplot desarrollado por Hernández y Vicente-Villardón [2013] y disponible en el CRAN o repositorio central, el cual será desgranado y analizado en el capítulo 8.

La determinación del número de factores apropiados en la solución final depende en gran medida de la bondad de ajuste de cada regresión logística asociada a las diferentes variables. Puesto que el objetivo en este ejemplo es la descripción del método y que en un espacio de dimensión 2 los indicadores pseudo-R^2 son en general altos hemos elegido 2 factores. Igualmente se ha calculado una solución para 3 factores, la cual eleva ligerísimamente la bondad de ajuste de las regresiones, pero complica más la interpretación y sólo debería ser considerada si el propósito es una comprensión más profunda y detallada del problema en cuestión.

Desde el punto de vista del MCA, un asunto que reviste bastante controversia es el porcentaje de varianza explicada por cada dimensión. Este problema ha sido investigado en profundidad por Blasius y Greenacre [1998], Greenacre [1993] y Greenacre [1988]. Atendiendo al porcentaje de varianza explicada por cada dimensión (figura 5.17), calculando los porcentajes por el método usual puede apreciarse que el primero de ellos es el más importante y acumula casi el 14 % de la misma, y el segundo un 7.7 % del total de la inercia. El tercer eje parece ser similar al segundo en cuanto a importancia puesto que representan un 7.5 % de la inercia. Como hemos comentado vamos a considerar una solución en dos dimensiones, que es capaz de capturar más de un 21 % de la inercia, que puede parecer pesimista, pero si tenemos en cuenta un ajuste de inercias que está explicado para un contexto práctico en Blasius y Greenacre [1998] y en Nenadić y Greenacre [2007], los porcentajes de inercia ajustados serían del 65.29 % y del 8.16 % para los dos primeros ejes. Fundamentalmente este ajuste consiste en que para cada valor propio σ_s

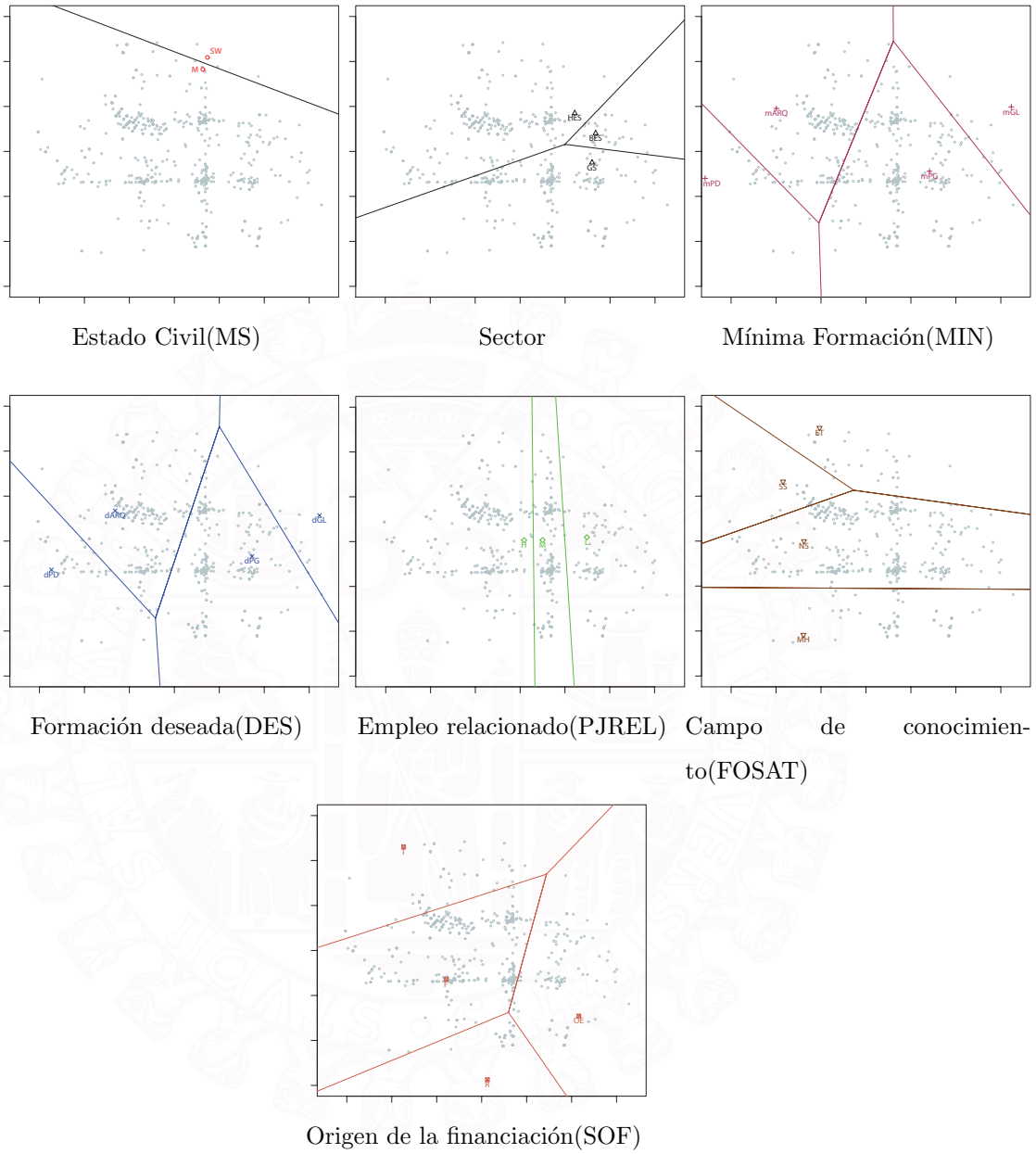


Figura 5.16: Teselaciones y puntos categoría para cada variable obtenidas con el NLB.

(obtenido del análisis de la matriz indicadora) que verifica la desigualdad $\sigma_s \leq \frac{1}{p}$ se calculan las inercias ajustadas con la fórmula $\sigma_s^{adj} = \left(\frac{p}{p-1}\right)^2 \left(\sigma_s - \frac{1}{p}\right)^2$. Estas se

expresan como un porcentaje de inercia media fuera de la diagonal, la cual puede ser calculada mediante un cálculo directo de las tablas fuera de la diagonal en la matriz de Burt, o de la inercia total de $\mathbf{C}$(matriz de Burt) utilizando:

$$\frac{p}{p-1}\left(\text{inercia}(\mathbf{C}) - \frac{L}{L-p}\right)$$

siendo inercia$(\mathbf{C}) = \sum_s \sigma_s^2$.

Por tanto, el plano $\mathcal{P}_{1-2}$ explica al menos el 73 % de la inercia total, que es suficiente para llevar a cabo una interpretación cualitativa que permita de alguna manera comparar diferentes aspectos de este método con el NLB. La fiabilidad de la escala de medida está garantizada de acuerdo con el valor del Alfa de Cronbach que toma un valor de 0.7.

Principal inertias (eigenvalues):

dim	value	%	cum%	scree plot
1	0.453797	13.8	13.8	*************************
2	0.252750	7.7	21.5	************
3	0.244874	7.5	29.0	************
4	0.201221	6.1	35.1	**********
5	0.178200	5.4	40.5	*********
6	0.166696	5.1	45.6	********
7	0.158030	4.8	50.4	*******
8	0.155400	4.7	55.1	*******
9	0.144312	4.4	59.5	*******
10	0.140719	4.3	63.8	******
11	0.134432	4.1	67.9	******
12	0.131665	4.0	71.9	******
13	0.123692	3.8	75.7	*****
14	0.121366	3.7	79.3	*****
15	0.114798	3.5	82.8	*****
16	0.104701	3.2	86.0	****
17	0.092307	2.8	88.8	****
18	0.090710	2.8	91.6	****
19	0.071735	2.2	93.8	**
20	0.064899	2.0	95.8	**
21	0.061441	1.9	97.6	**
22	0.045646	1.4	99.0	*
23	0.032322	1.0	100.0	
Total: 3.285714 100.0				

Figura 5.17: Porcentaje de inercia acumulada y gráfico de sedimentación del MCA calculado con la matriz indicadora.

La técnica del análisis de correspondencias múltiple, que está muy extendida, tiene algunas similaridades con el biplot logístico nominal, aunque la naturaleza de la técnica es bastante diferente. En el MCA las distancias entre dos individuos o dos variables no se miden de la forma usual. Se tiene en cuenta lo que se llaman perfiles medios de un individuo (fila) o de una variable (columna), y estas distancias se calculan utilizando los perfiles fila y columna, de tal forma que se pondera cada línea con una cantidad llamada masa que denota la importancia relativa de cada línea en el conjunto de datos. El perfil medio estará situado en el origen de coordenadas. Utilizando la distancia de Mahalanobis, la media de las distancias al cuadrado de cada línea (fila o columna) al centro de gravedad (origen de coordenadas) se llamaban inercia de las filas o columnas, de tal forma que los puntos cerca del origen son puntos de baja inercia, lo cual denota que son similares, mientras que inercias altas indican mayores diferencias entre la respectiva fila o columna respecto del perfil medio correspondiente. Es decir, lo que es muy frecuente estará cerca del origen y aquellas características menos comunes, con bajas frecuencias, se situarán lejos del centro de gravedad. Estas consideraciones hacen que la interpretación del MCA deba ser muy cuidadosa.

Es patente, según la tabla 5.4, y su representación gráfica (figura 5.21), que las variables SECT, MIN y DES contribuyen positivamente a definir el primer factor (eje horizontal) del espacio reducido, y en menor medida PJREL y el resto de ellas. El segundo factor (eje vertical) agrupa principalmente a las variables MIN y DES. Aunque este factor no lo explica suficientemente, las variables MS, FOSAT y SOF tienen una calidad de representación reducida, especialmente la primera de ellas, que tiene valores muy bajos, lo cual está en concordancia con lo que veremos con la técnica del NLB.

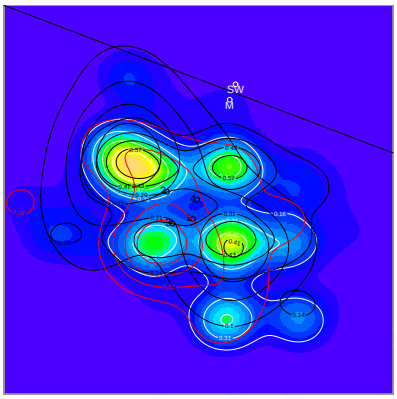
Puede apreciarse en la figura 5.19b que la categoría “Bajo” (L) de la variable “Empleo relacionado con formación” (PJREL) no está en el medio de la escala, sino hacia el menor valor de la misma, por lo menos en cuanto a la percepción de los doctorados que han respondido. De esta forma, podríamos admitir que el empleo

principal de los doctorados no está relacionado con la formación de los mismos para los sectores público (GS) y empresarial (BES).

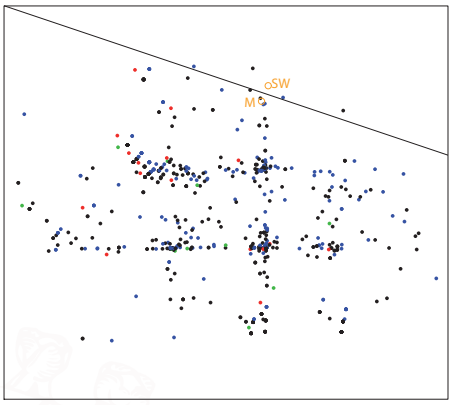
Igualmente se observa que la frecuencia de los casos en los que el nivel mínimo requerido o deseado para el empleo que ocupa el doctorado es un post-doctorado es pequeño, puesto que estos puntos aparecen muy alejados del origen.

Analizando la figura 5.18, esta muestra la estimación no paramétrica para la función de densidad en 2 dimensiones ("kernel density plot") de la variable "Estado Civil", con una coloración más clara cuánto más alta es la densidad de los puntos, y las líneas de contorno de una superficie tridimensional correspondiente a la densidad definida por los puntos correspondientes a cada una de las categorías. De esta forma se resume la información de los individuos de una manera clara. La superposición de ambas representaciones revela las categorías de la variable completamente mezcladas, razón por la que no es posible inferir conclusiones con rigor, e incluso 2 niveles están ocultos en la teselación del NLB. En dicha configuración sólo aparece una línea que separa los casados (M) de los que son solteros o viudos (SW), porque las parejas de hecho y los separados o divorciados (MLR y SD) nunca se predicen.

NLB es capaz de proporcionar y distinguir aquellas modalidades visibles de las variables, de tal forma que si una categoría no está presente en el gráfico es porque la probabilidad de la misma es menor que la de cualquier otra en el plano de representación y por tanto nunca se predecirá para un individuo de la muestra. Esta característica es distintiva y propia de esta herramienta. En la figura 5.19a puede verse la representación de nuestro método de forma conjunta sin presentar las teselaciones para cada variable. Además, como hemos comentado, la interpretación se hace en términos de la distancia euclídea, de tal forma que aquellos individuos cercanos a ciertas modalidades significa que tienen altas probabilidades de presencia en ellos, y por tanto lo que caracteriza a cada individuo son las categorías de las variables que tiene más cerca. Dos niveles de diferentes variables que están cercanos denotan que es probable que estén presentes en los mismos individuos, al

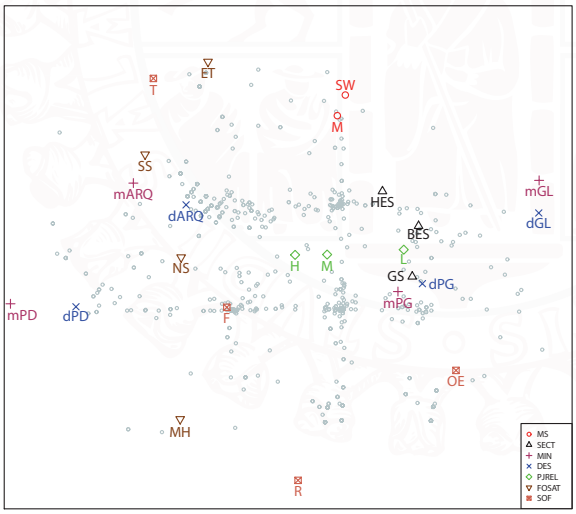


(a)Gráficos de contorno para cada categoría de la variable Estado Civil y teselación obtenida con la metodología del biplot logístico nominal.

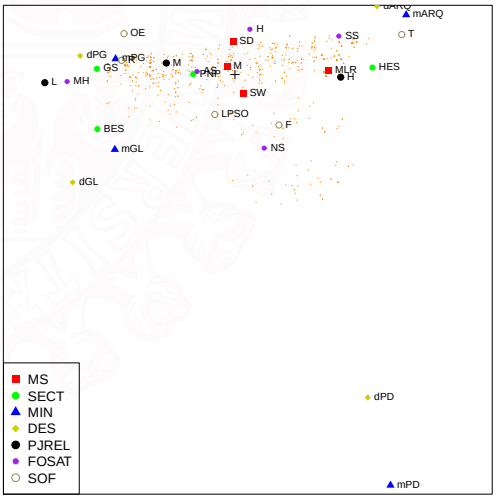


(b)Coloración de los doctorados según su Estado Civil

Figura 5.18: Estudio de la variable Estado Civil en el plano 1-2



(a)Biplot Logístico Nominal(NLB)



(b)Análisis de Correspondencias Múltiples

Figura 5.19: Representaciones NLB y MCA, en el espacio reducido, de los doctorados en Castilla y León

igual que ocurre en MCA, aunque siempre hay que tener en cuenta cómo de bueno es el ajuste de las variables.

La figura 5.20 muestra las regiones de predicción asociadas a los CLPs tal y como están descritos en Gower y Hand [1996], como parte del análisis MCA. La nube de puntos para los individuos se muestra para entender qué regiones pueden aportar información sobre el colectivo en cada variable, y está claro que no todas ellas son útiles puesto que todas las variables presentan niveles que no se utilizan en el plano. Si comparamos los porcentajes de clasificaciones correctas para cada variable con los que obtenemos con el NLB (tabla 5.5), todas ellas son más pequeñas, indicando que parece construir mejores regiones de predicción nuestra técnica.

Como ya apuntaban Gower y col. [2011], las regiones de predicción para el MCA tienen varios problemas. Para los CLPs existe una región convexa en el espacio reducido que está más cerca de cada punto categoría que del resto. Estas regiones se caracterizan porque cualquier punto de dichas regiones será predicho de tal forma que tenga la categoría asociada al punto categoría. Un problema de estas representaciones es que cada variable categórica ocupa o se extiende a todo el espacio de representación, y no sólo a un eje biplot. Esto hace impracticable, aunque no imposible, representar las regiones de predicción para más de una variable en un único diagrama o gráfico. Gower y Hand [1996] desarrollaron un algoritmo para calcular las regiones de predicción. Una alternativa sencilla es colorear cada pixel según el CLP más cercano y cubrir todo el plano para visualizar los gráficos. Es importante entender que los CLPs no se sitúan en los centroides de sus regiones de vecindad más cercanas ni tampoco son sus proyecciones. Estos autores especifican también que las proyecciones ni siquiera tienen que estar situadas dentro de las regiones de predicción y que algunos CLP pueden no tener una región de predicción en el espacio aproximado puesto que puede que se oculten detrás de otras regiones de predicción. Por lo tanto, no es obvio en la figura 5.19(b) saber qué CLP tiene una región de predicción asociada, porque es necesario analizar los mapas de las

regiones e incluso en este caso es difícil extraer la información relevante. Puesto que los CLP no necesariamente se situan dentro de sus regiones de predicción, como puede comprobarse facilmente en la figura 5.20, la interpretación es bastante complicada y tediosa.

NLB es capaz de resolver este tipo de problemas ya que la técnica averigua y calcula aquellas modalidades que se predicen y no tiene en cuenta el resto, limpiando el gráfico y facilitando la interpretación, a la vez que elimina la problemática de tener que saber si cada punto categoría pertenece o no a una región.

La figura 5.19(a) muestra que el sector educación superior (HES) abarca doctorados de ciencias sociales e ingeniería y tecnología, mientras que es menos probable la presencia de expertos en ciencias médicas. Dicho sector engloba empleos que pueden ser ocupados por personal menos cualificado que con un nivel de doctorado, aunque los doctorados con más experiencia y más cualificados es más probable que recalen en este sector que en cualquier otro. En las empresas, los doctorados ocupan puestos que podrían ser desempeñados por postgraduados, graduados o ingenieros, e igual ocurre en el sector público. Además, en el sector empresarial (BES) los doctorados están trabajando en puestos que no parecen tener una fuerte relación con sus estudios avanzados, atendiendo a la cercanía de los puntos BES y L de las variables “sector” y “empleo relacionado” respectivamente. Se puede apreciar que los doctorados que recalan laboralmente en el sector público pertenecen al campo de las ciencias médicas y de la salud y financiaron sus doctorados utilizando préstamos u otros tipos de empleos diferentes de la actividad docente. Se observa también que en ciencias naturales los doctorados es más probable que hayan disfrutado de beca escolar que en ciencias médicas. Por otra parte, los doctorados de ingeniería y tecnología es probable que financiaran su preparación mediante algun tipo de enseñanza o docencia.

En la representación mediante MCA aparecen cerca del origen categorías como casado o soltero/divorciado, y en general las relacionadas con el estado civil, el sector instituciones privadas sin fines de lucro, disciplinas AS y H y un poco

más lejanas las categorías relativas a la financiación, precisamente porque tienen poca relavancia en este plano o solución bidimensional. Para inferir conclusiones más claras de estos aspectos se deberían tener en cuenta dimensiones posteriores a la segunda que nos proporcionaran más información. Esta circunstancia ya la conocíamos por el NLB, ya que muchas de estas modalidades estaban ocultas en el gráfico o bien la calidad de representación no era muy alta y no podíamos concluir con certeza afirmaciones sobre ellas.

Es bastante interesante comprobar que las categorías con mayores calidades de representación[8] en el MCA, según la figura 5.22, son prácticamente las que se representan en el NLB (figura 5.16), es decir, aquellas que se predicen en el plano de representación, lo cual reafirma las capacidades de este método de simplificar la lectura e interpretación y de ser capaz de seleccionar y mostrar sólo aquellas variables realmente importantes.

[8]En la figura 5.22 las columnas marcadas con "cor" se refieren a los cuadrados de las correlaciones para la dimensión 1 y 2, las cuales se pueden sumar para determinar la calidad de representación, dada por la columna "qlt".

Cuadro 5.2: Predicciones para las variables categóricas de la tabla 5.1 dadas por una aproximación en 2 dimensiones. T denota el valor verdadero en dicha tabla y MCA, Mirt y AM son las predicciones utilizando las coordenadas de las filas estimadas mediante MCA, Mirt y nuestro Método Alternado(AM).

Granja	Humedad				Gestión				Pastizales				Abonos			
	T	MCA	Mirt	AM	T	MCA	Mirt	AM	T	MCA	Mirt	AM	T	MCA	Mirt	AM
1	1	2*	2*	2*	1	1	1	1	2	2	2	2	4	4	4	4
2	1	1	5*	1	2	3*	3*	1*	2	2	1*	2	2	2	0*	2
3	2	2	2	2	1	1	1	1	2	2	2	2	4	4	4	4
4	2	2	2	2	1	1	1	1	2	2	2	2	4	4	4	4
5	1	1	5*	1	3	3	3	3	1	3*	1	1	2	1*	0*	2
6	1	1	5*	1	3	3	3	3	2	2	1*	2	2	2	0*	2
7	1	1	1	1	3	3	3	3	3	1*	3	3	3	1*	3	3
8	5	1*	1*	5	3	3	3	3	3	1*	3	3	3	3	3	3
9	4	1*	1*	1*	3	3	1*	3	1	3*	2*	1	1	1	3*	1
10	2	1*	5*	1*	2	3*	4*	1*	1	1	1	1	1	1	0*	1
11	1	1	5*	1	2	3*	3*	3*	3	3	3	3	1	1	2*	3*
12	3	1*	5*	1*	1	1	3*	1	2	2	3*	2	2	2	3*	2
13	5	2*	1*	5	1	1	1	4*	2	2	2	1*	3	4*	3	0*
14	5	5	5	5	4	4	4	4	3	1*	1*	3	0	0	0	0
15	5	5	5	5	4	4	4	4	2	1*	1*	2	0	0	0	0
16	5	5	1*	5	1	1	1	4*	3	2*	3	3	3	3	3	3
17	2	5*	5*	5*	4	4	4	4	1	1	1	1	0	0	0	0
18	1	5*	5*	5*	4	4	4	4	1	1	1	1	0	0	0	0
19	5	5	5	5	4	4	4	4	1	1	1	1	0	0	0	0
20	5	5	5	5	4	4	4	4	1	1	1	1	0	0	0	0
Errores	0	8	13	6	0	3	5	5	0	7	6	1	0	3	7	2

Cuadro 5.3: Variables seleccionadas para el estudio de los doctorados en Castilla-León.

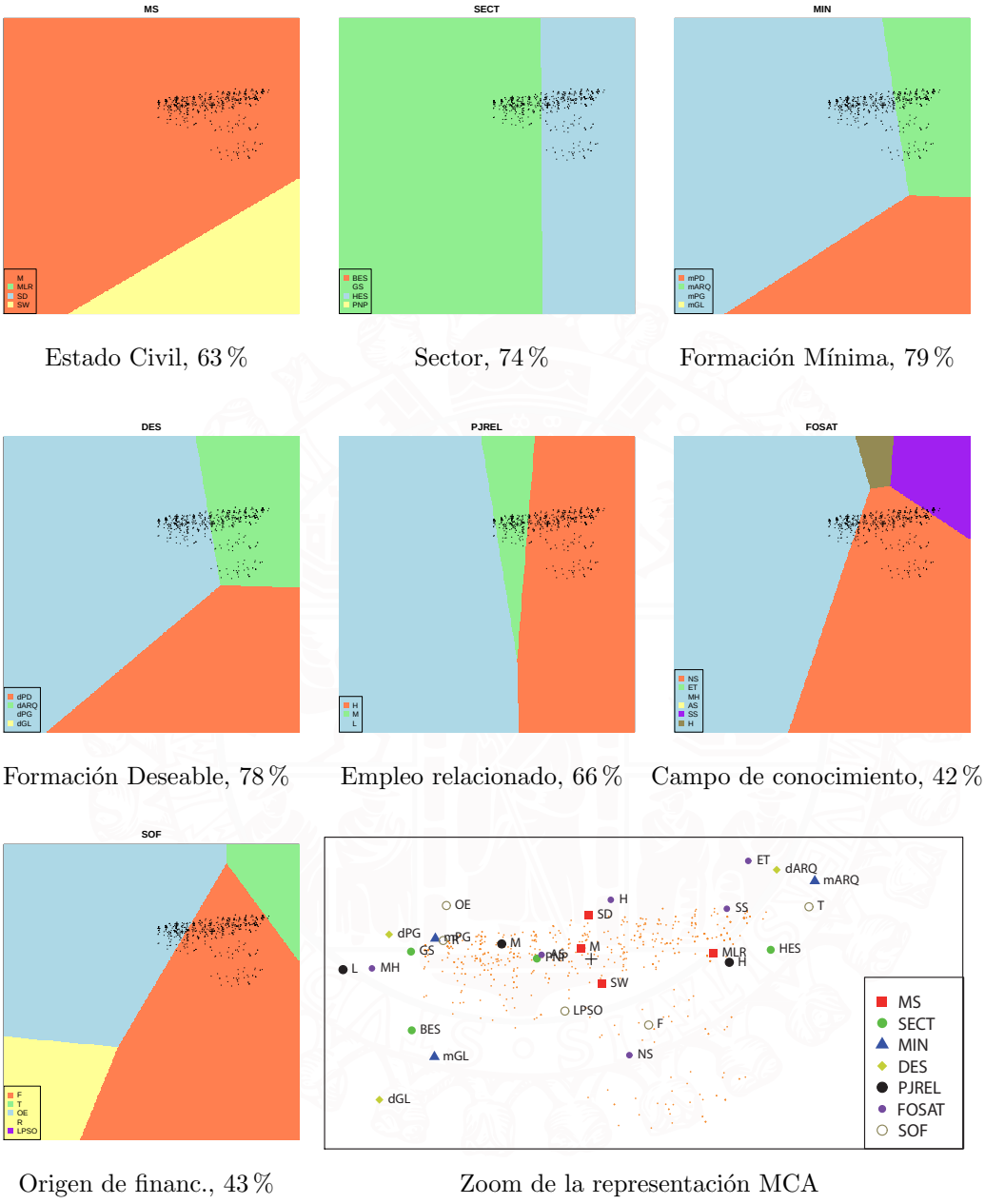
Variable	Descripción	Pregunta	Etiquetas de cada categoría
MS	Estado Civil	A.1.7	M(Casado) MLR(Pareja de hecho) SD (Separado o divorciado) SW(Viudo o soltero)
SECT	Sector de empleo	C.5.4	BES(Empresas) GS (Gobierno) HES(Educación superior) PNP(Instit. Privadas sin fines de lucro)
MIN	Nivel de formación mínimo requerido para el empleo principal	C.6.1	mPD(Postdoctorado) mARQ(Cualifiación investigadora avanzada) mPG(Post-graduado) mGL(Graduado o inferior)
DES	Nivel de formación deseado requerido para el empleo principal	C.6.2	dPD(Postdoctorado) dARQ(Cualifiación investigadora avanzada) dPG(Post-graduado) dGL(Graduado o inferior)
PJREL	Su trabajo principal está relacionadocon su grado de cualificación investigadora avanzada	C.6.3	H(Alto) M(Medio) L(Bajo)
FOSAT	Campo de ciencia y tecnología	B.2	NS(Ciencias Naturales) ET(Ingeniería y tecnología) MH(Ciencias Médicas y de la salud) AS(Agricultura) SS(Ciencias Sociales) H(Humanidades)
SOF	Fuente principal de financiación durante sus estudios de investigación	B.7	F(Beca) T(Enseñanza) OE (Otros empleos) R(Préstamos reembolsables) LPSO (Préstamos peronales u otros)

Cuadro 5.4: Medidas de discriminación proporcionadas por el MCA

Variable	Descripción	Dim 1	Dim 2
MS	Estado civil	0.009	0.015
SECT	Sector de empleo	0.608	0.032
MIN	Nivel educativo mínimo	0.666	0.682
DES	Nivel educativo deseable	0.720	0.756
PJREL	Empleo relacionado con formación	0.479	0.006
FOSAT	Campo de ciencia y tecnología	0.328	0.167
SOF	Principal fuente de financiación	0.366	0.111

Cuadro 5.5: Indicadores de bondad de ajuste para las variables seleccionadas obtenidos con la técnica NLB

Variable	Descripción	Nagelkerke R^2	% Clasificaciones correctas
MS	Estado Civil	0.04	64
SECT	Sector de empleo	0.85	85
MIN	Mínimo nivel de formación requerido para el empleo principal	0.80	84
DES	Nivel de formación deseable requerido para el empleo principal	0.78	88
PJREL	¿Está tu principal empleo relacionado con tu grado de cualificación de investigación avanzada?	0.74	76
FOSAT	Campo de conocimiento y tecnología	0.69	58
SOF	Principal fuente de financiación durante los estudios de investigación	0.34	51



Estado Civil, 63 %

Sector, 74 %

Formación Mínima, 79 %

Formación Deseable, 78 %

Empleo relacionado, 66 %

Campo de conocimiento, 42 %

Origen de financ., 43 %

Zoom de la representación MCA

Figura 5.20: Regiones de predicción del MCA con el porcentaje de clasificaciones correctas. Los gráficos de esta figura se han hecho con el paquete de R Bbipl, disponible en el libro "Understanding Biplots" de Gower y col. [2011]. Se ha llamado a una función con el nombre MCABipl con un valor en el argumento zoomval de 0.7 para la última imagen.

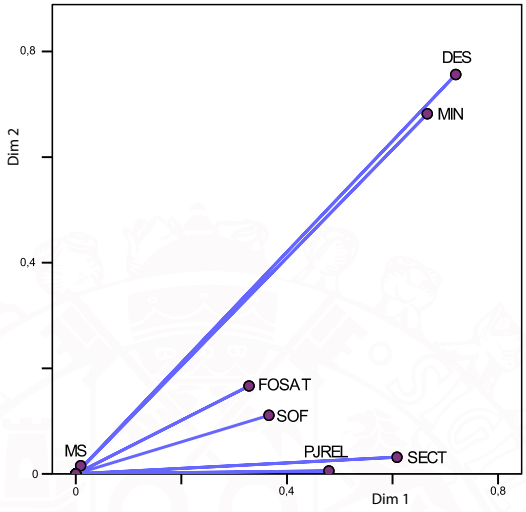


Figura 5.21: Interpretación gráfica de las medidas de discriminación.

Cuadro 5.6: Matriz indicadora $\mathbf{G}_{n\times L}$ construída a partir de $\mathbf{X}_{n\times p}$.

Ind.\ Categ.	1	2	...	L	Total
1	g_{11}	g_{12}	$\cdots$	g_{1L}	
2	g_{21}	g_{22}	$\cdots$	g_{2L}	
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
n	g_{n1}	g_{n2}	$\cdots$	g_{nL}	
Total	g_{+1}	g_{+2}	$\cdots$	g_{+L}	

							Contribuciones de				
							Puntos a la inercia de la dim		la dimensión a la inercia del punto		
					dim1	dim2					
Categoría	Frec	Masa	Inercia	qlt	cor	cor	1	2	1	2	Total
MS.M	432	,091	,052	,011	,004	,007	,000	,002	,004	,010	,013
MS.MLR	17	,004	,139	,007	,007	,000	,002	,000	,007	,000	,007
MS.SD	13	,003	,140	,001	,000	,001	,000	,001	,000	,002	,002
MS.SW	219	,046	,097	,011	,001	,010	,000	,005	,001	,014	,015
Total		,143	,429				,003	,009			
SECT.BES	70	,015	,128	,092	,071	,020	,020	,016	,071	,031	,102
SECT.GS	257	,054	,089	,382	,381	,001	,075	,001	,381	,004	,385
SECT.HES	334	,070	,073	,603	,600	,003	,096	,001	,600	,002	,602
SECT.PNP	20	,004	,139	,002	,002	,000	,001	,000	,002	,000	,002
Total		,143	,429				,192	,018			
MIN.mPD	41	,009	,134	,691	,051	,640	,015	,310	,051	,583	,634
MIN.mARQ	241	,051	,092	,645	,530	,116	,108	,039	,530	,107	,637
MIN.mPG	357	,075	,068	,534	,517	,016	,078	,007	,517	,027	,544
MIN.mGL	42	,009	,134	,053	,031	,022	,009	,030	,031	,056	,087
Total		,143	,429				,210	,386			
DES.dPD	68	,014	,129	,749	,065	,685	,018	,318	,065	,625	,689
DES.dARQ	290	,061	,082	,698	,494	,205	,089	,062	,494	,191	,684
DES.dPG	292	,061	,082	,609	,593	,016	,107	,008	,593	,025	,618
DES.dGL	31	,007	,136	,074	,041	,033	,012	,040	,041	,074	,115
Total		,143	,429				,227	,428			
PJREL.H	364	,076	,066	,424	,424	,000	,062	,000	,424	,000	,424
PJREL.M	179	,038	,105	,058	,055	,003	,013	,001	,055	,003	,058
PJREL.L	138	,029	,114	,304	,303	,001	,076	,002	,303	,004	,307
Total		,143	,286				,151	,003			
FOSAT.NS	186	,039	,104	,131	,011	,120	,002	,057	,011	,139	,150
FOSAT.ET	64	,013	,129	,084	,050	,035	,014	,017	,050	,034	,083
FOSAT.MH	153	,032	,111	,270	,269	,001	,066	,000	,269	,000	,269
FOSAT.AS	43	,009	,134	,003	,003	,000	,001	,000	,003	,000	,003
FOSAT.SS	120	,025	,118	,095	,076	,019	,020	,006	,076	,014	,090
FOSAT.H	115	,024	,119	,026	,001	,025	,000	,013	,001	,028	,030
Total		,143	,714				,103	,094			
SOF.F	247	,052	,091	,122	,036	,086	,007	,033	,036	,091	,127
SOF.T	127	,027	,116	,232	,210	,022	,054	,010	,210	,021	,231
SOF.OE	164	,034	,108	,160	,129	,031	,031	,017	,129	,040	,169
SOF.R	118	,025	,118	,091	,089	,002	,023	,001	,089	,002	,090
SOF.LPSO	25	,005	,138	,004	,001	,004	,000	,002	,001	,004	,005
Total		,143	,571				,115	,063			

Figura 5.22: Indicadores de la calidad del ajuste y contribuciones de las variables del MCA.

Capítulo 6

Biplot de Variables Ordinales

What I give form to in daylight is only one per cent of what I have seen in darkness.

– M. C. Escher

Cuando es necesario analizar en un conjunto de datos variables cuyas categorías están ordenadas, los biplots lineales, binarios o los logísticos nominales tampoco son adecuados y técnicas como el CATPCA ó la IRT serían más convenientes. Una revisión de los modelos más extendidos de teoría de respuesta al ítem para variables ordinales puede encontrarse en van der Linden y Hambleton [1997].

Existen modelos que acomodan variables con un orden en sus categorías, tales como el modelo de respuesta graduada, generalizado del modelo probit/normal o logit/normal para respuestas binarias y el modelo de crédito parcial generalizado del modelo de Rasch o del modelo logit/normal. Samejima [1969] desarrolló un modelo de variables latentes para variables ordenadas como una generalización del

modelo logit/normal. Su modelo unidimensional fue el punto de partida de investigaciones posteriores, como las de Muraki [1990] y Muraki y Carlsson [1995] sobre modelos multidimensioneales y estimación por máxima verosimilitud de modelos de respuesta graduada. Una discusión, dentro del marco de las distribuciones de la familia exponencial, sobre el modelo de respuesta graduada puede encontrarse en Moustaki [2000].

En este capítulo vamos a extender el concepto de biplot a aquellas situaciones en las que aparezcan este tipo de datos, resultando un método que llamaremos "Biplot Logístico Ordinal(OLB)". Haremos uso de lo que se conoce como modelo de odds proporcionales para construir un modelo multidimensional, que en el ámbito de la IRT se llama modelo de respuesta graduada. Estudiaremos la geometría de tales representaciones e implementaremos algoritmos computacionales para la estimación de los parámetros y de las direcciones de predicción.

6.1. El modelo ordinal

Sea $\mathbf{X}_{I\times J}$ una matriz de datos con información de I individuos medidos sobre J variables ordinales con $K_j, (j = 1, \ldots, J)$, categorías ordenadas, y sea $\mathbf{P}_{I\times L}$ la matriz indicadora, con $L = \sum_j K_j$ columnas. La matriz indicadora de dimensiones $I \times K_j$ para cada variable categórica $\mathbf{P}_j$ contiene los indicadores binarios para cada categoría, siendo $\mathbf{P} = (\mathbf{P}_1, \ldots, \mathbf{P}_J)$. Cada fila de $\mathbf{P}_j$ suma 1 y cada columna de $\mathbf{P}$ suma J. Por tanto $\mathbf{P}$ es la matriz de probabilidades observadas para cada categoría de cada variable.

Consideremos $\pi^*_{ij(k)} = P(x_{ij} \leq k)$ como la probabilidad acumulada (esperada) de que el individuo i tenga un valor igual ó más pequeño que k en la j-ésima variable ordinal, y sea $\pi_{ij(k)} = P(x_{ij} = k)$ la probabilidad(esperada) de que el individuo i tome el k-ésimo valor de la variable ordinal j-ésima. Entonces $\pi^*_{ij(K_j)} = P(x_{ij} = K_j) = 1$ y $\pi_{ij(k)} = \pi^*_{ij(k)} - \pi^*_{ij(k-1)}$ (con $\pi^*_{ij(0)} = 0$). Un modelo logístico de respuesta latente multidimensional (S-dimensional) para las probabilidades acumuladas se

puede escribir de la siguiente forma, siendo $(1 \leq k \leq K_j - 1)$:

$$\pi^*_{ij(k)} = \frac{1}{1 + e^{-\left(d_{jk} + \sum_{s=1}^{S} a_{is} b_{js}\right)}} = \frac{1}{1 + e^{-(d_{jk} + \mathbf{a}_i' \mathbf{b}_j)}} \tag{6.1}$$

donde $\mathbf{a}_i = (a_{i1}, \ldots, a_{iS})'$ es el vector de puntuaciones de la respuesta latente para el i-ésimo individuo y d_{jk} y $\mathbf{b}_j = (b_{j1}, \ldots, b_{jS})'$ los parámetros para cada ítem o variable. Se observa que lo que hacemos es definir un conjunto de modelos logísticos binarios, uno para cada categoría, donde ahora hay un término independiente (intercepto) para cada una, pero un conjunto común de pendientes para todos. En el contexto de la Teoría de la Respuesta al Ítem(IRT), esto se conoce con el nombre de *Modelo de respuesta Graduada* o *Modelo de Samejima* [Samejima, 1969]. La principal diferencia con los modelos característicos de la IRT es que no tenemos la restricción de que la probabilidad de obtener una categoría más alta deba crecer a lo largo de las dimensiones. Nuestras variables no son necesariamente ítems de un test, aunque los modelos, formalmente hablando son los mismos para ambos casos. En el caso unidimensional lo que tendríamos es un modelo con un único parámetro de discriminación b_j para todas las categorías y diferentes umbrales, fronteras o parámetros de dificultad $d_{j(k)}$. En la figura 6.1 puede verse un modelo acumulativo bidimensional. Las puntuaciones $\mathbf{a}_i$ se pueden representar en un diagrama de dispersión y usarse para detectar similaridades y diferencias entre individuos o para buscar clusters o agrupaciones de los mismos que tengan características homogéneas, es decir, la representación es similar a la obtenida con un método de escalamiento multidimensional. Veremos que los parámetros $\mathbf{b}_j$ se pueden representar sobre el gráfico como direcciones en el espacio de las puntuaciones que mejor predicen las probabilidades y se usarán para ayudar en la búsqueda de las variables responsables de las diferencias entre los individuos.

En escala logit, el modelo puede escribirse

$$logit(\pi^*_{ij(k)}) = d_{j(k)} + \sum_{s=1}^{S} a_{is} b_{js} = d_{j(k)} + \mathbf{a}_i' \mathbf{b}_j \ , \ k = 1, \ldots, K_j - 1 \tag{6.2}$$

Esta expresión define un Biplot Logístico Binario para las categorías acumuladas.

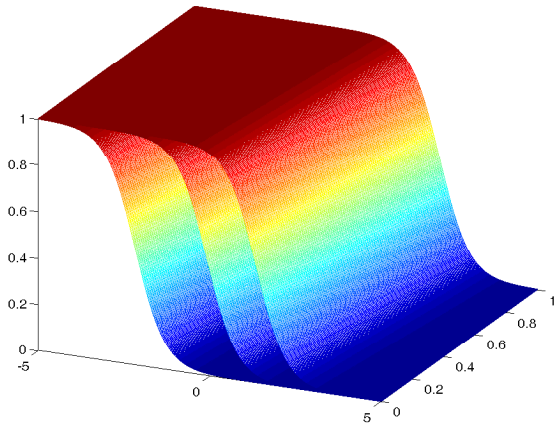


Figura 6.1: Curvas de respuesta acumuladas para un modelo de respuesta latente en dos dimensiones y para una variable con 4 categorías.

En forma matricial tenemos:

$$logit(\Pi^*) = \mathbf{1}_I \mathbf{d}' + \mathbf{A}\mathbf{B}' \tag{6.3}$$

donde $\Pi^* = (\Pi_1^*, \ldots, \Pi_J^*)$ es la matriz de tamaño $I \times (L - J)$ de probabilidades acumuladas esperadas, $\mathbf{1}_I$ es un vector de unos, $\mathbf{d} = (\mathbf{d}_1', \ldots, \mathbf{d}_J')$, con $\mathbf{d}_j' = (d_{j(1)}, \ldots, d_{j(K_j-1)})$, es el vector que contiene los umbrales, $\mathbf{A} = (\mathbf{a}_1', \ldots, \mathbf{a}_I')'$ con $\mathbf{a}_i' = (a_{i1}, \ldots, a_{iS})$ es la matriz $I \times S$ que contiene las puntuaciones de los individuos y $\mathbf{B} = (\mathbf{B}_1', \ldots, \mathbf{B}_J')'$ con $\mathbf{B}_j = \mathbf{1}_{K_j-1} \otimes \mathbf{b}_j'$ y $\mathbf{b}_j' = (b_{j1}, \ldots, b_{jS})$, es la matriz de $(L - J) \times S$ que contiene las pendientes para todas las variables. Esta expresión define un biplot para los odds que llamaremos "Biplot Logístico Ordinal". Cada ecuación del biplot acumulado comparte la geometría descrita para el caso binario [Vicente-Villardón y col., 2006], y además, todas las curvas tienen la misma dirección cuando se proyectan sobre el biplot. El conjunto de parámetros $\{d_{jk}\}$ proporcionan un umbral diferente para cada categoría acumulada, y la segunda parte de 6.2 no depende de una categoría en particular, por lo que por tanto las $K_j - 1$ curvas comparten las mismas pendientes.

La probabilidad esperada de que el individuo i responda la categoría k de la

variable j, con $(k = 1, \ldots, K_j)$, que denotamos por $\pi_{ij(k)} = P(x_{ij} = k)$ se obtendrá restando las probabilidades acumuladas:

$$\pi_{ij(k)} = \pi^*_{ij(k)} - \pi^*_{ij(k-1)}$$

y utilizando las ecuaciones dadas por 6.1 tenemos:

$$\begin{aligned}
\pi_{ij(1)} &= P(x_{ij} = 1) = \frac{1}{1 + e^{-(d_{j(1)} + \mathbf{a}_i' \mathbf{b}_j)}} \\
\pi_{ij(k)} &= P(x_{ij} = k) = P(x_{ij} \leq k) - P(x_{ij} \leq (k-1)) \\
&= \frac{1}{1 + e^{-(d_{j(k)} + \mathbf{a}_i' \mathbf{b}_j)}} - \frac{1}{1 + e^{-(d_{j(k-1)} + \mathbf{a}_i' \mathbf{b}_j)}} \\
&= \frac{e^{-(\mathbf{a}_i' \mathbf{b}_j)} (e^{-d_{j(k-1)}} - e^{-d_{j(k)}})}{(1 + e^{-(d_{j(k)} + \mathbf{a}_i' \mathbf{b}_j)})(1 + e^{-(d_{j(k-1)} + \mathbf{a}_i' \mathbf{b}_j)})}, \ 1 < k < K_j \\
\pi_{ij(K_j)} &= P(x_{ij} = K_j) = 1 - \frac{1}{1 + e^{-(d_{j(K_j - 1)} + \mathbf{a}_i' \mathbf{b}_j)}}
\end{aligned} \tag{6.4}$$

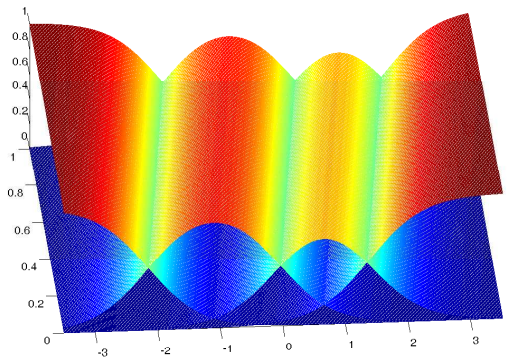


Figura 6.2: Curvas de respuesta para una variable ordinal con 4 categorías.

Si las puntuaciones de las filas o individuos se conocieran, la obtención de los parámetros del modelo en 6.4 es equivalente al ajuste de un modelo de odds proporcionales utilizando cada variable como la variable respuesta y las puntuaciones de las filas como los regresores. Las superficies de respuesta para tal modelo podrían corresponderse en un caso sencillo a las que se muestran en la figura 6.2.

6.2. Geometría y obtención de la representación biplot.

En las siguientes secciones detallaremos la geometría en el caso general y un algoritmo para llevar a cabo los cálculos.

6.2.1. Descripción de la Geometría

Las superficies que podíamos ver en la figura 6.2 no son sigmoides, aunque las curvas se cortan en líneas rectas, por lo que el conjunto de puntos sobre la representación (generados por las columnas de $\mathbf{A}$) que predicen un valor particular de la probabilidad de una categoría se sitúan en una línea recta, y diferentes probabilidades para todas las categorías de una variable se sitúan en líneas rectas paralelas. Una perpendicular a todas estas líneas se puede utilizar como el "eje del biplot", como en Gower y Hand [1996], de tal forma que esta es la dirección que mejor predice las probabilidades de todas las categorías, en el sentido de que proyectando cualquier individuo sobre dicha dirección, obtendríamos la predicción óptima de las probabilidades asociadas a cada categoría. Como todas las categorías comparten la misma dirección en el biplot, sería muy difícil situar una escala graduada para cada una y lo que haremos es representar únicamente los segmentos de la recta en los que la probabilidad de una categoría es mayor que las probabilidades del resto. Esto nos llevará, excepto para algunos casos patológicos o muy extraños, a tener tantos segmentos como categorías (K_j), separados por $K_j - 1$ puntos en los que las probabilidades de dos categorías contiguas son iguales. En la figura 6.3 pueden apreciarse las líneas rectas paralelas representando los puntos que predicen las mismas probabilidades para dos categorías adyacentes y una línea, perpendicular a todas ellas que es el "Eje del biplot". Las tres líneas paralelas dividen el espacio de aproximación en 4 regiones, cada una de las cuales predice una categoría de la variable. Para el biplot no necesitamos el conjunto completo de rectas, sino sólo el

eje y los puntos sobre él, como hemos comentado, que son las intersecciones de las fronteras de las regiones de predicción.

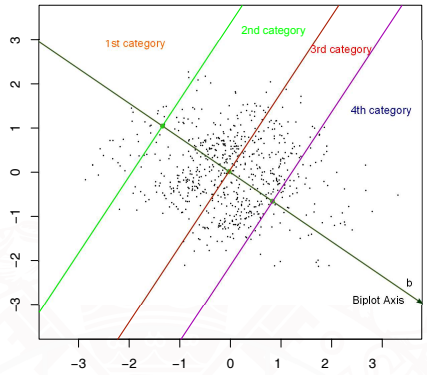


Figura 6.3: Regiones de predicción determinadas por tres líneas rectas paralelas para una variable ordinal con 4 categorías.

Llamenos (x, y) a uno de los puntos en los que se cortan una de las líneas paralelas con el eje del biplot, el cual deberá estar en la dirección del biplot, es decir,

$$y = \frac{b_{j2}}{b_{j1}} x \tag{6.5}$$

y la probabilidad de dos categorías, posiblemente contiguas (por ejemplo l y m) en este punto debe ser la misma,

$$\pi_{j(l)} = \pi_{j(m)} \quad (\pi^*_{j(l)} - \pi^*_{j(l-1)} = \pi^*_{j(m)} - \pi^*_{j(m-1)}). \tag{6.6}$$

Omitimos el índice i porque las probabilidades son para un punto general y no para un individuo en particular. Utilizando la condición dada por 6.5 podemos reescribir las probabilidades acumuladas (o sus *logit*) como

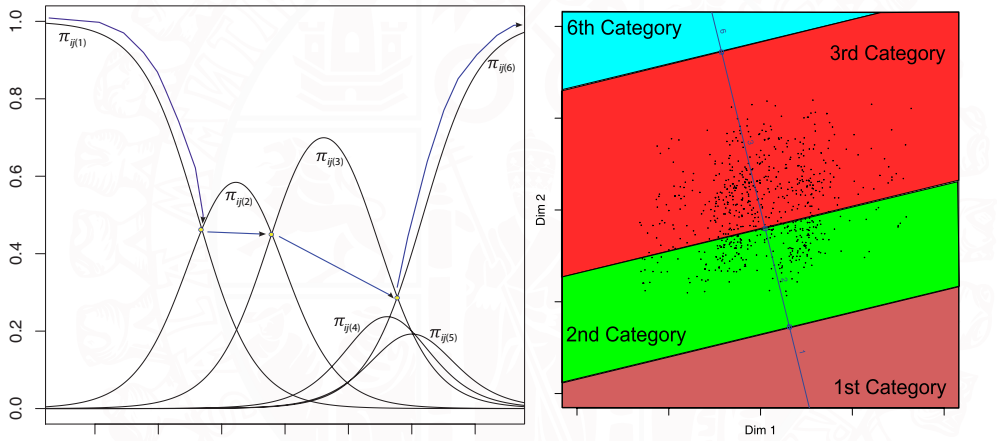
$$logit(\pi^*_{j(k)}) = d_{j(k)} + xb_{j1} + yb_{j2} = d_{j(k)} + z \tag{6.7}$$

con

$$z = x \left(\frac{b^2_{j1} + b^2_{j2}}{b_j} \right) \tag{6.8}$$

Cambiando los valores de z podemos obtener las probabilidades de cada categoría a lo largo del eje del biplot. Por tanto, encontrar el punto (x, y) es equivalente a encontrar los valores de z en los que se verifica 6.6. A partir de estos valores, el punto original se obtiene resolviendo x en 6.8 y calculando después y de 6.5.

Pueden presentarse casos patológicos en los que la probabilidad de una o varias categorías no sean nunca superiores a las probabilidades del resto, de tal forma que esas categorías serán "ocultas" o "no predichas" y, consecuentemente, el número de puntos que se sitúan en el eje biplot será menor que $K_j - 1$ (figura 6.4). Estos casos tienen que tenerse en cuenta a la hora de calcular los puntos de intersección.



(a) (b)

Figura 6.4: Curvas de probabilidad para una variable con 6 categorías en la que dos de ellas (la 4 y la 5) están ocultas. (a) Representación de las curvas de respuesta en el plano correspondiente al eje del biplot. (b) Representación biplot final sin las categorías ocultas.

La existencia de casos anormales significa o supone que no sólo hay que comparar categorías contiguas, sino que todos los pares de categorías se deben estudiar, y por tanto pueden presentarse diversos casos en este proceso de comparación, en el que intervienen las siguientes categorías:

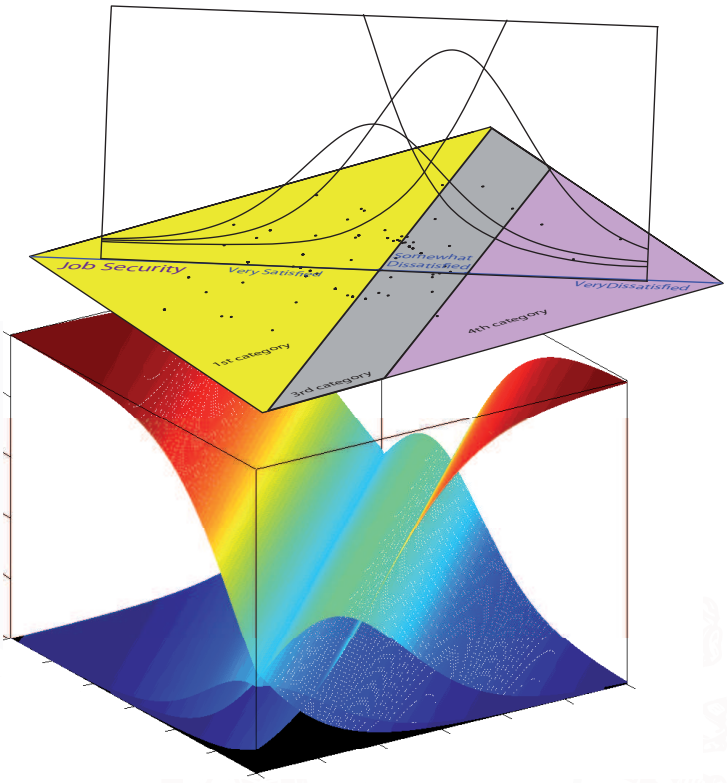


Figura 6.5: Representación-3D de las superficies de probabilidad de una variable con 4 categorías en la que la segunda no se predice nunca.

1. 1-2
2. 1-$l(2 < l < K_j)$
3. 1-K_j
4. l-$K_j(l > 1)$
5. l-$(l + 1)$ con $l > 1$
6. l-j con $j > (l + 1), l > 1$
7. $(K_j - 1)$-K_j

Las ecuaciones detalladas de cada escenario se mostrarán posteriormente en la sección 6.2.2.

Así pues, una ilustración en 3D de la geometría de estos biplots sería la que muestra la figura 6.5, y que de alguna forma resume los conceptos que se han desarrollado en un ejemplo sencillo, viendo lo que ocurre para una variable ordinal con 4 categorías, como parte de un análisis de un conjunto de ellas en un espacio reducido de dimensión 2.

Llegados hasta este punto, un procedimiento para calcular la representación de una variable ordinal en el biplot sería el siguiente:

1. Calcular el eje del biplot con la ecuación $y = \frac{b_{j2}}{b_{j1}}x$.

2. Calcular los puntos de intersección z y después (x, y) del eje del biplot con las líneas rectas paralelas utilizadas como fronteras de las regiones de predicción para cada par de categorías, en el siguiente orden:

$$\pi_{j(1)} = \pi_{j(2)}$$
$$\pi_{j(l-1)} = \pi_{j(l)} \ , 1 < l < (K_j - 1)$$
$$\pi_{j(K_j-1)} = \pi_{j(K_j)}$$

3. Si los valores de z están ordenados, entonces no se presentarán categorías ocultas y los cálculos finalizan.

4. Si los valores de z no están ordenados, podemos proceder de la siguiente forma:

   a) Calcular los valores de z para todos los pares de curvas y las probabilidades asociadas para las dos categorías involucradas.

   b) Comparar cada categoría con las siguientes (comenzando con la primera), de tal forma que la siguiente que se representará será la que tenga una probabilidad más alta en el punto de intersección.

c) Si la categoría siguiente resulta ser la K_j el proceso ha concluído, y sino, volvemos al paso anterior, comenzando con la nueva categoría.

Se podría alternativamente desarrollar un algoritmo más simple para evitar la resolución explícita de las ecuaciones de la sección 6.2.2, de la siguiente forma:

1. Calcular la categoría predicha para un conjunto de valores de z. Por ejemplo, para una secuencia desde -6 hasta 6 en intervalos de 0.001. (La precisión del procedimiento se podría cambiar con la longitud de los intervalos)
2. Buscar los valores de z en los que la predicción cambia de una categoría a otra.
3. Calcular la media de los dos valores de z obtenidos en el paso anterior y a partir de ella calcular los valores (x, y). Estos serían pues los puntos que estábamos buscando.

Las categorías ocultas serían aquellas que tienen frecuencias cero en las predicciones obtenidas por el algoritmo.

6.2.2. Ecuaciones que definen los puntos de las regiones del biplot ordinal.

Como ya comentamos, debido a la existencia de casos patológicos es necesario efectuar todas las comparaciones posibles entre las categorías:

1. 1-2
2. 1-$l (2 < l < K_j)$
3. 1-K_j
4. l-$K_j (l > 1)$
5. l-$(l + 1)$ with $l > 1$

6. l-j with $j > (l+1), l > 1$

7. $(K_j - 1)$-K_j

Llamando (x, y) a uno de esos puntos de intersección, sabemos que debe estar en la dirección del eje del biplot, es decir:

$$y = \frac{b_{j2}}{b_{j1}} x \tag{6.9}$$

Luego las probabilidades acumuladas, utilizando la expresión anterior, serían: $logit(\pi^*_{j(k)}) = d_{j(k)} + xb_{j1} + yb_{j2} = d_{j(k)} + z$, con

$$z = x\left(\frac{b_{j1}^2 + b_{j2}^2}{b_{j1}}\right) \tag{6.10}$$

Decíamos también que modificando los valores de z podíamos obtener las probabilidades de cada categoría a lo largo del eje del biplot. Luego para situar el punto (x, y) hay que encontrar los valores de z en los que se cumple $\pi_{j(l)} = \pi_{j(m)}$.

Analizamos pues cada uno de estos casos de forma separada, imponiendo la igualdad de las probabilidades de dos categorías en estos puntos:

$$\pi_{j(l)} = \pi_{j(m)} \quad (\pi^*_{j(l)} - \pi^*_{j(l-1)} = \pi^*_{j(m)} - \pi^*_{j(m-1)}) \tag{6.11}$$

y prescindimos el subíndice i puesto que los cálculos no son de ningún individuo, sino que corresponden a un punto general.

6.2.2.1. Caso 1-2

$$\begin{aligned}
\pi_{j(1)} &= \pi_{j(2)} \\
\frac{1}{1+e^{-(d_{j(1)}+z)}} &= \frac{e^{-z}(e^{-d_{j(2-1)}} - e^{-d_{j(2)}})}{(1+e^{-(d_{j(2)}+z)})(1+e^{-(d_{j(2-1)}+z)})} \\
1 + e^{-(d_{j(2)}+z)} &= e^{-(d_{j(1)}+z)} - e^{-(d_{j(2)}+z)} \\
1 &= e^{-z}(-2e^{-d_{j(2)}} + e^{-d_{j(1)}}) \\
-z &= log\left(\frac{1}{e^{-d_{j(1)}} - 2e^{-d_{j(2)}}}\right) \\
z &= log\left(e^{-d_{j(1)}} - 2e^{-d_{j(2)}}\right)
\end{aligned}$$

Por tanto, en algunas ocasiones no se cortarán, pues para ello es necesario que la expresión $e^{-d_{j(1)}} - 2e^{-d_{j(2)}}$ sea positiva, que no siempre tiene por qué cumplirse.

6.2.2.2. Caso 1-l $(2 < l < K_j)$

Al comparar la primera categoría de la variable con la l-ésima categoría, tenemos que resolver $\pi_{j(1)} = \pi_{j(l)}$, es decir,

$$\frac{1}{1+e^{-(d_{j(1)}+z)}} = \frac{e^{-z}(e^{-d_{j(l-1)}} - e^{-d_{j(l)}})}{(1+e^{-(d_{j(l)}+z)})(1+e^{-(d_{j(l-1)}+z)})}$$

Llamando

$$w = e^{-z}$$

hay que resolver la ecuación cuadrática

$$\alpha w^2 - \beta w - 1 = 0$$

con $\alpha = (e^{-(d_{j(1)}+d_{j(l-1)})} - e^{-(d_{j(1)}+d_{j(l)})} - e^{-(d_{j(l-1)}+d_{j(l)})})$ y $\beta = 2e^{-d_i}$.

Si las raíces de la ecuación son ambas negativas, entonces las dos curvas no se cortan. Si la ecuación tiene una raíz positiva, podemos calcular el punto de intersección obteniendo dicha raíz w y deshaciendo las transformaciones citadas anteriormente para obtener (x, y).

6.2.2.3. Caso 1-K_j

Es sencillo deducir que

$$z = \frac{-(d_{j(K_j-1)} + d_{j(1)})}{2}$$

porque la igualdad $\pi_{j(1)} = \pi_{j(K_j)}$ implica:

$$\frac{e^{(d_{j(1)}+z)}}{1+e^{(d_{j(1)}+z)}} = \frac{1}{(1+e^{(d_{j(K_j-1)}+z)})}$$

$$1+e^{(d_{j(1)}+z)} = e^{(d_{j(1)}+z)} + e^{(d_{j(1)}+d_{j(K_j-1)}+2z)}$$

$$2z + d_{j(1)} + d_{j(K_j-1)} = 0$$

6.2.2.4. Caso l-K_j $(l > 1)$

Tenemos que resolver $\pi_{j(l)} = \pi_{j(K_j)}$, es decir

$$\frac{e^{-z}(e^{-d_{j(l-1)}} - e^{-d_{j(l)}})}{(1 + e^{-(d_{j(l)}+z)})(1 + e^{-(d_{j(l-1)}+z)})} = \frac{e^{-(d_{j(K_j-1)}+z)}}{1 + e^{-(d_{j(K_j-1)}+z)}}$$

Operando y reordenando los términos, si llamamos

$$w = e^{-z}$$

obtenemos la ecuación de segundo grado en w:

$$e^{-(d_{j(l)}+d_{j(l-1)}+d_{j(K_j-1)})}w^2 + 2e^{-(d_{j(l)}+d_{j(K_j-1)})}w - e^{-d_{j(l-1)}} + e^{-d_{j(l)}} + e^{-d_{j(K_j-1)}} = 0$$

Resolviendo esta ecuación y teniendo en cuenta una discusión similar a la del caso 1-l $(l < K_j)$ obtenemos:

$$z = -log(w)$$

6.2.2.5. Caso l-$(l+1)$, $(l > 1)$

La igualdad $\pi_{j(l)} = \pi_{j(l+1)}$, o escrita de otra forma:

$$\frac{e^{-z}(e^{-d_{j(l-1)}} - e^{-d_{j(l)}})}{(1 + e^{-(d_{j(l)}+z)})(1 + e^{-(d_{j(l-1)}+z)})} = \frac{e^{-z}(e^{-d_{j(l)}} - e^{-d_{j(l+1)}})}{(1 + e^{-(d_{j(l+1)}+z)})(1 + e^{-(d_{j(l)}+z)})}$$

arroja, reordenando términos, y llamando $w = e^{-z}$:

$$w = \frac{e^{-d_{j(l-1)}} - 2e^{-d_{j(l)}} + e^{-d_{j(l+1)}}}{e^{-(d_{j(l-1)}+d_{j(l)})} - 2e^{-(d_{j(l-1)}+d_{j(l+1)})} + e^{-(d_{j(l+1)}+d_{j(l)})}}$$

, pudiendo calcularse z como en los casos anteriores.

6.2.2.6. Caso l-m ,$(m > (l+1), l > 1)$

En este caso tenemos que resolver la igualdad $\pi_{j(l)} = \pi_{j(m)}$, es decir:

$$\frac{e^{-z}(e^{-d_{j(l-1)}} - e^{-d_{j(l)}})}{(1 + e^{-(d_{j(l)}+z)})(1 + e^{-(d_{j(l-1)}+z)})} = \frac{e^{-z}(e^{-d_{j(m-1)}} - e^{-d_{j(m)}})}{(1 + e^{-(d_{j(m)}+z)})(1 + e^{-(d_{j(m-1)}+z)})}$$

Si agrupamos los términos adecuadamente y operamos, sustituyendo $w = e^{-z}$, se obtiene la ecuación cuadrática:

$$(e^{-(d_{j(l-1)}+d_{j(m-1)}+d_m)} - e^{-(d_{j(l)}+d_{j(m)}+d_{j(m-1)})} - e^{-(d_{j(l-1)}+d_{j(l)}+d_{j(m-1)})}$$

$$+e^{-(d_{j(l)}+d_{j(m)}+d_{j(l-1)})})w^2+$$

$$2(e^{-(d_{j(m)}+d_{j(l-1)})} - e^{-(d_{j(l)}+d_{j(m-1)})})w-$$

$$e^{-d_{j(l)}} - e^{-d_{j(m-1)}} + e^{-d_{j(l-1)}} + e^{-d_{j(m)}} = 0$$

6.2.2.7. Caso $(K_j - 1)$-K_j

De nuevo, tenemos que resolver $\pi_{j(K_j-1)} = \pi_{j(K_j)}$,

$$\frac{e^{-z}(e^{-d_{j(K_j-2)}} - e^{-d_{j(K_j-1)}})}{(1+e^{-(d_{j(K_j-1)}+z)})(1+e^{-(d_{j(K_j-2)}+z)})} = \frac{e^{-(d_{j(K_j-1)}+z)}}{(1+e^{-(d_{j(K_j-1)}+z)})}$$

La solución de la ecuación, con $w = e^{-z}$, es:

$$w = \frac{e^{-d_{j(K_j-2)}} - 2e^{-d_{j(K_j-1)}}}{e^{-(d_{j(K_j-1)}+d_{j(K_j-2)})}}$$

6.2.3. Propiedad geométrica del máximo de las curvas en el modelo de respuesta graduada.

Es sencillo demostrar que $\pi_{j(1)}$ crece con z porque su derivada parcial respecto de z es siempre mayor que cero. De la misma forma se puede comprobar que $\pi_{j(K_j)}$ decrece a lo largo de z.

También es obvio, teniendo en cuenta la expresión de $\pi_{j(k)}$, con $K_j > k > 1$, que se verifica que $d_k > d_{k-1}$, porque:

$$\pi_{j(k)} = P(x_{ij} = k) = P(x_{ij} \leq k) - P(x_{ij} \leq (k-1)) = \frac{1}{1+e^{-(d_{j(k)}+z)}} - \frac{1}{1+e^{-(d_{j(k-1)}+z)}}$$

$$\pi_{j(k)} > 0 \Leftrightarrow 1+e^{-(d_{j(k)}+z)} < 1+e^{-(d_{j(k-1)}+z)} \Leftrightarrow d_k > d_{k-1}$$

Este resultado es de sobra conocido, puesto que los umbrales en el modelo de respuesta gradual o graduada están ordenados.

Calculemos ahora el máximo de las curvas $\pi_{j(k)}$, con $K_j > k > 1$:

$$\pi_{j(k)} = \frac{e^{-z}(e^{-d_{j(d-1)}} - e^{-d_{j(k)}})}{(1+e^{-(d_{j(k)}+z)})(1+e^{-(d_{jk-1)}+z)})} = \frac{A}{B \cdot C}$$

$$0 = \frac{\partial \pi_{j(k)}}{\partial z} = \frac{\frac{\partial A}{\partial z} \cdot B \cdot C - A \cdot \frac{\partial (B \cdot C)}{\partial z}}{((1+e^{-(d_{j(k)}+z)})(1+e^{-(d_{jk-1)}+z)}))^2} \Leftrightarrow$$

$$-e^{-z}(e^{-d_{j(k-1)}} - e^{-d_{j(k)}})(1+e^{-(z+d_{j(k)})})(1+e^{-(z+d_{j(k-1)})}) - e^{-z}(e^{-d_{j(k-1)}} - e^{-d_{j(k)}})\frac{\partial (B \cdot C)}{\partial z} = 0$$

Como $\frac{\partial (B \cdot C)}{\partial z} = -2 - e^{-(z+d_{j(k-1)})} - e^{-(z+d_{j(k)})}$, sustituyendo en la última expresión y operando tenemos:

$$1 = e^{-(2z+d_{j(k)}+d_{j(k-1)})} \Leftrightarrow 2z + d_{j(k)} + d_{j(k-1)} = 0 \Leftrightarrow$$

$$z = \frac{d_{j(k)} + d_{j(k-1)}}{2}$$

Puede comprobarse con la segunda derivada de $\pi_{j(k)}$ que estos puntos son máximos, es decir, que $\left(\frac{\partial^2 \pi_{j(k)}}{\partial z^2} < 0\right)$.

Con este resultado, como los umbrales están ordenados ascendentemente, los máximos de las curvas también están ordenados, es decir, el máximo de la i-ésima curva no puede situarse a la izquierda de los máximos de las curvas precedentes.

6.3. Estimación de los parámetros del modelo.

Utilizando como referencia el algoritmo alternado para variables binarias detallado en Vicente-Villardón y col. [2006] y reemplazando las regresiones logísticas

binarias por regresiones logísticas ordinales, con las particularidades que ya comentabamos al inicio de la sección 5.1.5 que detallaba la estimación de parámetros para el caso nominal, se podrían estimar los parámetros del modelo propuesto.

Siguiendo un planteamiento análogo al caso nominal maximizaremos:

$$L_j(\mathbf{P}|\,\mathbf{d}_j, \mathbf{b}_j) - \lambda \left(\|\mathbf{d}_j\|^2 + \|\mathbf{b}_j\|^2 \right). \tag{6.12}$$

en lugar de $L_j(\mathbf{P}|\,\mathbf{d}_j, \mathbf{b}_j)$, suponiendo conocidos los valores de $\mathbf{A}$.

Por otra parte, cuando los parámetros de las variables sean conocidos se estimarán las habilidades de los individuos mediante una alternativa del algoritmo EM que se presentará en la sección 7.2. Los detalles de este procedimiento de estimación están descritos en el capítulo 7.7.

6.4. Bondad de ajuste del modelo ordinal

El logaritmo de la verosimilitud se puede utilizar como medida de la bondad de ajuste global, especialmente cuando se llevan a cabo comparaciones entre diferentes modelos que contengan, por ejemplo, distinto número de parámetros o dimensiones. Se podrían utilizar tests estadísticos similares a los propuestos en el contexto de los modelos de la IRT, en particular, los tests basados en regresiones logísticas ordinales. Mair y col. [2008] proponen tests de este tipo para variables binarias cuya generalización al caso de variables cuyas categorías están ordenadas sería sencillo. Por una parte, este tipo de tests estadísticos tienen numerosos problemas, pero por otra parte nosotros estudiamos el procedimiento como un modelo exploratorio descriptivo, y por tanto estamos menos interesados en tests estadísticos globales y mucho más en encontrar indicadores de la bondad de ajuste o tests para cada variable de forma separada. Para los modelos de la IRT, los ítems o variables están muy relacionados con una o varias dimensiones latentes y todas son útiles para describir dichos factores latentes, de tal manera que es necesario que haya una bondad de ajuste global adecuada. En una situación exploratoria más general

algunas de las variables no serán adecuadas para describir el problema, lo cual conllevará un peor ajuste que incluso podría producir interpretaciones erróneas del fenómeno que se está estudiando. Demey y col. [2008] llevaron a cabo un estudio de simulación en el que se añadieron variables irrelevantes y de ruido a una estructura conocida, y mostraban que era posible identificar esa estructura incluso cuando la bondad de ajuste global no era alta. Utilizando índices de ajuste para cada variable eran capaces de identificar las variables relevantes y eliminar aquellas que no aportaban nada. Un resultado similar puede encontrarse en Gabriel [2002].

Considerando que cada variable se ha ajustado con una regresión logística ordinal con el modelo de odds proporcionales se pueden definir para cada variable diversos tests e índices de bondad de ajuste. En este estudio utilizamos el test de razón de verosimilitud para comparar el modelo con probabilidad constante(sin dimensiones latentes) con el modelo completo de la misma forma que en la regresión logística estándar. El test se debe interpretar con cautela en esta ocasión, puesto que las variables latentes se estiman dentro del procedimiento y su variación no se considera explicitamente en el test; no obstante es una indicación de la significación de las variables para describir los datos. Para los biplots clásicos lineales no se llevan a cabo este tipo de tests, sino que se calculan indicadores de bondad de ajuste. Gardner-Lubbe y col. [2008] definen lo que ellos llaman *predictivities* o predictividades como el porcentaje de la varianza de cada variable explicada por las dimensiones, es decir, es una medida de la precisión de la predicción en el biplot. La predictividad se utiliza como una medida de bondad de ajuste por el paquete de R BiplotGUI(la Grange y col. [2009]). En el contexto del CA dichas cantidades se llamaban contribuciones relativas de los ejes a los elementos(variables o individuos) (ver Benzécri [1973], citeGreenacre1984), extendidas a los biplots por Galindo [1986] o Greenacre [1984]. El paquete de software MULTBIPLOT (Vicente-Villardón [2010]) utiliza y calcula las contribuciones de esta forma.

Desde otro punto de vista, las predictividades son los coeficientes de determinación R_j^2 para las regresiones en 3.9. Para respuestas dadas por variables ordinales

se puede utilizar un indicador pseudo R_j^2 como el de Cox-Snell o Nagelkerke.

6.5. Un estudio empírico.

El estudio real que nos servirá para ilustrar el funcionamiento del biplot logístico ordinal es el que describíamos en la sección 5.2.2.3 y que utilizaba la encuesta sobre las carreras profesionales de los doctorados en España.

La riqueza de datos que aporta la encuesta que estamos analizando llevaría en sí misma a disponer de material suficiente para presentar un apartado exclusivamente descriptivo, que no es el objetivo de esta sección. Por ello se presentan algunas pinceladas descriptivas y algunos resultados que son especialmente destacables.

Al margen de las descripciones más usuales tales como que el 54 % de los doctores son varones y el 46 % mujeres; que más de la mitad de los doctores tienen entre 35 y 45 años de edad, que el 45 % trabajaba en centros de enseñanza superior y un 36 % en la Administración Pública, que 1 de cada 4 doctores pertenecen al campo "ciencias naturales" (matemáticas, física, química, biología, informática, medio ambiente, etc...), que el 96.4 % de ellos estaba en activo a 31 de diciembre de 2006, o que la distribución de los doctorados en áreas de conocimiento está repartida en Ciencias Naturales(29,2 %), Ciencias Médicas(22,6 %), Ciencias Sociales(20,8 %), Humanidades(14 %), Ingeniería y Tecnologia(9,6 %) y Ciencias de Agricultura(4 %), las cuales además pueden consultarse en la nota de prensa que publica el INE en su página web, se exponen otros resultados significativos:

1.- En todos los campos hay "paridad" en el número de doctores según el sexo, salvo en "ingeniería y tecnología" donde hay más de 2 varones por cada mujer. Esto queda perfectamente explicado por el hecho de que la mujer se incorporó tarde a este campo de estudio, que hasta hace poco fue coto casi exclusivo de varones.

2.- La financiación principal en "ciencias naturales" e "ingeniería y tecnología" proviene de becas tanto públicas como privadas, mientras que en "ciencias médicas" y "humanidades" dos tercios no se han financiado con becas.

3.- En el último decenio del siglo XX la relación entre doctores según sexo (V/M) fue de 22/17, pero en el siglo XXI esta relación ha ido cambiando hasta llegar a 21/21.

4.- De los doctores ocupados que investigaban a finales de 2006, el 8 % lo hacía en la "empresa privada" y el 60 % en la "enseñanza".

5.- El salario es siempre una variable influyente; un dato que explica en sí mismo la valoración objetiva que se hace de los distintos campos de estudio e investigación. Podemos destacar que los doctores del campo "ciencias médicas" del sector "administraciones" tienen el salario medio más alto, mientras que los doctores de "humanidades" en el sector "empresa" tienen el más bajo (tabla 6.1). El sector "enseñanza" es el que tiene el salario medio más bajo. La mayor diferencia de salario entre sexos se da en "ciencias sociales" en el sector "empresa"; las menores en "humanidades" en los sectores "enseñanza" e "Instituciones Privadas Sin Fines de Lucro (IPSFL)".

Cuadro 6.1: Media de salarios por disciplina científica y sector de actividad.

Disciplina Científica	**Media por Disciplina**	**Salario medio por sector**			
		Empresas	**Sector Público**	**Enseñanza Superior**	**IPSFL**
Ciencias Naturales	31287	31808	30592	31763	29025
Ing. y tecnología	35137	40534	32824	34549	32869
Ciencias Médicas	41187	38981	43545	32448	40320
Ciencias Agrarias	30991	28708	31763	31287	29490
Ciencias Sociales	33521	35298	35157	32669	33573
Humanidades	29208	21979	29997	30282	25792
Total	34098	34456	36495	32188	32495

6.- Los doctores en "humanidades" tardan el triple de tiempo en encontrar un trabajo relacionado con su preparación que los de "ingeniería y tecnología".

7.- Dos de cada tres doctores trabajan en un puesto para el que ser doctor no es un requisito indispensable.

Particularmente destaca la información reflejada en las variables "edad al graduarse doctor" y "tiempo de duración del doctorado". Podemos ver en la tabla siguiente (tabla 6.2) que la edad media al graduarse en las disciplinas de "letras puras" es superior a las de "ciencias" debido a que la duración del doctorado es mayor en las primeras. Los graduados en las disciplinas de "ciencias sociales" y "humanidades" tardan, de media, un año más en acabar el doctorado que el resto, sin embargo se gradúan con algo más de tres años de edad biológica. Esta diferencia es similar en ambos sexos y parece que sólo puede explicarse suponiendo que los doctores en "ciencias sociales" y "humanidades" empiezan con más edad el doctorado. Por ello analicemos ambas variables desde el punto de vista de la "edad" del doctor y de la "edad x disciplina".

El término PhD, que aparecerá en algunas tablas, se refiere al proceso que han seguido las personas que están en posesión del título de doctor, desde los cursos de doctorado hasta la lectura de la tésis doctoral. Hablaremos de su duración, de cuándo se iniciaron estos estudios, de la edad con la que se concluyeron, etc..., aunque se usa también indistintamente para hacer referencia a las personas que son doctores en alguna disciplina.

Podemos comparar las variables antes mencionadas teniendo en cuenta la "generación de doctores", es decir, teniendo en cuenta la media por grupos de edad al doctorarse independientemente de la edad biológica actual (tabla 6.3). Podemos contestar a preguntas del tipo: "los doctores que tienen entre 55 y 64 años cumplidos que se doctoraron cuando tenían menos de 35 años, ¿a qué edad media lo hicieron?". Los datos recogidos en la encuesta, como se señaló anteriormente, se centran en las personas doctoradas con posterioridad a 1990, y este dato hay que tenerlo presente a la hora de sacar conclusiones descriptivas de la tabla. Este "inconveniente" restringe las posibles comparaciones a realizar.

Las últimas generaciones de jóvenes licenciados se doctoraron con 29.17 años

Cuadro 6.2: Edades medias y medianas de los doctorados por disciplina científica.

Disciplina Científica	**Edad al doctorarse**	**Edad al doctorarse**	**Duración PhD**	**Duración PhD**
	Media (años)	**Mediana** (años)	**Media** (meses)	**Mediana** (meses)
Ciencias Naturales	31	30	65	60
Ing. y tecnología	34	32	69	60
Ciencias Médicas	35	34	72	60
Ciencias Agrarias	33	31	65	57
Ciencias Sociales	36	33	77	68
Humanidades	37	34	84	72
Total	34	32	72	60

de media (tabla 6.3), mientras que los doctores que tienen en la actualidad entre 45 y 54 años y que se doctoraron también 'antes de los 35 años', lo hicieron con 32.12 años de media.

Esta tendencia se mantiene también (tabla 6.3) si nos fijamos en la edad biológica a la que comenzaron los cursos de doctorado.

Siguiendo con la comparación, si nos fijamos en las edades más avanzadas (tabla 6.3), podemos detectar que la tendencia se mantiene a pesar de la diferencia de edad. Las generaciones que hoy tienen una edad entre 45-54 años que se doctoraron cuando tenían entre 45 y 54 años, lo hicieron con casi 4 años menos de edad que los doctores más mayores que también se doctoraron entre 45 y 54 años de edad.

Una vez analizadas algunas de las características de la muestra lo que haremos en este ejemplo es focalizar nuestra atención en los aspectos relacionados con la satisfacción laboral. Este concepto ha sido abordado por diversos autores, como Locke [1976] que lo entiende como un estado emocional positivo que resulta de la

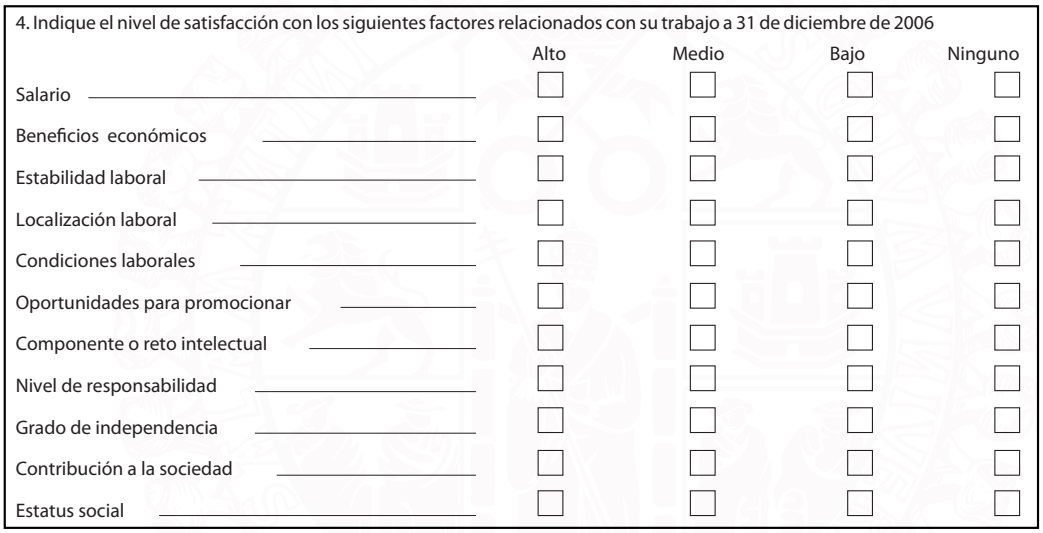
Cuadro 6.3: Edades actuales, duraciones y edades al comienzo de los PhD para las distintas generaciones.

Edad actual del Doctor	**Generaciones de PhD**	**Edad al doctorarse**		**Duración del PhD**		**Edad de inicio del PhD**	
		Media (años)	**Mediana** (años)	**Media** (meses)	**Mediana** (meses)	**Media** (añoss)	**Mediana** (meses)
Edad <35	PhD <35	29,17	29	57,81	55	24,35	24
35 <Edad <44	PhD <35	30.29	30	61.68	60	25.15	25
	35 <PhD <44	37.39	37	95.17	96	29.46	29
45 <Edad <54	PhD <35	32.12	32	63.53	60	26.83	26.92
	35 <PhD <44	39.53	39	85.91	72	32.38	32.83
	45 <PhD <54	47.82	47	95.94	84	39.83	40.5
55 <Edad <64	35 <PhD <44	42.34	43	88.45	72	34.97	36
	45 <PhD <54	49.32	49	96.28	73	41.30	42.50
	55 <PhD <64	57.61	57	114.58	108	48.06	49
65 <Edad <70	45 <PhD <54	51.71	52	107.15	72	42.78	45
	55 <PhD <64	59.10	59	111.56	72	49.81	52.25
	65 <PhD <70	67.21	68	60.46	60	62.17	62

valoración que cada persona tiene de su experiencia profesional, o como Spector [1976], para el que es el sentimiento de las personas hacia su trabajo, que contiene tanto aspectos satisfactorios como insatisfactorios del mismo. La problemática real, independientemente de las posibles definiciones es su medición y la identificación de los condicionantes que la determinan. Existen trabajos en los que se asigna en una escala un valor a la satisfacción laboral, mientras que en otros, como los de Fabra y Camisón [2009] la satisfacción laboral se calcula como una media aritmética de la valoración del entrevistado a siete aspectos del puesto de trabajo. En nuestro caso, en la encuesta citada sobre la que estamos trabajando, existe en el módulo C una

pregunta, la 6.4, que trata de conocer el nivel de satisfacción de los doctorados en relación a algunos aspectos relacionados con su trabajo. En este caso el objetivo es analizar el comportamiento de cada ítem dado su carácter ordinal, y no establecer una medida numérica y única relativa a la satisfacción laboral.

En esta ocasión no nos restringiremos a un conjunto pequeño de doctorados, sino que trabajaremos con la matriz completa de todos ellos. Esta pregunta tiene 11 apartados o ítems en escala likert de 1 a 4(ver figura 6.6) que serán considerados como variables ordinales.


4. Indique el nivel de satisfacción con los siguientes factores relacionados con su trabajo a 31 de diciembre de 2006

	Alto	Medio	Bajo	Ninguno
Salario	☐	☐	☐	☐
Beneficios económicos	☐	☐	☐	☐
Estabilidad laboral	☐	☐	☐	☐
Localización laboral	☐	☐	☐	☐
Condiciones laborales	☐	☐	☐	☐
Oportunidades para promocionar	☐	☐	☐	☐
Componente o reto intelectual	☐	☐	☐	☐
Nivel de responsabilidad	☐	☐	☐	☐
Grado de independencia	☐	☐	☐	☐
Contribución a la sociedad	☐	☐	☐	☐
Estatus social	☐	☐	☐	☐

Figura 6.6: Pregunta C.6.4 del cuestionario de doctores

La distribución de las respuestas en cada uno de los apartados según las 4 categorías puede verse en la tabla 6.4. El salario se configura como el tercer aspecto con menor alta satisfacción, es decir, entre los condicionantes sobre los que los doctorados se muestran más satisfechos el salario aparece en una de las últimas posiciones. Además, las oportunidades para promocionar y progresar en la carrera laboral son cuestiones sobre las que los doctorados se muestran muy pesimistas y poco satisfechos. Por otra parte, los doctorados parecen estar mucho más satisfechos con la contribución a la sociedad del trabajo que desarrollan, al igual que con

la estabilidad y seguridad laboral, toda vez que la mayoría de ellos están en el sector público. También la localización, así como el nivel de responsabilidad y el reto que supone su actividad tienen valores elevados de satisfacción, lo cual confiere al escenario, en general, de una percepción de valoración positiva, podríamos decir.

Cuadro 6.4: Distribución porcentual de los ítems de la satisfacción según sus categorías.

	Ninguna	**Baja**	**Media**	**Alta**
Salario	2.4	30.1	53	14.5
Beneficios	15.8	34.8	40	9.5
Seguridad Laboral	7.4	15.7	19.5	57.4
Ubicación laboral	1.6	9.5	28.9	60.1
Condiciones de trabajo	1.8	15	46.4	36.8
Oportunidades de progresar	12.6	36.1	37.4	13.8
Reto intelectual	3	12.1	30.1	54.9
Nivel de responsabilidad	0.8	7.2	38.1	53.9
Grado de independencia	1.8	10.4	38.7	49
Contribución a la sociedad	0.8	5.9	36	57.3
Estatus social	2.2	13.3	63.2	21.3

Canal Domínguez [2013], sobre datos de la misma encuesta, trata de agrupar los aspectos de la satisfacción inicialmente mediante un análisis factorial, no encontrando ninguna estructura de factores subyacentes, cosa, por otra parte lógica, puesto que esta técnica no es adecuada para este tipo de datos en escala tipo likert de variables ordinales, sino para datos en escala numérica de intervalos. Analizando, por otra parte, los coeficientes de correlación de Pearson entre estas variables puede verse que estas se agrupan en dos conjuntos, que estarían formados, uno por aspectos relacionados con el puesto de trabajo (salario, estabilidad laboral, localización, condiciones laborales, oportunidades para promocionar, nivel de responsabilidad y grado de independencia) y el segundo por cuestiones que van más

allá del propio puesto (reto intelectual, contribución a la sociedad y estatus social), lo cual va a estar en concordancia con los cálculos que vamos a realizar con la metodología del biplot ordinal.

En la tabla 6.5 figuran los indicadores resultantes de la estimación de los parámetros del modelo con el algoritmo alternado en un espacio bidimensional, y las cargas en cada uno de los ejes y comunalidades pueden verse en la tabla 6.6. Los Porcentajes de Clasificaciones Correctas (PCC) son muy altos para las variables "Salario" y "Reto intelectual", presentando pseudo-R^2 de Nagelkerke cercanos a uno. Esto nos da un pista de que el problema de la separación, incluso en un caso con una matriz enorme, puede afectar a la solución, por tanto es esencial que el algoritmo sea eficiente en este sentido.

En virtud de dichas cargas, el primer factor tiene pesos altos en las variables "Oportunidades de progresar", "Grado de independencia", "Reto intelectual" y "Nivel de responsabilidad" y "Contribución a la sociedad", que son características de la actividad investigadora y en parte del puesto de trabajo en sí mismo. El segundo factor presenta valores elevados en las variables "Salario", "Beneficios", "Seguridad laboral" y "Condiciones laborales", todas ellas relacionadas con aspectos económicos y de estabilidad en el puesto de trabajo. Las variables "Ubicación laboral" y "Estatus social" tienen puntuaciones similares en ambos factores y participan en los dos.

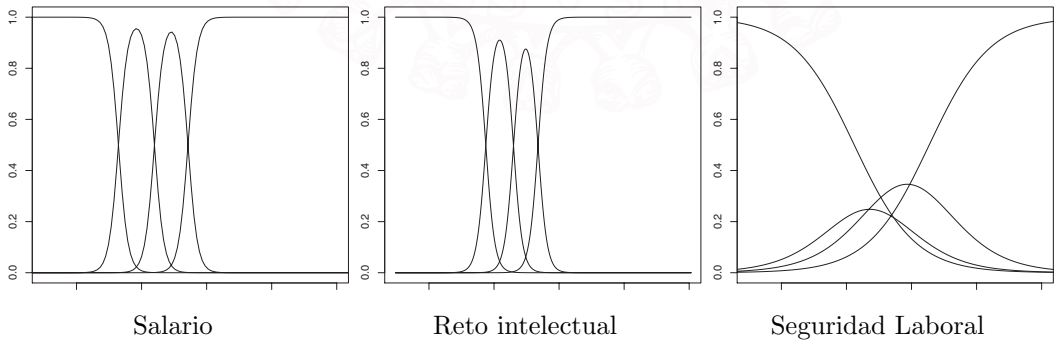


Figura 6.7: Curvas de respuesta de los items para cada una de las variables.

Cuadro 6.5: Indicadores de bondad de ajuste para las 11 variables.

Variable	logLik	Deviance	df	p-valor	PCC	Nagelkerke
Salario	-2829.422	5658.845	2	0	0.937	0.957
Beneficios	-8158.998	16317.996	2	0	0.799	0.826
Seguridad Laboral	-12689.867	25379.735	2	0	0.582	0.151
Ubicación laboral	-11037.486	22074.972	2	0	0.603	0.098
Condiciones de trabajo	-10072.833	20145.666	2	0	0.630	0.465
Oportunidades de progresar	-11963.385	23926.771	2	0	0.561	0.543
Reto intelectual	-1845.119	3690.237	2	0	0.962	0.971
Nivel de responsabilidad	-9857.231	19714.462	2	0	0.592	0.243
Grado de independencia	-10049.245	20098.490	2	0	0.609	0.386
Contribución a la sociedad	-9407.924	18815.847	2	0	0.623	0.247
Estatus social	-8991.158	17982.315	2	0	0.708	0.465

Cuadro 6.6: Cargas factoriales y comunalidades.

Variable	F1	F2	Comunalidades
Salario	0.105	0.991	0.994
Beneficios	0.109	0.986	0.984
Seguridad Laboral	0.287	0.858	0.819
Ubicación laboral	0.684	0.442	0.664
Condiciones de trabajo	0.613	0.749	0.938
Oportunidades de progresar	0.876	0.403	0.930
Reto intelectual	0.988	-0.137	0.995
Nivel de responsabilidad	0.902	0.173	0.843
Grado de independencia	0.911	0.275	0.906
Contribución a la sociedad	0.922	0.018	0.851
Estatus social	0.732	0.626	0.929

Las funciones de información de los ítems para tres de las variables pueden observarse en la figura 6.7.

En la variable “Seguridad laboral” la segunda categoría está oculta, que corresponde con parcialmente satisfecho, o satisfacción media, concentrando por tanto toda la información en las otras tres opciones de tal forma que podría entenderse que la satisfacción es o muy alta o muy baja, lo cual parece tener sentido con el hecho de que la organización de la administración pública española concentra el empleo de la mayoría de los doctorados.

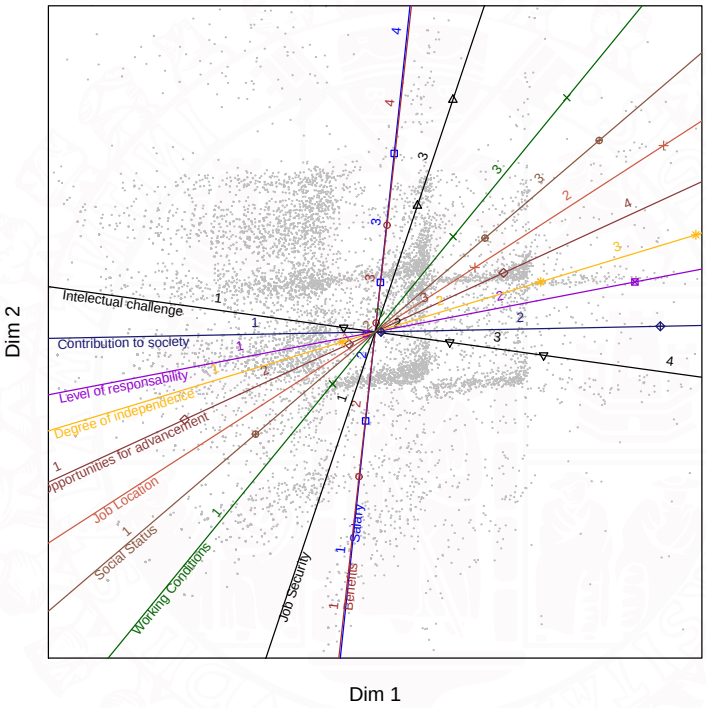


Figura 6.8: Biplot Logístico Ordinal. Satisfacción de los doctorados con su principal empleo en España.

El biplot logístico ordinal puede verse en la figura 6.8, en el que se muestran sobre las dimensiones del biplot las posiciones de cada variable así como los puntos característicos del corte de las curvas para aquellas categorías que no son ocultas. En este gráfico, el ángulo entre el eje de abcisas y algunas variables como el “Salario”, los “Beneficios” y la “Seguridad laboral” es casi de 90°, de modo que

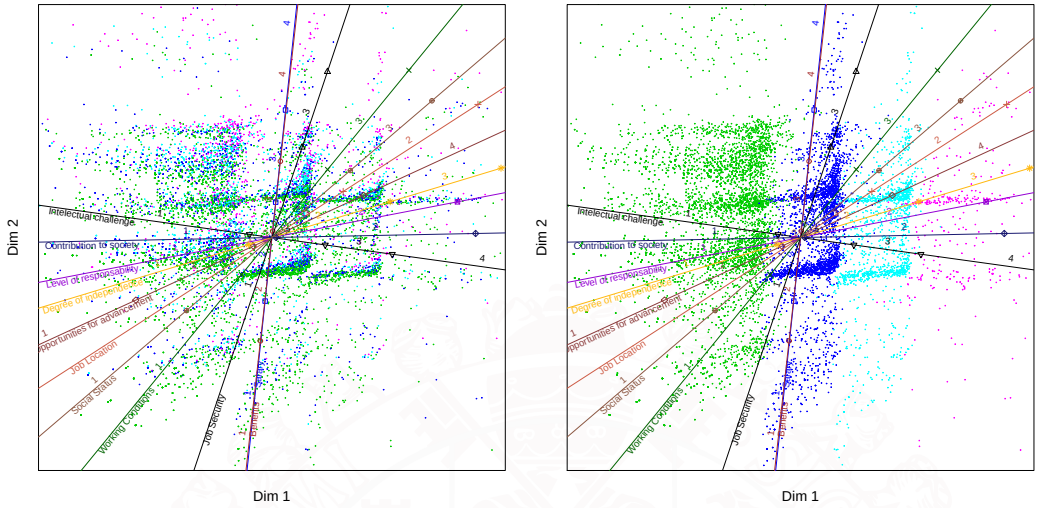
aparece una región en el primer cuadrante alejada del origen que se corresponde con doctorados que están muy poco satisfechos con dichas variables y también con otras relacionadas con la actividad investigadora. La representación muestra que la satisfacción con los ingresos no parece estar correlacionada con la que los doctorados perciben sobre aspectos intelectuales, como el reto intelectual o el grado de responsabilidad, aspecto que aparece también en otros países europeos, como se hace patente en el caso austriaco [Schwabe, 2011].

Al igual que ocurría con la matriz reducida de datos, ahora aparecen variables con un comportamiento similar en el plano estudiado, como el "Nivel de responsabilidad" o la "Contribución a la sociedad", en las que los puntos que delimitan cada categoría en el eje biplot se situan en posiciones parecidas para ambas y las pendientes son también similares.

Si coloreamos cada individuo de acuerdo con la respuesta a las variables citadas en las curvas de información del ítem, la situación de cuasi-separación es muy patente en la variable "Reto intelectual" y se ve que no existe en la "Seguridad laboral" con un comportamiento individual mucho más disperso y mezclado (ver figura 6.9). Si dibujamos el salario, también apreciamos este problema, con franjas horizontales correspondientes a cada categoría (figura 6.10).

Estas dos variables (Salario y Reto intelectual) parecen ser importantes en la interpretación de la información y comprensión de ciertos aspectos de la nube de puntos de los doctorados.

En investigaciones recientes, como las de Canal Domínguez [2013] sobre este colectivo tan específico y particular, se ha podido comprender, analizando una regresión por mínimos cuadrados ordinarios de la media aritmética de las 11 respuestas a la pregunta de satisfacción frente a una batería amplia de variables del cuestionario, que existen condicionantes que parecen estar muy ligados a la satisfacción laboral, como el sector en el que trabaja el doctorado, apareciendo los que se sitúan en la empresa privada como más satisfechos que los del sector público. Otro es la edad, que parece ejercer un efecto positivo aunque no significativo y



Seguridad laboral Reto intelectual

Figura 6.9: Coloración según la categoría respondida por los doctorados en el biplot logístico ordinal.

refleja que las mujeres están más satisfechas globalmente, o que cuanto mayor es la relación entre formación y empleo mayor es la satisfacción del trabajador. No obstante, los puestos que aglutinan a personal más cualificado de lo que dichos puestos requieren tiene un efecto negativo sobre la satisfacción, lo cual está en concordancia con la literatura tradicional.

Además parece haber un aspecto fundamental cuyo efecto es determinante dados los coeficientes que presenta, que son los ingresos del doctorado, de manera que a medida que aumenten estos mayor es el nivel de satisfacción, afectando estos positivamente a dicha variable, como vamos a comprobar al final de este apartado. El intervalo de salarios entre 20 y 30 mil euros aglutina al 24,6 % de la muestra y más de la mitad de los doctorados cobran menos de 35.000 euros. El intervalo de ganancias de más de 50.000 euros congrega a porcentajes de doctorados muy similares a los que estan en el tramo de 10 a 20 mil euros(12 %).

Se han estudiado en el gráfico conjunto algunas variables ya conocidas, por

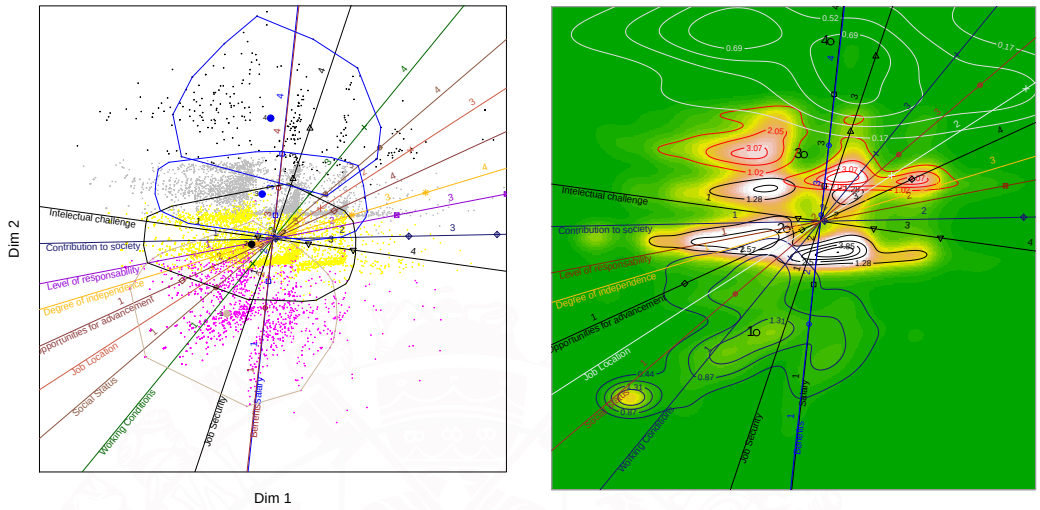
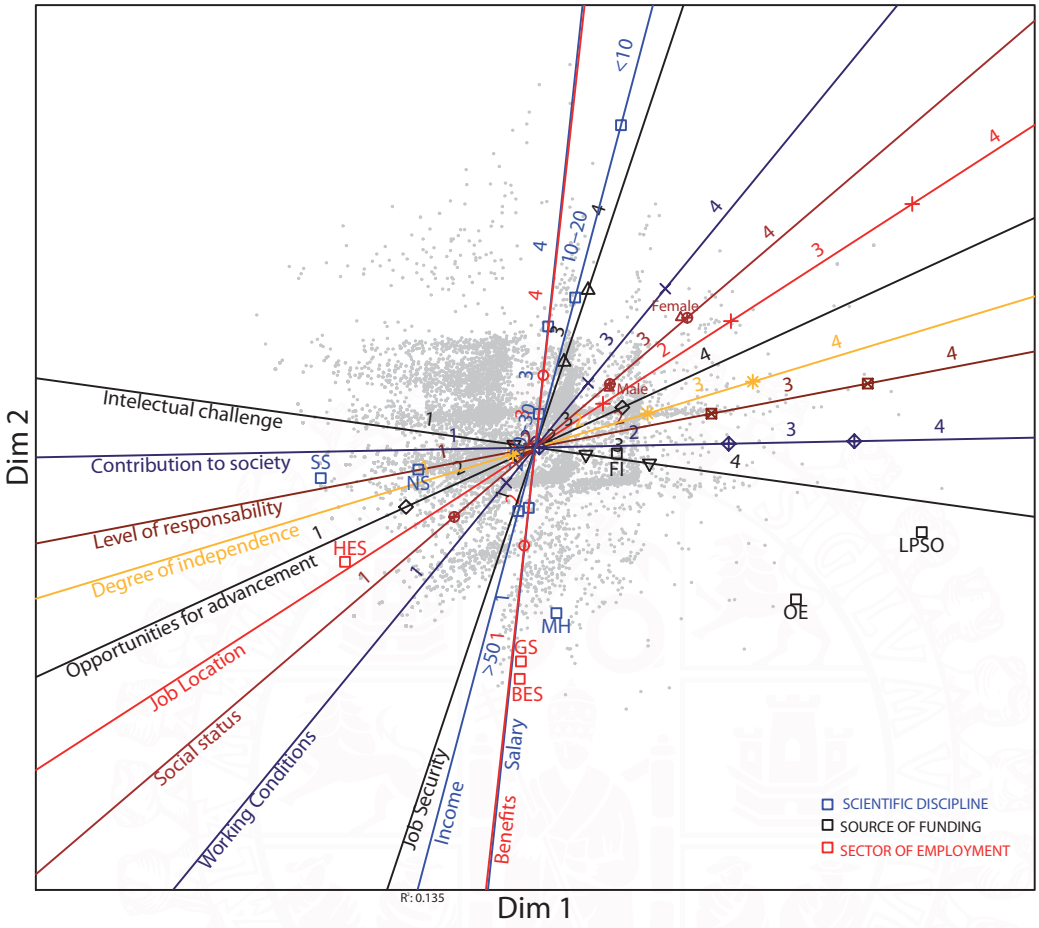


Figura 6.10: Envolturas convexas y gráfico de densidad con las líneas de contorno para la variable Salario

ejemplo el sector en el que trabajan los individuos o la fuente de financiación, así como el sexo del doctorado con el objetivo de situar las posiciones de las categorías predichas en relación con la satisfacción en cada una de sus vertientes. Según el sector de empleo, aparece una diferenciación clara entre el sector educación superior y el resto de sectores. Para el primero de ellos parece existir un acentuado reto intelectual, con un alto grado de independencia y evidentes oportunidades de mejora laboral. En términos del salario y los beneficios, el sector empresas(BES) muestra un componente más atractivo para los doctorados(figura 6.11). Parece además, que el hecho de que estas expectativas salariales penalicen la carrera profesional universitaria hace que el destino profesional de los doctores se orienta cada vez más hacia las empresas privadas y hacia otros ámbitos de la administración pública, como muestran autores como Canal y Muñiz [2012].

La posición de los hombres y mujeres, calculando los centroides sobre las coordenadas del biplot, es practicamente indistinguible (figura 6.12(c)) dada la gran



Scientific Discipline(PCC: 30%,Nagelkerke:0.033): SS(Ciencias Sociales), NS(Ciencias Naturales), MH(Ciencias Médicas y de la Salud)
Source of Funding(PCC:41%,Nagelkerke:0.02): FI(Beca en España),OE(Otros empleos # enseñanza),LPSO(Préstamos y ahorros personales)
Sector of Employment(PCC:54%,Nagelkerke:0.11):HES(Enseñanza Superior), BES(Sector Empresas), GS(Administración Pública)
Sex(PCC:55%, Nagelkerke:0.02)

Figura 6.11: OLB con la variable ordinal Ingresos(Salario Bruto Anual) ajustada sobre los resultados del biplot, y con las variables nominales Sexo, Disciplina Científica, Fuente de financiación y Sector de Empleo superpuestas. Pueden consultarse las categorías de las mismas en la tabla 5.3.

muestra sobre la que trabajamos y que hay casi paridad, pero si tratamos la variable como nominal y ajustamos la variable nominal con las dimensiones del biplot

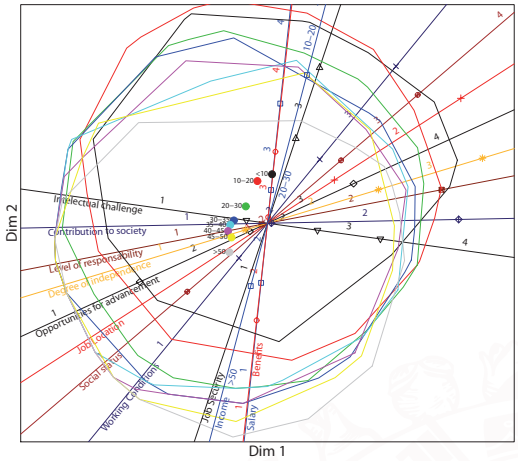
obtenemos posiciones distintas para ambos sexos, como se ve en el mismo gráfico, lo cual puede relacionarse con variables como los ingresos y el salario, confirmando la conocida brecha salarial entre unos y otros y que concuerda con estudios sobre este aspecto, como los de Canal y Rodríguez [2012].

En las figuras 6.12(a,b) se han calculado los centroides de los diversos tramos salariales para los doctores de acuerdo a sus coordenadas en el biplot bidimensional en las dos primeras, y en la figura 6.11 está superpuesta al biplot la variable ingresos del cuestionario, tratada como variable ordinal (Income), y ajustada sobre los ejes del biplot, y con las mismas categorías que presenta en el cuestionario[1], pudiendo apreciarse su relación con el eje que resume lo relativo al puesto de trabajo en si mismo, y poniendo de manifiesto que no todas las categorías se predicen, sino sólo algunas de ellas.

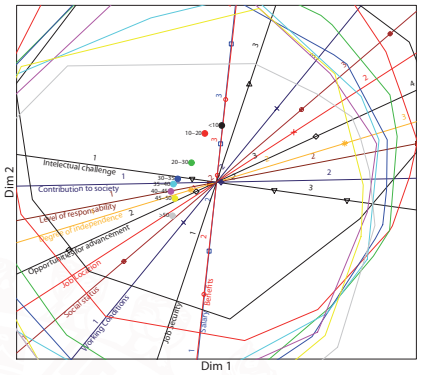
Si en lugar de haber considerado la variable ingresos como ordinal la hubiéramos catalogado como nominal se podría haber ajustado, mediante la función disponible en el software, sobre las dimensiones de la representación, obteniendo una configuración realmente similar a la ya conocida para el caso ordinal y con las mismas categorías predichas (ver figura 6.13).

Ahora nos podríamos plantear de qué forma influye el salario en la concepción de los elementos de satisfacción del cuestionario, para lo cual consideremos la figura 6.14, que muestra el biplot logístico ordinal calculado sobre el conjunto de doctorados que se sitúan en el intervalo de ganancias más bajo (por debajo de 10000 euros). Sobre él se han posicionado algunas variables nominales, como son la edad, el sector de empleo y la disciplina científica asociadas a este colectivo. Es significativo, como cabría esperar, que en este contexto se muevan los doctorados más jóvenes, puesto que solo se predicen los dos tramos de menor edad (tramo 1: <34 años; tramo 2; entre 35 y 45 años;tramo 3; entre 45 y 55 años;tramo 4; entre

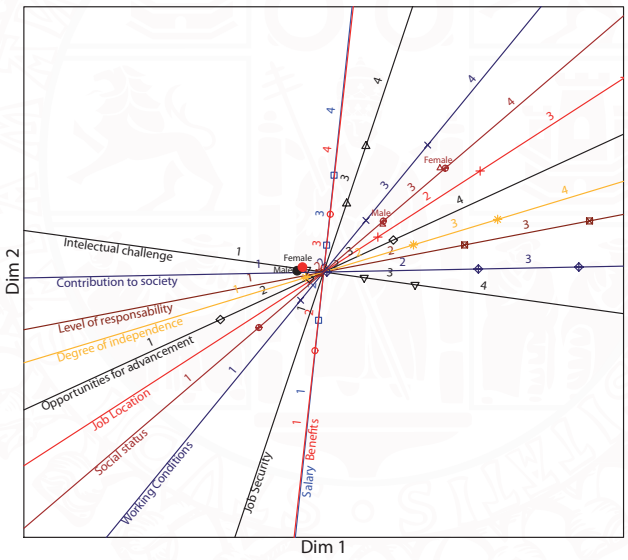
[1] Intervalo 1=Menos de 10000 euros; Intervalo 2=Desde 10000 hasta 20000 euros; Intervalo 3=Desde 20000 hasta 30000 euros; Intervalo 4=Desde 30000 hasta 35000 euros; Intervalo 5=Desde 35000 hasta 40000 euros; Intervalo 6=Desde 40000 hasta 45000 euros; Intervalo 7=Desde 45000 hasta 50000 euros; Intervalo 8=Más de 50000 euros;



(a) Centros de las categorías de la variable ingresos y ajuste ordinal de la variable. Sólo 4 categorías se predicen en este plano.



(b) Zoom de los centros de la imagen anterior



(c) Centros y ajuste de la variable nominal Sexo sobre el biplot logístico ordinal

Figura 6.12: Biplots Logísticos Ordinales con variables con información externa situadas en los gráficos.

55 y 65 años;tramo 5; entre 65 y 70 años), los cuales en su mayoría están en el sector enseñanza superior(HES) o empresarial(BES), y cuya orientación científica

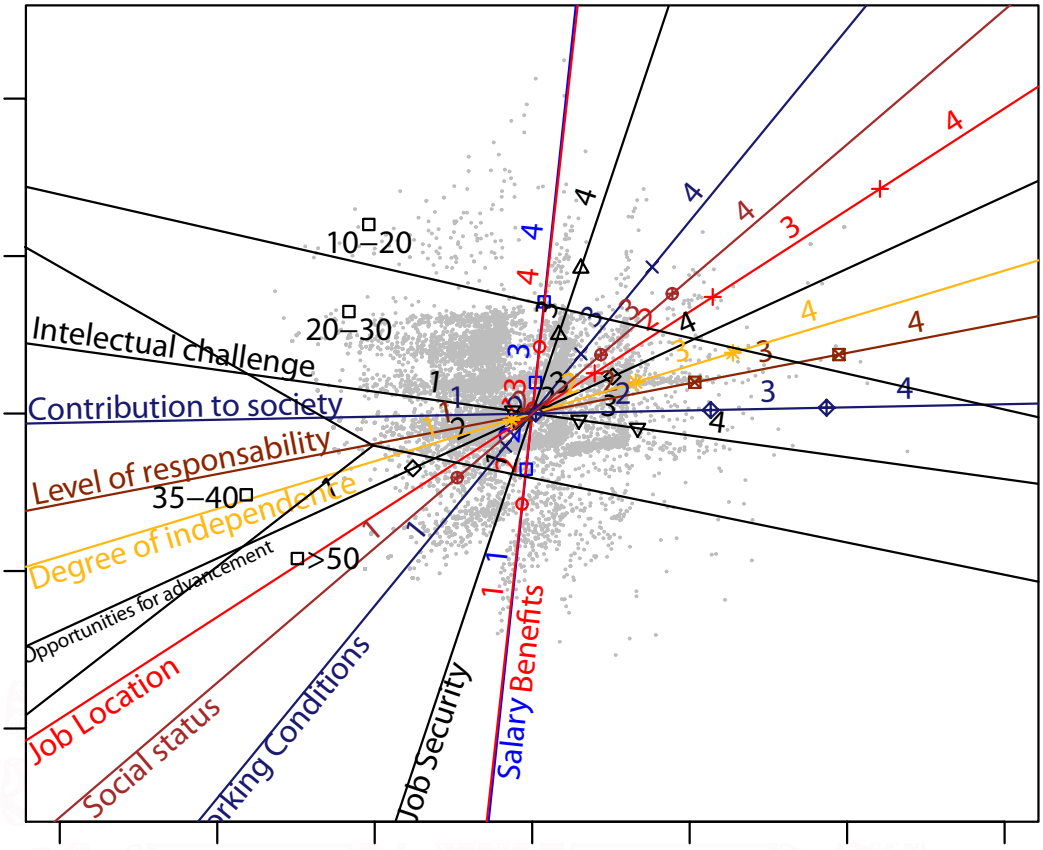


Figura 6.13: OLB con la variable Ingresos(Salario Bruto Anual) considerada como variable nominal y ajustada sobre los resultados del biplot.

viene dada por las ciencias de la naturaleza(NS), humanidades(H) y ciencias sociales(SS). Existen determinantes de la satisfacción que parecen estar en este tramo salarial y humano bastante unidos, como muestra el eje 2 con todo aquello relativo al salario, a la seguridad laboral, condiciones de trabajo y oportunidades para promocionar, aglutinando el sector empresarial mejores expectativas y satisfacción. Por otra parte aparecen aspectos como el reto intelectual, nivel de responsabilidad o contribución a la sociedad en los que ocurre, al contrario que en el caso anterior,

que el sector enseñanza superior es capaz de proporcionar grados de satisfacción más altos en los doctorados.

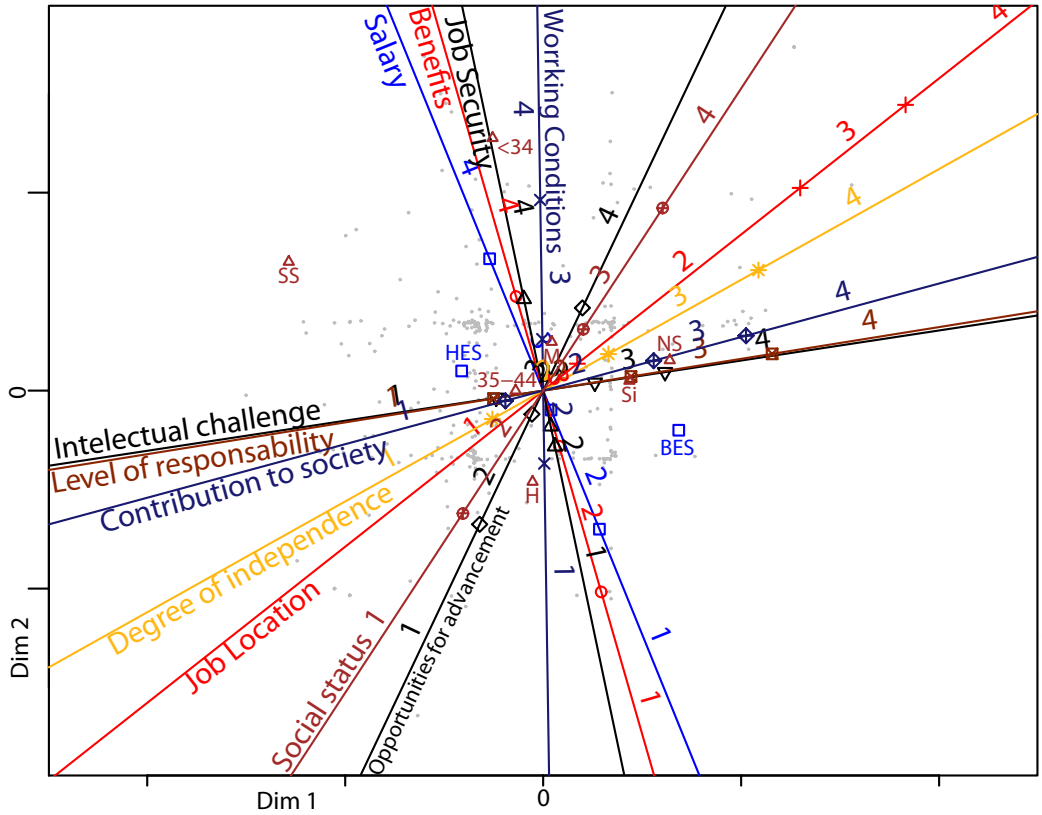


Figura 6.14: OLB de los doctorados con menores ingresos, con las variables nominales sector, edad y disciplina científicas ajustadas sobre el biplot.

Se podría ir analizando cómo varian las percepciones del colectivo a medida que aumenta el nivel de ingresos, puesto que estas cambian, como muestra la figura 6.15, en la cual el biplot logístico se ha ajustado al tramo de ingresos más elevado, en el cual los doctorados perciben más de 50000 euros. Sería interesante poder analizar a una generación de doctorados para ver cómo van variando sus apreciaciones y perspectivas, pero con estos datos no es posible puesto que la información

de este colectivo es transversal, por lo que la composición de la muestra es muy diversa en los tramos salariales y las comparaciones hay que realizarlas con cierta cautela. En este gráfico se han ajustado variables nominales sobre el biplot, como el sexo, la financiación o la disciplina científica, resultando que sólo se pueden predecir características para los hombres y no las mujeres, así como que los principales mecanismos de financiación de los estudios de doctorado son o bien una beca (de la institución donde realizó el doctorado, administración pública, empresa o institución sin fines de lucro; FI), o bien una ocupación a tiempo parcial o completo(OE). Por otra parte, para este colectivo parecen poder predecirse disciplinas científicas asociadas a las ciencias sociales(SS) y a las ciencias médicas(MH), si bien es verdad que los ajustes no son satisfactorios, indicando que es necesario un análisis más pormenorizado y teniendo en cuenta alguna dimensión adicional que sea capaz de captar variabilidad añadida al complejo colectivo que estamos tratando.

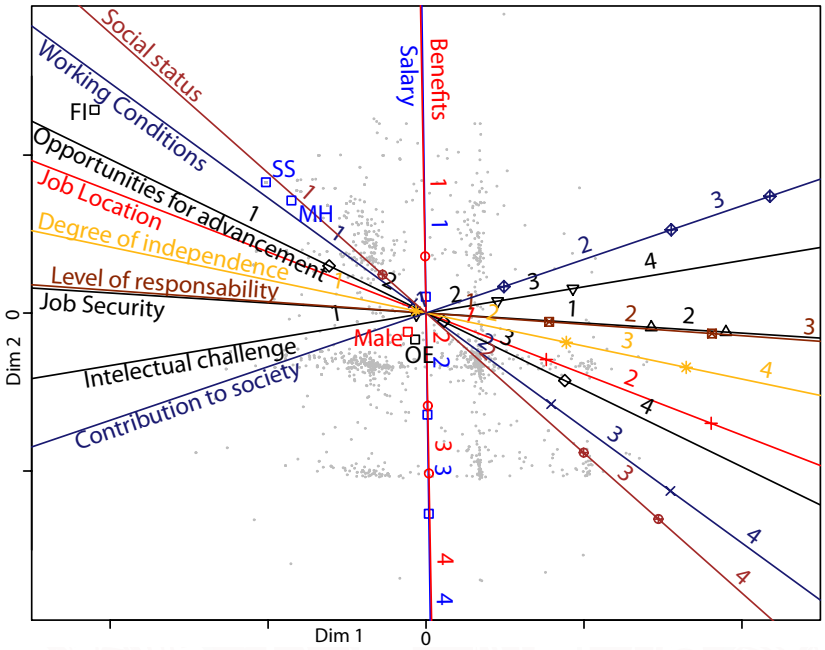


Figura 6.15: OLB de los doctorados con mayores ingresos, con las variables nominales sexo, financiación y sector ajustadas sobre el biplot.

Capítulo 7

El Algoritmo EM y el procedimiento de estimación de los parámetros

We all have dreams, in order to make dreams come into reality, it takes an awful lot of determination, dedication, self-discipline and effort.

– Jesse Owens

Olympic gold-medalist runner

En los modelos de la IRT construidos de forma general existe un problema consistente en que la habilidad de los individuos aparece en ellos como un parámetro molesto que no puede eliminarse de la verosimilitud aunque se condicione el modelo a un estadístico suficiente de igual forma que se ha

propuesto para el modelo logístico de Rasch con un parámetro[Andersen, 1980]. Incluso, en general, no es posible la estimación conjunta por máxima verosimilitud de los parámetros de individuos y variables puesto que el número de parámetros aumenta con los individuos y no es posible aplicar los métodos clásicos a estos casos.

Para poder estimar cuando aparece un parámetro aleatorio problemático como el que comentábamos anteriormente una buena aproximación es integrar sobre la distribución de dicho parámetro, estimando los parámetros estructurales por máxima verosimilitud sobre la distribución marginal(Máxima Verosimilitud Marginal (MML)). Para un modelo de dos parámetros con una distribución normal Bock y Lieberman [1970] estimaron los umbrales para las variables y las puntuaciones factoriales suponiendo que los individuos eran una muestra aleatoria extraída de una distribución $N(0,1)$ de la habilidad. Mediante la cuadratura de Gauss-Hermite pudieron llevar a cabo la integración necesaria y obtuvieron estimadores estables de estos parámetros para tan sólo cinco variables.

No obstante, el método de Bock y Lieberman no parece práctico para su uso general, aunque se puede aplicar a casi cualquier tipo de modelo de respuesta al ítem y fue utilizado por Bock [1972] para el modelo logístico de respuesta múltiple nominal. El motivo de ello es que computacionalmente el método es muy pesado debido a que para resolver las ecuaciones de MML utilizando Newton-Raphson, con n variables, es necesaria la generación e inversión de la matriz de información, que tiene dimensiones $2n$x$2n$, y en la que cada elemento de la misma es la suma de 2^n términos. Dado que es necesario construir e invertir esta matriz varias veces en las iteraciones del método de Newton-Raphson, el proceso deja de ser operativo. Incluso desde el punto de vista estadístico es discutible este método porque asume que la forma de la distribución de la habilidad que ha sido realmente muestreada se conoce de antemano. Puesto que los estudios de calibración de las variables se llevan a cabo sobre muestras seleccionadas arbitrariamente, es difícil especificar la distribución “a priori” de la habilidad en la población muestreada.

Bock y Aitkin [1981] reformularon las ecuaciones de verosimilitud de Bock-Lieberman, de forma que consiguieron una solución computacionalmente factible tanto para un número pequeño como alto de variables. Dempster y col. [1977] y otros autores posteriormente, como Hsu y col. [1999] muestran que el método obtenido está relacionado con el algoritmo EM para estimaciones de máxima verosimilitud. Esta formulación de las ecuaciones de verosimilitud pone de manifiesto que no es necesario que se conozca de antemano la forma de la distribución de la habilidad. En su lugar, se puede estimar mediante una distribución discreta sobre un número finito de puntos. Los parámetros correspondientes a las variables se estiman entonces integrando sobre la distribución empírica, lo cual libera al método de suposiciones arbitrarias sobre la distribución de la habilidad en la población muestreada.

A pesar de los avances de Bock y Aitkin [1981], trabajos posteriores, como los de Bock y col. [1988] y Muraki y Carlsson [1995] muestran que este método resulta apropiado para soluciones con un número de factores bajo o moderado con tal de que el número de cuadraturas por dimensión decrezca a la vez que el número de factores aumente. Cuando el número de dimensiones crece esta técnica se comporta de manera ineficiente, puesto que el número de puntos de las cuadraturas necesarios para estimar en la etapa “E” aumenta exponencialmente.

Schilling y Bock [2005] trabajaron en una solución para un número moderado de dimensiones utilizando una cuadratura adaptativa con el objetivo de conseguir una mejor precisión cuando se utilizaba un número pequeño de cuadraturas por dimensión, pero el problema de soluciones basadas en muchas dimensiones aún permanecía. Este problema ha sido estudiado recientemente mediante el empleo de métodos de estimación estocásticos para análisis exploratorios y confirmatorios de las variables. Edwards [2010] y Sheng [2010] han investigado sobre métodos bayesianos MCMC(Markov Chain Monte Carlo), y ambos autores han publicado software para estimar los parámetros para modelos de respuesta politómicos y dicotómicos, respectivamente.

Por otra parte, el algoritmo de Metropolis-Hastings Robbins-Monro para análisis factorial exploratorio de ítems o variables [Cai, 2010a] para calcular estimadores MML es adecuado cuando el espacio solución considerado tiene un número de dimensiones alto, presentando ventajas sobre los procedimientos existentes, como el algoritmo EM basado en cuadraturas numéricas.

Puesto que uno de los objetivos de este trabajo es desarrollar una metodología para el tratamiento de variables categóricas en un espacio reducido(normalmente de 2 ó 3 dimensiones a lo sumo) el algoritmo EM es una opción válida, aunque consideramos necesaria una adaptación del mismo para, como veremos, tener en cuenta la resolución del problema de la separación en regresión logística. Esta matización no parece estar contemplada en trabajos muy recientes sobre este campo, como los de Chalmers [2012].

7.1. El Algoritmo EM.

El algoritmo EM es una técnica iterativa que pretende realizar una estimación de máxima verosimilitud de parámetros en situaciones en las que existen *datos ocultos* y fue desarrollado por primera vez por Dempster y col. [1977]. Se utiliza en situaciones en las que se quiere estimar un conjunto de parámetros que describen una distribución de probabilidad subyacente, θ , disponiendo sólo de una parte observada de los datos completos producidos por la distribución.

El nombre EM (Expectation-Maximization) es debido a que el procedimiento consiste en definir una esperanza o expectativa particular que posteriormente se maximiza. Es una técnica iterativa, que se inicia con un valor inicial de los parámetros, los cuales se van actualizando en cada iteración de forma que maximizan la expectativa en esa iteración particular. Vamos a presentar a continuación una formalización del procedimiento.

Consideremos $\mathbf{X}_{I\times J}$ una matriz de datos que contiene los valores de J variables categóricas(pueden ser nominales u ordinales), cada una con K_j $(j = 1, \ldots, J)$

categorías, medidas sobre I individuos, y sea $\mathbf{G}_{I\times L}$ la matriz indicadora construida como

$$\mathbf{G}_{I\times L} = [\mathbf{G}_1, \mathbf{G}_2, \mathbf{G}_3, \ldots, \mathbf{G}_J]$$

con $L = K_1 + K_2 + \cdots + K_J$, y siendo cada $\mathbf{G}_i$ una matriz de dimensión $I \times K_i$. Todas las filas de $\mathbf{G}$ suman J, y las sumas por columnas dan las frecuencias de todos los niveles de todas las categorías de las variables. Podemos considerar esta matriz desde el punto de vista de los individuos y denotarla de la siguiente forma

$$\mathbf{G}_{I\times L} = (\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_I)$$

siendo cada componente un patrón de respuesta posible de la muestra de individuos, es decir, $\mathbf{g}_i = (g_{i1}, g_{i2}, \ldots, g_{iL})'$, denotando g_{ij} la puntuación del individuo i sobre la variable j-ésima, que es binaria dada la construcción de $\mathbf{G}$. Tenemos entonces que

$$g_{ij} = \begin{cases} 1 & \text{si el individuo } i \text{ posee la característica } j \\ 0 & \text{en otro caso} \end{cases} \tag{7.1}$$

Llamamos, al igual que en otras secciones $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_L)$ a la matriz de parámetros de las variables, donde los componentes del vector $\mathbf{b}_\ell$ para cada variable dependen del modelo de ajuste particular, y $\mathbf{A}$ a la de las habilidades de los individuos, siendo

$$\mathbf{A} = \begin{pmatrix} \mathbf{a}_1' \\ \vdots \\ \mathbf{a}_I' \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1S} \\ \vdots & \ddots & \ldots & \vdots \\ a_{I1} & a_{I2} & \ldots & a_{IS} \end{pmatrix}$$

para un espacio reducido de dimensión S.

Antes de proponer el procedimiento para el conjunto de individuos vamos a centrarnos en uno en concreto para familiarizarnos con algunas expresiones. Para un individuo genérico con una habilidad $\bar{\mathbf{a}}$, y suponiendo que g representa la puntuación sobre una de las variables binarias, que será 1 ó 0, se cumple que la función de probabilidad de g es

$$P(g = \mathcal{G}|\mathbf{a} = \bar{\mathbf{a}}, \mathbf{b}) = P(g = 1|\mathbf{a} = \bar{\mathbf{a}}, \mathbf{b})^{\mathcal{G}}(1 - P(g = 1|\mathbf{a} = \bar{\mathbf{a}}, \mathbf{b}))^{(1-\mathcal{G})}$$

siendo $\mathcal{G} = 0$ ó 1. Además, asumiendo la existencia de independencia local, considerando la respuesta a las L variables binarias del vector $\mathbf{g}_v = (g_1, \ldots, g_L)$ es inmediato que

$$P(\mathbf{g}_v = \overline{\mathbf{g}}|\mathbf{a} = \overline{\mathbf{a}}, \mathbf{b}) = \prod_{l=1}^{L} P(g_l = \mathcal{G}_l|\mathbf{a} = \overline{\mathbf{a}}, \mathbf{b}_l)$$

Para simplificar la notación posteriormente denotamos también

$$p_j(\overline{\mathbf{a}}_i) \equiv P(g_{ij} = 1|\mathbf{a}_i = \overline{\mathbf{a}}_i, \mathbf{b}_j)$$

La probabilidad condicionada de que una matriz $\overline{\mathbf{G}}$ de dimensión $I \times L$ se presente para un conjunto de I individuos con habilidades $\overline{\mathbf{a}}_i$ que responden de manera independiente es:

$$\begin{aligned} P(\mathbf{G} = \overline{\mathbf{G}}|\mathbf{A} = \overline{\mathbf{A}}, \mathbf{B}) &= \prod_{i=1}^{I} P(\mathbf{g}_i = \overline{\mathbf{g}}_i|\mathbf{a}_i = \overline{\mathbf{a}}_i, \mathbf{B}) \\ &= \prod_{i=1}^{I}\prod_{j=1}^{L} P(g_{ij} = 1|\mathbf{a}_i = \overline{\mathbf{a}}_i, \mathbf{b}_j)^{g_{ij}}(1 - P(g_{ij} = 1|\mathbf{a}_i = \overline{\mathbf{a}}_i, \mathbf{b}_j))^{1-g_{ij}} \\ &= \prod_{i=1}^{I}\prod_{j=1}^{L} p_j(\overline{\mathbf{a}}_i)^{g_{ij}}(1 - p_j(\overline{\mathbf{a}}_i))^{1-g_{ij}} \end{aligned} \tag{7.2}$$

Si los valores de los I individuos corresponden al grupo de estudio de interés, entonces el parámetro $\overline{\mathbf{A}}$ presente en la expresión de la probabilidad condicionada en 7.2 se considera como una matriz no estocástica de parámetros incidentales, el cual se puede estimar por Máxima Verosimilitud (ML) conjuntamente con los parámetros de las variables mediante la observación de la matriz $\overline{\mathbf{G}}$, teniendo en cuenta que $\overline{\mathbf{A}}$ es desconocido pero fijo.

La función de verosimilitud es precisamente la probabilidad condicional conjunta de la expresión 7.2, es decir,

$$M(\mathbf{A}, \mathbf{B}|\overline{\mathbf{G}}) = P(\mathbf{G} = \overline{\mathbf{G}}|\mathbf{A} = \overline{\mathbf{A}}, \mathbf{B})$$

y por tanto puede obtenerse un conjunto de estimadores ML para $\mathbf{A}$ y $\mathbf{B}$ de forma simultánea resolviendo las ecuaciones

$$\frac{\partial L(\mathbf{A}, \mathbf{B}|\overline{\mathbf{G}})}{\partial b_z} = 0, \tag{7.3}$$

donde b_z es un elemento de $\mathbf{B}$, y

$$\frac{\partial L(\mathbf{A}, \mathbf{B}|\overline{\mathbf{G}})}{\partial a_{is}} = 0 \ \forall i, s \tag{7.4}$$

siendo a_{is} un elemento de la matriz $\mathbf{A}$,para el logaritmo de la función de verosimilitud $L(\mathbf{A}, \mathbf{B}|\overline{\mathbf{G}}) = lnM(\mathbf{A}, \mathbf{B}|\overline{\mathbf{G}})$.

Puesto que el sistema de ecuaciones de verosimilitud no es lineal se necesita un método iterativo, por ejemplo, el método de Newton-Raphson, pero no es práctico el cálculo de la inversa de la matriz Hessiana de segundas derivadas que necesita este procedimiento cuando I(que es el número de individuos) es grande.

Bock y Lieberman [1970] propusieron un procedimiento para estimar los parámetros de los ítems o variables para el caso de un modelo de ogiva normal, y utilizándolos se podían estimar los niveles de habilidad de cada individuo [Holland, 1990]. Vamos a presentar el proceso de estimación de una manera sencilla, aunque pueden encontrarse más detalles del mismo en Bock y Lieberman [1970], Bock y Shilling [1997] and Hsu y col. [1999].

Consideremos los patrones de respuesta observados en una muestra aleatoria de I sujetos a los que se miden J variables categóricas(que transformadas según la matriz indicadora se convierten en L variables binarias, por lo que habrá 2^L patrones de respuesta posibles), indexados por $\ell = 1, 2, \cdots, u$, donde $u \leq min(I, \prod_{j=1}^{J} K_j)$, y $\mathbf{g}_\ell = (g_{\ell 1}, \ldots, g_{\ell L})$. El número de individuos o sujetos que contestan el patrón ℓ lo denotamos por r_ℓ, con $\sum_{\ell=1}^{u} r_\ell = I$

Suponiendo que las habilidades de los individuos son independientes e idénticamente distribuidas con función de densidad $g(\overline{\mathbf{a}}|\boldsymbol{\beta})$ perteneciente a una familia indexada por el vector de parámetros de la distribución $\boldsymbol{\beta}$ (que la mayoría de las ocasiones será una distribución normal multivariante), la función de probabilidad

marginal de un patrón de respuesta genérico $\overline{\mathbf{g}}$ es:

$$\begin{aligned} P(\mathbf{g}=\overline{\mathbf{g}}|\mathbf{B},\boldsymbol{\beta}) &= \int f(\overline{\mathbf{g}},\overline{\mathbf{a}}|\mathbf{B},\boldsymbol{\beta})\mathbf{d}\overline{\mathbf{a}} = \\ &= \int P(\mathbf{g}=\overline{\mathbf{g}}|\mathbf{a}=\overline{\mathbf{a}},\mathbf{B})g(\overline{\mathbf{a}}|\boldsymbol{\beta})d\overline{\mathbf{a}}, \end{aligned} \tag{7.5}$$

Para la estimación de los parámetros $\mathbf{B}$ y $\boldsymbol{\beta}$, la función de verosimilitud(marginal) es la probabilidad de observar una matriz de datos respuesta $\mathbf{G}_\wp$ que resulta tener, para $\{r_1,\cdots,r_u\}$ una distribución multinomial,

$$M \equiv P(\mathbf{G}=\mathbf{G}_\wp|\mathbf{B},\boldsymbol{\beta}) = \frac{I!}{r_1!r_2!\cdots r_u!}\prod_{\ell=1}^{u} P(\mathbf{g}=\mathbf{g}_\ell|\mathbf{B},\boldsymbol{\beta})^{r_\ell} \tag{7.6}$$

El logaritmo de la verosimilitud será

$$L = lnM = lnP(\mathbf{G}=\mathbf{G}_\wp|\mathbf{B},\boldsymbol{\beta}) = log(I!) - \sum_{\ell=1}^{u} log(r_\ell!) + \sum_{\ell=1}^{u} r_\ell lnP(\mathbf{g}=\mathbf{g}_\ell|\mathbf{B},\boldsymbol{\beta}) \tag{7.7}$$

Los estimadores MML se obtienen resolviendo las ecuaciones

$$\frac{\partial L}{\partial b_{vt}} = 0, \qquad \frac{\partial L}{\partial \beta_\nu} = 0 \tag{7.8}$$

siendo b_{vt} y β_ν elementos de la matriz $\mathbf{B}$ y del vector de parámetros $\boldsymbol{\beta}$ respectivamente.

Considerando la primera ecuación de 7.8, es decir, diferenciando respecto a un parámetro correspondiente a una variable $v(v=1,\cdots,L)$ obtenemos

$$\begin{aligned} \frac{\partial L}{\partial b_{vt}} &= \sum_{\ell=1}^{u} \frac{r_\ell}{P(\mathbf{g}=\mathbf{g}_\ell|\mathbf{B},\boldsymbol{\beta})} \frac{\partial P(\mathbf{g}=\mathbf{g}_\ell|\mathbf{B},\boldsymbol{\beta})}{\partial b_{vt}} \\ &= \sum_{\ell=1}^{u} \frac{r_\ell}{P(\mathbf{g}=\mathbf{g}_\ell|\mathbf{B},\boldsymbol{\beta})} \frac{\partial \int P(\mathbf{g}=\mathbf{g}_\ell|\mathbf{a}=\overline{\mathbf{a}},\mathbf{B})g(\overline{\mathbf{a}}|\boldsymbol{\beta})d\overline{\mathbf{a}}}{\partial b_{vt}} \\ &= \sum_{\ell=1}^{u} \frac{r_\ell}{P(\mathbf{g}=\mathbf{g}_\ell|\mathbf{B},\boldsymbol{\beta})} \int g(\overline{\mathbf{a}}|\boldsymbol{\beta}) \frac{\partial P(\mathbf{g}=\mathbf{g}_\ell|\mathbf{a}=\overline{\mathbf{a}},\mathbf{B})}{\partial b_{vt}} d\overline{\mathbf{a}} \end{aligned}$$

$$= \sum_{\ell=1}^{u} \frac{r_\ell}{P(\mathbf{g}=\mathbf{g}_\ell|\mathbf{B},\boldsymbol{\beta})} \int g(\overline{\mathbf{a}}|\boldsymbol{\beta}) P(\mathbf{g}=\mathbf{g}_\ell|\mathbf{a}=\overline{\mathbf{a}},\mathbf{B}) \frac{\partial log(P(\mathbf{g}=\mathbf{g}_\ell|\mathbf{a}=\overline{\mathbf{a}},\mathbf{B}))}{\partial b_{vt}} d\overline{\mathbf{a}}$$

$$= \sum_{\ell=1}^{u} \frac{r_\ell}{P(\mathbf{g}=\mathbf{g}_\ell|\mathbf{B},\boldsymbol{\beta})} \int g(\overline{\mathbf{a}}|\boldsymbol{\beta}) P(\mathbf{g}=\mathbf{g}_\ell|\mathbf{a}=\overline{\mathbf{a}},\mathbf{B}) \cdot$$

$$\frac{\partial \sum_{n=1}^{L}[g_{\ell n} log(P(g_{\ell n}=1|\mathbf{a}=\bar{\mathbf{a}},\mathbf{b}_n)+(1-g_{\ell n})log(1-P(g_{\ell n}=1|\mathbf{a}=\bar{\mathbf{a}},\mathbf{b}_n)))]}{\partial b_{vt}} d\bar{\mathbf{a}}$$

$$\begin{aligned}
&= \sum_{\ell=1}^{u} \frac{r_\ell}{P(\mathbf{g}=\mathbf{g}_\ell|\mathbf{B},\boldsymbol{\beta})} \int g(\bar{\mathbf{a}}|\boldsymbol{\beta}) P(\mathbf{g}=\mathbf{g}_\ell|\mathbf{a}=\bar{\mathbf{a}},\mathbf{B}) \cdot \\
&\left(\frac{g_{\ell v}}{P(g_{\ell v}=1|\mathbf{a}=\bar{\mathbf{a}},\mathbf{b}_v)} \frac{\partial P(g_{\ell v}=1|\mathbf{a}=\bar{\mathbf{a}},\mathbf{b}_v)}{\partial b_{vt}} + \right. \\
&\left. \frac{1-g_{\ell v}}{1-P(g_{\ell v}=1|\mathbf{a}=\bar{\mathbf{a}},\mathbf{b}_v)} \frac{\partial (1-P(g_{\ell v}=1|\mathbf{a}=\bar{\mathbf{a}},\mathbf{b}_v))}{\partial b_{vt}} \right) d\bar{\mathbf{a}} \\
&= \sum_{\ell=1}^{u} \frac{r_\ell}{P(\mathbf{g}=\mathbf{g}_\ell|\mathbf{B},\boldsymbol{\beta})} \int g(\bar{\mathbf{a}}|\boldsymbol{\beta}) P(\mathbf{g}=\mathbf{g}_\ell|\mathbf{a}=\bar{\mathbf{a}},\mathbf{B}) \cdot \\
&\left(\frac{g_{\ell v}}{p_v(\bar{\mathbf{a}})} \frac{\partial p_v(\bar{\mathbf{a}})}{\partial b_{vt}} + \frac{1-g_{\ell v}}{1-p_v(\bar{\mathbf{a}})} \frac{\partial (1-p_v(\bar{\mathbf{a}}))}{\partial b_{vt}} \right) d\bar{\mathbf{a}}
\end{aligned}$$

$$= \sum_{\ell=1}^{u} \int \frac{r_\ell g(\bar{\mathbf{a}}|\boldsymbol{\beta}) P(\mathbf{g}=\mathbf{g}_\ell|\mathbf{a}=\bar{\mathbf{a}},\mathbf{B})}{P(\mathbf{g}=\mathbf{g}_\ell|\mathbf{B},\boldsymbol{\beta})} \left[\frac{g_{lv}-p_v(\bar{\mathbf{a}})}{p_v(\bar{\mathbf{a}})(1-p_v(\bar{\mathbf{a}}))} \right] \frac{\partial p_v(\bar{\mathbf{a}})}{\partial b_{vt}} d\bar{\mathbf{a}}$$

Mediante un procedimiento similar, considerando β_ν como un elemento del vector $\boldsymbol{\beta}$ de parámetros correspondiente a la distribución de $\bar{\mathbf{a}}$, puede observarse que:

$$\begin{aligned}
\frac{\partial L}{\partial \beta_\nu} &= \sum_{\ell=1}^{u} \frac{r_\ell}{P(\mathbf{g}=\mathbf{g}_\ell|\mathbf{B},\boldsymbol{\beta})} \frac{\partial P(\mathbf{g}=\mathbf{g}_\ell|\mathbf{B},\boldsymbol{\beta})}{\partial \beta_\nu} \\
&= \sum_{\ell=1}^{u} \int \frac{r_\ell g(\bar{\mathbf{a}}|\boldsymbol{\beta}) P(\mathbf{g}=\mathbf{g}_\ell|\mathbf{a}=\bar{\mathbf{a}},\mathbf{B})}{P(\mathbf{g}=\mathbf{g}_\ell|\mathbf{B},\boldsymbol{\beta})} \frac{\partial ln(g(\bar{\mathbf{a}}|\boldsymbol{\beta}))}{\partial \beta_\nu} d\bar{\mathbf{a}}
\end{aligned} \tag{7.9}$$

Para el caso de un modelo de ogiva normal de dos parámetros (unidimensional) y la distribución normal estándar de $\mathbf{A}$, puede utilizarse el método de Newton-Raphson y la cuadratura S-dimensional de Gauss-Hermite para resolver las ecuaciones de verosimilitud [Bock y Lieberman, 1970]

$$\frac{\partial L}{\partial b_{vt}} = 0,$$

aproximándolas mediante:

$$\frac{\partial L}{\partial b_{vt}} \approx \sum_{qS=1}^{Q} \cdots \sum_{q1=1}^{Q} \sum_{\ell=1}^{u} \frac{r_\ell P(\mathbf{g}=\mathbf{g}_\ell|\mathbf{Y}=y_{q1\cdots qS},\mathbf{B})}{\tilde{P}(\mathbf{g}=\mathbf{g}_\ell|\mathbf{B})} \left[\frac{g_{lv}-p_v(y_{q1\cdots qS})}{p_v(y_{q1\cdots qS})(1-p_v(y_{q1\cdots qS}))}\right] \cdot \frac{\partial p_v(y_{q1\cdots qS})}{\partial b_{vt}} \psi(y_{q1})\ldots\psi(y_{qS}), \tag{7.10}$$

dónde la integral se aproxima por la cuadratura S-dimensional de Gauss-Hermite, que denotamos por $\mathbf{Y}$, obtenida como el producto de S cuadraturas unidimensionales sobre la escala de $\bar{\mathbf{a}}$, $(y_1,\ldots,y_Q)$ con Q nodos cada una, y siendo también $\{\psi(y_q): q=1,\cdots,Q\}$ los pesos asociados de la cuadratura e $y_{q1\cdots qS} \overset{not}{=} y$ cada punto de la cuadratura S-dimensional. Por tanto

$$\tilde{P}(\mathbf{g}=\mathbf{g}_\ell|\mathbf{B}) = \sum_{qS=1}^{Q} \cdots \sum_{q1=1}^{Q} P(\mathbf{g}=\mathbf{g}_\ell|\mathbf{Y}=y,\mathbf{B})\psi(y_{q1})\ldots\psi(y_{qS}) \tag{7.11}$$

Podemos escribir la aproximación dada por (7.10) de la siguiente forma, denotando por $q \overset{not}{=} (q1,\cdots,qS)$ para facilitar la notación:

$$\frac{\partial L}{\partial b_{vt}} \approx \sum_{qS=1}^{Q} \cdots \sum_{q1=1}^{Q} \frac{\bar{r}_{qv}-\bar{n}_q p_v(y)}{p_v(y)(1-p_v(y))} \frac{\partial p_v(y)}{\partial b_{vt}} \psi(y_{q1})\ldots\psi(y_{qS}), \tag{7.12}$$

de tal manera que si utilizamos la distribución normal multivariante estándar, y que las habilidades son independientes queda

$$\bar{n}_q = \sum_{\ell=1}^{u} \frac{r_\ell P(\mathbf{g}=\mathbf{g}_\ell|\mathbf{Y}=y,\mathbf{B})}{\tilde{P}(\mathbf{g}=\mathbf{g}_\ell|\mathbf{B})} \tag{7.13}$$

$$\bar{r}_{qv} = \sum_{\ell=1}^{u} \frac{r_\ell P(\mathbf{g}=\mathbf{g}_\ell|\mathbf{Y}=y,\mathbf{B}) g_{\ell v}}{\tilde{P}(\mathbf{g}=\mathbf{g}_\ell|\mathbf{B})} \tag{7.14}$$

Entonces el procedimiento EM puede describirse de la siguiente forma:

1. Utilizando valores provisionales de los parámetros de los individuos, de los parámetros de la distribución y nodos preestablecidos, calcular $\tilde{P}(\mathbf{g}=\mathbf{g}_\ell|\mathbf{B})$ para los patrones de respuesta $\ell, \ell=1,\cdots,u$.

2. (etapa "E") Calcular $\bar{n}_q$ y $\bar{r}_{qv}$ para cada variable v y cada punto o nodo de la cuadratura.

3. (etapa "M") Obtener los estimadores de los parámetros de las variables resolviendo por Newton-Raphson el sistema de ecuaciones de verosimilitud para todos ellos mediante la ecuación (7.12), donde $\bar{n}_q$ y $\bar{r}_{qv}$ se tratan como constantes.

4. Volver al paso 1 si no se ha alcanzado el criterio de convergencia elegido.

La etapa "E" del algoritmo EM consiste en encontrar (7.13) y (7.14) considerando **A** como un dato conocido provisionalmente. La etapa posterior "M" se centra en encontrar la raíz **0** de la expresión 7.12 de forma independiente para cada variable. El procedimiento EM se repite hasta que el cambio entre una iteración y otra es más pequeño que una tolerancia elegida con antelación.

En relación a la convergencia del método hay dos resultados destacables que podemos comentar; el primero es que se ha demostrado que, bajo leves condiciones, la convergencia del algoritmo está garantizada a un máximo local del logaritmo de la verosimilitud(Boyles [1983]; Dempster y col. [1977],Redner y Walker [1984] y Wu [1983]), siendo la convergencia monótona, es decir, $l(\theta^{(k+1)}) \geq l(\theta^{(k)})$, con $\theta^{(k)}$ el valor del vector de parámetros en la iteración k-ésima; y el segundo es que si consideramos el algoritmo EM como una aplicación $\theta^{(k+1)} = M(\theta^{(k)})$ que tiene un punto fijo $\theta^* = M(\theta^*)$ entonces se demuestra que $\theta^{(k+1)} - \theta^* \approx \frac{\partial M(\theta^*)}{\partial \theta^*}(\theta^{(k)} - \theta^*)$ cuando $\theta^{(k+1)}$ está cerca de θ^* y por tanto

$$\left\|\theta^{(k+1)} - \theta^*\right\| \leq \left\|\frac{\partial M(\theta^*)}{\partial \theta^*}\right\| \left\|\theta^{(k)} - \theta^*\right\|$$

siendo $\left\|\frac{\partial M(\theta^*)}{\partial \theta^*}\right\|$ casi seguro. Por tanto este algoritmo es de orden[1] 1, lo cual conlleva a tener ciertas precauciones sobre él, puesto que esto supone un inconveniente

[1]Un algoritmo iterativo se dice que tiene una tasa de convergencia local de orden $q \geq 1$ si $\frac{||\theta^{(k+1)} - \theta^*||}{||\theta^k - \theta^*||^q} \leq r + o(||\theta^k - \theta^*||)$ para k suficientemente grande.

según citan algunos autores como Redner y Walker [1984] que proponen otro tipo de métodos de segundo orden para ciertas situaciones.

El algoritmo evoluciona de forma lenta hasta alcanzar la solución, aunque se han propuesto soluciones para agilizar los cálculos, como emplear un factor de aceleración para resolver las ecuaciones implícitas [Ramsay, 1975]. Otra podría ser la ordenación de los vectores de respuesta formando grupos de puntuaciones y calculando las verosimilitudes para los patrones, cambiando los factores sólo cuando los unos o ceros difieren entre patrones, resultando esta forma de proceder en un ahorro de tiempo, computacionalmente hablando.

7.2. Una alternativa al procedimiento EM

Las etapas del algoritmo EM se pueden derivar como una extensión del principio de información perdida utilizado por Dempster y col. [1977] para obtener estimadores de máxima verosimilitud cuando el modelo de probabilidad pertenece a la familia exponencial. En este caso existiría un estadístico suficiente para $\mathbf{a}$, y, según el principio citado, los valores esperados de este estadístico dados los datos observados se sustituirían en la etapa de maximización del algoritmo.

En el caso de no existir dichos estadísticos suficientes para $\mathbf{a}$ puede considerarse una aproximación a la verosimilitud y reemplazar cada observación correspondiente a cada individuo $\mathbf{a}_i$ por su esperanza condicionada, dado el valor observado $\mathbf{g}_i$. Teniendo en cuenta el teorema de Bayes , la distribución condicional de $\mathbf{a}$ dado $\mathbf{g} = \mathbf{g}_i$ es

$$g(\mathbf{a}|\mathbf{g}_i, \boldsymbol{\beta}) = \frac{P(\mathbf{g} = \mathbf{g}_i|\mathbf{a})g(\mathbf{a}, \boldsymbol{\beta})}{P(\mathbf{g} = \mathbf{g}_i)}$$

y por tanto, la esperanza condicionada de $\mathbf{a}$ dado $\mathbf{g} = \mathbf{g}_i$, considerando que

$$P(\mathbf{g} = \mathbf{g}_i) = \int P(\mathbf{g} = \mathbf{g}_i|\mathbf{a})g(\mathbf{a}|\boldsymbol{\beta})d\mathbf{a}$$

y que,siguiendo una notación análoga a la descrita en la sección 7.1

$$\begin{aligned} P(\mathbf{g}=\mathbf{g}_i|\mathbf{a}_i) &= \prod_{n=1}^{L}[P(g_{in}=1|\mathbf{a}=\mathbf{a}_i)]^{g_{in}}[1-P(g_{in}=1|\mathbf{a}=\mathbf{a}_i)]^{1-g_{in}} \\ &= \prod_{n=1}^{L}[\Phi_n(\mathbf{a}_i)]^{g_{in}}[1-\Phi_n(\mathbf{a}_i)]^{1-g_{in}} \end{aligned} \tag{7.15}$$

llamando, para simplificar la notación, $P(g_{in}=1|\mathbf{a}=\mathbf{a}_i) \overset{not}{=} \Phi_n(\mathbf{a}_i)$, tiene la siguiente expresión:

$$E(\mathbf{a}|\mathbf{g}_i) = \frac{\int \mathbf{a} g(\mathbf{a}) \prod_{n=1}^{L}[\Phi_n(\mathbf{a})]^{g_{in}}[1-\Phi_n(\mathbf{a})]^{1-g_{in}} d\mathbf{a}}{\int g(\mathbf{a}) \prod_{n=1}^{L}[\Phi_n(\mathbf{a})]^{g_{in}}[1-\Phi_n(\mathbf{a})]^{1-g_{in}} d\mathbf{a}} \tag{7.16}$$

Aproximando las integrales por sumas de Q puntos, y recodificando al i-ésimo sujeto al l-ésimo patrón de puntuaciones, tenemos según (7.10) que la expresión de la esperanza marginal a posteriori es:

$$E(\mathbf{a}|\mathbf{g}_\ell) \cong \frac{\sum_{qS=1}^{Q} \cdots \sum_{q1=1}^{Q} y_{q1\cdots qS} P(\mathbf{g}=\mathbf{g}_\ell|\mathbf{Y}=y_{q1\cdots qS},\mathbf{B})\psi(y_{q1})\ldots\psi(y_{qS})}{\tilde{P}(\mathbf{g}=\mathbf{g}_\ell|\mathbf{B})}, \tag{7.17}$$

siendo $y_{q1\cdots qS}$ los puntos de la cuadratura multivariante S-dimensional, tal y como fue denotada anteriormente, y $\tilde{P}(\mathbf{g}=\mathbf{g}_\ell|\mathbf{B})$ dada por (7.11).

Hay u patrones de puntuación diferentes, por tanto habrá u valores calculados mediante la expresión anterior: $\{E(\mathbf{a}|\mathbf{g}_\ell); \ell = 1, \cdots, u\}$. La habilidad para el individuo i que tiene un patrón ℓ, tiene S componentes (tantas como dimensiones del espacio solución, es decir, $(\mathbf{a}_i = (a_{i1}, \cdots, a_{iS}))$, y cada una $\{a_{is}, s = 1, \cdots, S\}$ será aproximada por la expresión 7.17, la cual depende de cada coordenada S-dimensional de $y_{q1\cdots qS}$. Pueden consultarse más detalles de este procedimiento en Bock y Aitkin [1981], que será el que hemos implementado (sección 7.7) con las correcciones necesarias que detallan los dos siguientes apartados.

7.3. Propiedades muestrales de los estimadores máximo verosímiles.

La determinación de los errores estándar para los parámetros estimados no es posible con estos algoritmos, siendo el procedimiento habitual en el caso de máxima verosimilitud y de mínimos cuadrados generalizados el cálculo de la matriz de covarianzas asintótica utilizando la matriz de información. Pero cuando el número de factores es elevado también esta alternativa es compleja, por lo que es necesario buscar soluciones a este problema.

Es conocido que los elementos de la matriz de información de Fisher evaluados en el punto solución, es decir, en el estimador calculado, proporcionan asintóticamente la matriz de covarianzas muestrales. Para un conjunto de parámetros $\boldsymbol{\beta}$ tenemos

$$Var(\hat{\boldsymbol{\beta}})^{-1} = E\left[-\frac{\partial^2 L}{\partial\beta_i\partial\beta_j}\right]_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}$$

con L el logaritmo de la verosimilitud, y

$$\frac{\partial^2 L}{\partial\beta_i\partial\beta_j} = \sum_{h=1}^{I}\left[\frac{1}{f}\frac{\partial^2 f}{\partial\beta_i\partial\beta_j} - \frac{1}{f^2}\frac{\partial f}{\partial\beta_i}\frac{\partial f}{\partial\beta_j}\right] \tag{7.18}$$

, denotando por $f \equiv f(\mathbf{x}_h)$ a la función de densidad conjunta de la muestra y $\mathbf{x}_h$ el vector de observaciones correspondiente al individuo h-ésimo. Calculando la esperanza, el primer término se anula, por tanto tendríamos

$$Var(\hat{\boldsymbol{\beta}})^{-1} = n\left[E\frac{1}{f^2}\frac{\partial f}{\partial\beta_i}\frac{\partial f}{\partial\beta_j}\right]_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}$$

Esta última esperanza, calculada sobre todos los posibles valores del vector $\mathbf{x}$ y sus posibles patrones, queda

$$\sum_{\mathbf{x}}\frac{1}{f(\mathbf{x})}\frac{\partial f(\mathbf{x})}{\partial\beta_i}\frac{\partial f(\mathbf{x})}{\partial\beta_j}$$

Si el número de variables es pequeño esta suma se puede calcular y por tanto se puede invertir la matriz resultante. Pero si no ocurre esto ese sumatorio resulta

muy problemático, e incluso puede que muchas de las probabilidades $f(\mathbf{x}_h)$ sean tan pequeñas que el cálculo de su inverso produzca errores computacionales. En estos casos se puede obtener una aproximación sustituyendo la esperanza de la matriz de información por su valor observado. Para obtener esto hay que calcular la expresión 7.18 y una inversa generalizada de la matriz resultante, es decir, quedaría

$$Var^*(\hat{\boldsymbol{\beta}}) = \left[\sum_{h=1}^{I} \frac{1}{f^2(\mathbf{x}_h)} \frac{\partial f(\mathbf{x}_h)}{\partial \beta_i} \frac{\partial f(\mathbf{x}_h)}{\partial \beta_j}\right]^{-1} \tag{7.19}$$

Esta aproximación dada por 7.19 parece ser buena para los errores estándar y menos fiable para las covarianzas. Albanese y Knott [1994] investigaron el comportamiento de los errores estándar de los estimadores de máxima verosimilitud para los modelos de un factor utilizando métodos bootstrap y se dieron cuenta de que cuando los parámetros de discriminación son pequeños la teoría asintótica funciona bien, pero a medida que aumentan los resultados clásicos no se cumplen.

En el caso de tener dos o más factores latentes el cálculo de estos errores requiere más atención. La situación requiere tener en cuenta el hecho de que existe una indeterminación debido a la existencia de diversas soluciones rotadas. Si se imponen suficientes restricciones a los parámetros de tal forma que se asegure la existencia de un único máximo se podrían utilizar las aproximaciones asintóticas estándar como en el caso de los modelos de variables latentes usuales. Sin ellas, el máximo no es único y la matriz de información será singular. No obstante, es posible transformar cualquier solución basada en máxima verosimilitud, por medio de rotaciones, en otra que verifique condiciones de ortogonalidad deseables para poder utilizar la teoría asintótica y encontrar los errores de la solución rotada.

7.4. El problema de la Separación en regresión logística.

La explicación de un fenómeno y de su posible comportamiento probabilístico se modeliza mediante la utilización de técnicas de regresión logística. Con frecuencia

ocurre que los datos están separados, de forma que por ejemplo, con variables binarias, aparecen los éxitos separados de los fracasos, lo cual implica que no es posible calcular los estimadores de máxima verosimilitud. Puesto que dichas técnicas permiten ordenar simultáneamente las variables, es muy probable que ocurra un problema de separación al estimar los parámetros del biplot.

Numerosos autores han tratado la existencia[2] de estimadores máximo-verosímiles en el modelo de regresión logística, como Haberman [1974], Wedderburn [1976], Silvapulle [1981] ó Albert y Anderson [1984]. Estos últimos autores, considerando las posibles configuraciones de n puntos muestrales en $\mathbb{R}^k$, examinan el problema de la maximización del logaritmo de la función máximo verosímil. Dichas configuraciones se pueden considerar desde tres puntos de vista mutuamente excluyentes, que son la separación completa, la cuasi-separación y el solapamiento.

En la regresión logística para el caso de una respuesta nominal (la binaria es un caso particular) se consideraba una categoría como base, supongamos la primera, y se restringían los parámetros de dicha categoría para que fueran cero. Si la variable respuesta tiene r grupos, el vector de parámetros del grupo era $\mathbf{b}_s = (b_{0s}, b_{1s}, \ldots, b_{ks})'$, con $(s = 2, \ldots, r)$. Llamamos $\mathbf{b}_1 = (b_{01}, b_{11}, \ldots, b_{k1})' = \mathbf{0}$ al vector de la categoría base. En la regresión logística binaria sólo teníamos un vector de parámetros $\mathbf{b} = (b_0, b_1, \ldots, b_k)'$, que se correspondía con la categoría 1 de presencia. El primer parámetro de cada vector es la constante y supone un traslado del centro de gravedad, lo cual no afecta al comportamiento del modelo.

El modelo logístico se puede utilizar como método de clasificación de un individuo dada una combinación de las variables independientes dadas por el vector $\mathbf{x}$ [Albert y Anderson, 1984]. Lo que se hace es asignar la observación $\mathbf{x}$ al grupo s si y sólo si

$$(\mathbf{b}_s - \mathbf{b}_t)'\mathbf{x} \geq 0 \qquad \text{con } t = 1, \ldots, r$$

.

En el caso de la regresión binaria, se asigna la observación $\mathbf{x}$ al grupo 1, o grupo

[2]Nos referimos a la ausencia de un máximo finito

de presencia si y sólo si $\mathbf{b}'\mathbf{x} \geq 0$ y al grupo 0 en caso contrario. Esta condición es equivalente a asignar al grupo 1 la observación $\mathbf{x}$ si $P(Y = 1/\mathbf{x}) \geq 0{,}5$ y al grupo 0 en otro caso. Es decir, el espacio de las variables independientes queda dividido en dos regiones, que son las que clasifican en el grupo 1 y 0 respectivamente, y ambas se separan por el hiperplano $\mathbf{b}'\mathbf{x} = 0$. Geométricamente la figura 4.3 plasmaba ambas regiones.

En el caso de variables respuesta nominales con más de 2 categorías la situación es más compleja, puesto que como veíamos en el capítulo 5, para cada categoría tenemos una superficie de respuesta, y la intersección de cada par de ellas era una línea recta (figura 5.2), con ecuaciones

$$(\mathbf{b}_s - \mathbf{b}_t)'\mathbf{x} = \mathbf{0} \qquad \text{con } \ s, t = 2, \ldots, r$$

si no se incluye la categoría base en la comparación, y $\mathbf{b}_s'\mathbf{x} = 0$ con $s = 2, \ldots, r$ en las comparaciones con la base. Dichas rectas definen las conocidas regiones convexas que permiten clasificar a cada individuo en una de ellas(figura 5.4). Esta idea de utilizar la regresión logística como método discriminante, así como su estimación máximo verosímil ha sido investigada por autores como Cox [1970], Anderson [1972] ó Anderson y Philips [1981]. Vamos a presentar las posibles configuraciones muestrales que citábamos anteriormente.

7.4.1. Separación completa.

Supongamos que medimos un conjunto de variables explicativas $X = (X_1, \ldots, X_J)$ sobre r grupos diferentes, teniendo así una variable respuesta multinomial Y. Llamemos E_s al conjunto de puntos que pertenecen al grupo observado s.

Decimos que el conjunto de datos X es completamente separable si existe un vector $\mathbf{b} \in \mathbb{R}^k$ tal que para todo $i \in E_j$ y $j, t = 1, \ldots, r (j \neq t)$ se verifica que

$$(\mathbf{b}_j - \mathbf{b}_t)'\mathbf{x}_i > 0$$

,es decir, existe un vector que clasifica correctamente todas las observaciones. Llamando A^c al conjunto de vectores que satisfacen la desigualdad anterior puede

demostrarse que dicho conjunto es un cono convexo en $\mathbb{R}^k$, o sea, que si $\mathbf{b} \in A^c$, entonces $\mathbf{b} + \Delta \in A^c$, donde $\Delta \neq k\mathbf{b}$ puede ser elegido tan pequeño como sea necesario para que verifique dicha ecuación. Si A^c contiene rectas $k\mathbf{b}$ contendrá también un haz de rectas. Albert y Anderson [1984] demuestran que si existe una separación completa de puntos, entonces la estimación máximo verosímil $\widehat{\mathbf{b}}$ no existe, y además $max_{\mathbf{b}\in\mathbb{R}^k}\mathbf{M}(X, \mathbf{b}) = 1$, siendo $\mathbf{M}$ la función de verosimilitud.

7.4.2. Separación cuasi-completa.

Si el conjunto de datos $\mathbf{X}$ no es completamente separable es necesario ampliar o matizar el concepto de separación para otras situaciones. Decimos que el vector $\mathbf{b} \in \mathbb{R}^k$ proporciona una separación cuasicompleta para un conjunto de datos X si para todo $i \in E_s$ y todo $s, t = 1, \ldots, g(s \neq t)$ se cumple

$$(\mathbf{b}_s - \mathbf{b}_t)'\mathbf{x}_i \geq 0$$

teniéndose la igualdad al menos para una terna (i, s, t).

Sea $s(i)$ el valor de s para el que $i \in E_s$, y sea $Q(\mathbf{b})$ el conjunto de valores $i \in E$,$(E = \cup E_s)$ que satisfacen la desigualdad anterior. Los puntos $\mathbf{x}_i$ se dice que están cuasi-separados con respecto a $\mathbf{b}$.

Puede demostrarse que los estimadores máximo verosímiles tampoco existen en este caso.

7.4.3. Solapamiento.

Si en el conjunto de los datos la separación no es ni completa ni cuasi-completa entonces habrá un solapamiento, es decir, para cualquier $\mathbf{b} \in \mathbb{R}^k$, existe una terna (i, s, t) donde $s, t \in \{1, \ldots, g\}, j \neq t, i \in E_s$, y;

$$(\mathbf{b}_s - \mathbf{b}_t)'\mathbf{x}_i < 0$$

.

La estimación máximo verosímil existe y es única en este caso, pero si el solapamiento no es muy pronunciado es posible que se presenten inestabilidades en la estimación de los parámetros. Esta propiedad esta demostrada para modelos de respuesta binomial por Silvapulle [1981] y también por Haberman [1974] en su trabajo sobre modelos logarítmico lineales, que incluyen la regresión logística binomial.

Estas tres situaciones se ilustran en la figura 7.1 para un caso en el que tenemos dos grupos solamente.

7.4.4. Detección de la separación.

Las estimaciones que maximizan la verosimilitud en el caso de que exista separación no tienen forma explícita y para calcularlas se utilizan algoritmos iterativos que hacen uso del método de Newton-Raphson con valores iniciales cero para los parámetros. La cuestión de la clasificación de los conjuntos de datos en los tres estadios comentados en las secciones anteriores se puede abordar de manera algebraica o empírica.

El enfoque algebraico utiliza la teoría de programación lineal. Con dos grupos, hay separación completa o cuasi-completa si existe $\mathbf{b_1} \neq 0$ tal que

$$\mathbf{X}^*\mathbf{b_1} \leq \mathbf{0} \tag{7.20}$$

, siendo $\mathbf{X}^*{}_{n\times(k+1)}$ con filas $-\mathbf{x}_i, i \in E_1$ y $\mathbf{x}_i, i \in E_2$. Para la separación cuasi-completa, la igualdad debe mantenerse por lo menos para un valor de i. Si la ecuación anterior no se satisface con $\mathbf{b_1} \neq 0$ entonces el conjunto de datos estará solapado.

De forma genérica podemos escribir

$$(\mathbf{X}^*, \mathbf{I_n})(\mathbf{b_1}, \mathbf{t}) = 0 \tag{7.21}$$

donde $\mathbf{I_n}$ es la matriz identidad de orden n y $\mathbf{t}$ es un vector fila de n variables adicionales. Por tanto la separación completa se produce si existe una solución

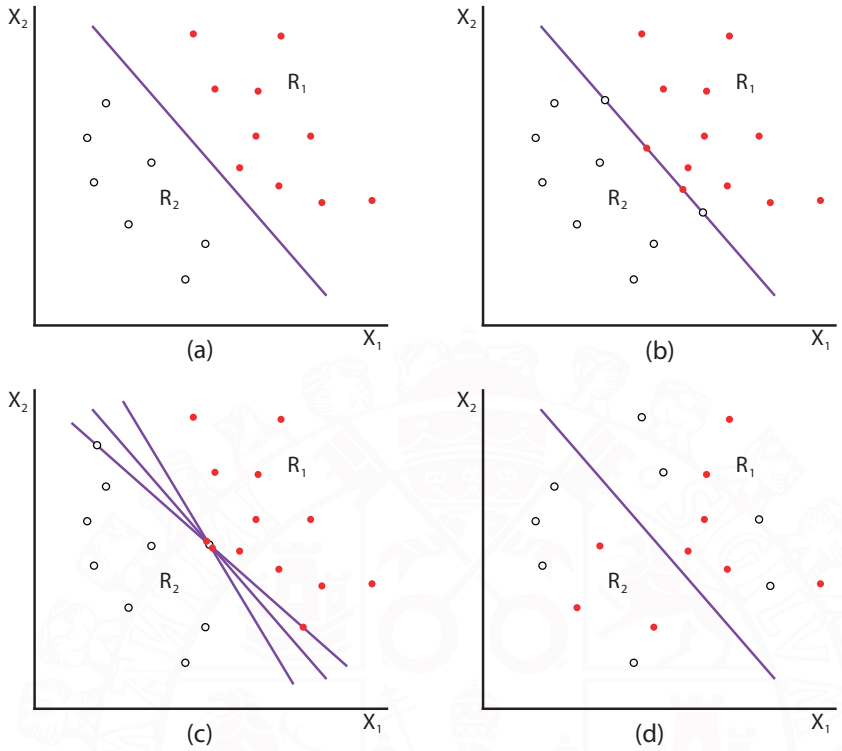


Figura 7.1: Configuraciones de puntos posibles, en un hipotético caso en el que se trabaja con dos variables y dos grupos, determinadas por las regiones R_1 y R_2. En el caso (a) existe Separación completa; en (b) hay cuasi-separación con un único hiperplano de separación; en (c) existe cuasi-separación con varios hiperplanos válidos, puesto que hay tres puntos en la intersección de las rectas; y en (d) se ilustra el solapamiento.

para 7.21 con todas las $t_i > 0$, $i = 1, \ldots, n$; la cuasi-separación completa si existe una solución con $t_i \geq 0$, $i = 1, \ldots, n$, dándose la igualdad para al menos un valor de i; y el solapamiento si se satisfacen alguna de las dos situaciones anteriores, es decir, todas las soluciones tienen algún t_i positivo y alguno negativo.

El conjunto de soluciones de la ecuación 7.21 se corresponde con encontrar

el conjunto de soluciones factibles de un problema de programación lineal, de tal forma que adaptando los métodos estándar se puede determinar si estos conjuntos son vacíos. La generalización al caso de más de dos grupos es inmediata.

Por otra parte, desde una perspectiva empírica, se pueden encontrar soluciones insertando una regla de parada al algoritmo si se encuentra separación completa, o cuando el número de condición de la matriz hessiana alcanza un valor determinado, e incluso se pueden corregir las probabilidades esperadas de manera que no alcancen los valores 0 ó 1, es decir, que no tengamos una predicción perfecta. En el caso de separación cuasi-completa la situación es diferente, pero para algunos puntos, cuando el proceso diverge, la probabilidad de pertenecer al grupo correcto tiende a uno rápidamente. Así que en cada iteración $t \geq T$ se busca el punto $\mathbf{x}$ con la probabilidad más alta de asignación correcta a través del conjunto de datos, $pr^m(\mathbf{b}_t)$, y hay que mostrar un mensaje de aviso si:

$$pr^m(\mathbf{b}_t) > min\{1 - \epsilon, pr^m(\mathbf{b}_{t-1})\} \tag{7.22}$$

Esta expresión muestra que la probabilidad de asignación correcta es muy cercana a 1 al menos para una observación $\mathbf{x}$. En este caso hay dos opciones: la primera es que existe solapamiento y esta observación es atípica en su grupo, lejos de la media, y por tanto esta advertencia no es necesaria, y el proceso por tanto puede continuar y se detiene cuando alcance su máximo; la segunda es que existe separación cuasi-completa y dicho punto es de los puntos totalmente separados, estando la matriz de dispersión asintótica no acotada. Lo que se requiere en una situación así es que el programa se ejecute de nuevo con todos los vectores de observación estandarizados y el proceso se detiene si cualquier elemento de la diagonal de la matriz de dispersión es mayor que 10^3, recomendándose la activación de las alertas tras varias iteraciones.

Es muy interesante el estudio comparativo que se lleva a cabo en el trabajo de Nieto y Vicente-Villardón [2012] con los resultados proporcionados por diversos paquetes estadísticos en situaciones en las que se utiliza la regresión logística con

distintos tipos de variables respuesta. Estos ponen de manifiesto, como veremos en el siguiente apartado, que los métodos penalizados estabilizan la estimación de los coeficientes en situaciones de separación, resultando sorprendente que algunos de esos paquetes de extendido uso no son capaces de realizar estimaciones correctas en casos límite.

7.4.5. Soluciones al problema de la separación.

Este problema surge de la observación de procesos de regresión logística en los que el modelo converge, pero al menos un parámetro diverge [Heinze y Schemper, 2002]. La separación tiene lugar en muestras pequeñas con varios factores de riesgo desequilibrados y con un alto nivel de predicción. Firth [1993] desarrolló un mecanismo para reducir el sesgo de las estimaciones por máxima verosimilitud, el cual trata de ofrecer una solución al problema de la separación.

Cuando se trabaja con conjuntos de datos medianos y pequeños, en regresión logística pueden presentarse situaciones en las que algunos de los parámetros son muy inestables, incluso aunque el método converja. Esto hace que por ejemplo los intervalos de confianza de Wald tiendan a infinito. En estudios como los de Heinze y Schemper [2002] muestran que la probabilidad de separación depende de varios factores, como la muestra, el número de factores de riesgo dicotómicos, la magnitud de los odds ratios asociados a ellos y el grado de equilibrio en su distribución.

Los procesos de Firth [1993], cuando se emplean en este problema, se pueden englobar dentro de un grupo más amplio de técnicas que se conocen con el nombre de regresión logística penalizada. En este contexto estarían la regresión ridge, que propusieron Hoerl y Kennard [1970] para la regresión lineal y extendida para la regresión logística por Le Cessie y Van Houwelingen [1992]. Dicha alternativa consiste en restringir la longitud del vector de parámetros, existiendo versiones en las que es posible la penalización de cada coeficiente individualmente [Harrel, 2001]. Estos métodos de regresión penalizada siguen evolucionando en la actualidad y se utilizan en contextos como la genética[Malo y col., 2008] y los microarrays(Zhu y

Hastie [2004];Sun y Wang [2012]).

Los métodos de regresión que utilizan técnicas de penalización se han asociado en la literatura al tratamiento de la colinealidad de los predictores, muestras pequeñas en relación al número de variables o tablas poco ocupadas. Vamos a detallar las ecuaciones básicas del procedimiento de Firth.

Procedimiento de Firth

La solución de las ecuaciones $\frac{\partial logM(\mathbf{b})}{\partial \mathbf{b}} \equiv U(\mathbf{b}) = 0$ proporciona las estimaciones de máxima verosimilitud de los parámetros de regresión $\widehat{\mathbf{b}}$, con M la función de verosimilitud. La utilización de Newton-Raphson consiste en actualizar el vector de parámetros en cada iteración como

$$\mathbf{b}_{l+1} = \mathbf{b}_l + \mathbf{I}(\mathbf{b}_l)^{-1}\mathbf{U}(\mathbf{b}_l) = \mathbf{b}_l + (\mathbf{X}'\mathbf{V}_l\mathbf{X})^{-1}\mathbf{X}'(\mathbf{y} - \boldsymbol{\pi}_l) \tag{7.23}$$

siendo $\mathbf{I}(\mathbf{b}_l)$ la matriz de información de Fisher, $\boldsymbol{\pi}_l = \frac{1}{1+e^{-\mathbf{x}_i'\mathbf{b}_l}}$ el vector de probabilidades estimadas en la iteración l y $\mathbf{V}_l = diag\{\hat{\pi}_{li}(1-\hat{\pi}_{li})\}$.

Firth propuso, con el objetivo de reducir el sesgo de las estimaciones, basar estas en ecuaciones modificadas para el j-ésimo parámetro

$$\mathbf{U}^*(b_j) = \mathbf{U}(b_j) + \frac{1}{2}traza\left[\mathbf{I}(\mathbf{b})^{-1}\left\{\frac{\partial \mathbf{I}(\mathbf{b})}{\partial b_j}\right\}\right] \tag{7.24}$$

La función $\mathbf{U}^*(b_j)$ está relacionada con la penalización de una función de verosimilitud que consigue eliminar el sesgo de los estimadores.

La idea de Firth se puede aplicar en el caso de trabajar con un modelo logístico, reemplazando el vector de primeras derivadas $\mathbf{U}(\mathbf{b}) = \mathbf{X}'(\mathbf{y} - \boldsymbol{\pi})$ por

$$\mathbf{U}^*(\mathbf{b}) = \mathbf{X}'(\mathbf{y} - \boldsymbol{\pi} + \frac{1}{2}\mathbf{h} - \mathbf{h} * \boldsymbol{\pi})$$

donde * es el producto elemento a elemento de vectores, $\mathbf{h}$ es el vector de los elementos de la diagonal de la matriz $\mathbf{H} = \mathbf{V}^{\frac{1}{2}}\mathbf{X}(\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{\frac{1}{2}}$, con $\mathbf{V} = diag\{\pi_i(1-\pi_i)\}$, y π_i el vector de probabilidades estimadas para un individuo.

7.5. Bondad de ajuste en los modelos de Teoría de la Respuesta al Ítem.

El contraste de la hipótesis que se plantea es $H_0 : \boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\alpha})$, siendo $\boldsymbol{\pi}$ la probabilidad teórica y real y $\boldsymbol{\pi}(\boldsymbol{\alpha})$ la probabilidad calculada con el modelo propuesto. Las proporciones de la muestra $\mathbf{P}$ se comparan con las estimadas, que son $\boldsymbol{\pi}(\hat{\boldsymbol{\alpha}})$, siendo $\hat{\boldsymbol{\alpha}}$ el estimador máximo verosímil o cualquier otro que se pudiera calcular. Si I es bastante más grande que 2^J, la frecuencia esperada para cada patrón de respuesta seguramente es suficientemente grande para llevar a cabo un test chi-cuadrado(χ^2) válido o bien un test de razón de verosimilitud(G^2) que comparen las frecuencias esperadas y observadas. Incluso, hasta un cierto punto, pequeñas frecuencias esperadas se podrían tratar agrupando patrones de respuesta de tal forma que sean, digamos, mayores que 5. En el caso no agrupado el número de grados de libertad será $2^J - J(S+1) - 1$. A medida que 2^J es más grande, la necesidad de agrupación podría ser tal que redujera el número de grados de libertad a cero, de tal forma que no exitirían tales tests, lo cual no es extraño, puesto que en muchas ocasiones se tienen más de 10 variables por ejemplo. Un enfoque alternativo podría ser pensar en el conjunto de datos como una tabla de contingencia de tamaño 2^J de tal forma que nuestro problema sería cómo medir la bondad de ajuste en tablas de contingencia dispersas.

Un ajuste inadecuado conduce a la obtención de unos parámetros de los ítems y de la habilidad que no son invariantes, por tanto es necesario disponer de un conjunto de tests de bondad de ajuste que proporcionen una herramienta para asegurar que se está utilizando el modelo adecuado. No obstante, Smith [2002] analizó la aplicación de algunos estadísticos de ajuste y decía que no existe un estadístico universal que sea óptimo para detectar cada tipo posible de perturbaciones de medida. Es decir, que cada estadístico presenta siempre unas bondades y a su vez debilidades, que habrá que analizar para utilizar uno u otro.

Los indicadores de bondad de ajuste dependen fuertemente del tamaño de la

muestra. Para algunos de ellos, como la chi-cuadrado del cociente de verosimilitud, las muestras que son muy grandes distorsionan el estadístico, aumentando artificialmente su valor y conduciendo a formular conclusiones erróneas sobre el conjunto de datos [Byrne, 2001]. Si el tamaño de muestra está entre 100 y 1000 la chi-cuadrado puede ser un indicador de bondad de ajuste adecuado, teniendo además la ventaja de que se conoce su distribución, pero para muestras pequeñas es problemático debido a su falta de potencia estadística [Hambleton, 2000].

Existen estudios basados en el método de Monte-Carlo que muestran que los procedimientos basados en la chi-cuadrado pueden identificar adecuadamente un modelo ajustado de forma apropiada con el modelo de Rasch en muestras de hasta 500 elementos y 50 variables [McKinley y Mills, 1995]. Estos autores realizaron pruebas con conjuntos de datos unidimensionales y multidimensionales de diferentes tamaños de muestra y distinto número de variables (hasta 2000 individuos y 75 variables) analizando el comportamiento de diversos indicadores basados en la chi-cuadrado, con el objetivo de determinar si estos eran capaces de identificar la falta de ajuste en algunos ítems. Mostraron que a medida que el tamaño de muestra aumentaba los estadísticos se distorsionaban, lo cual era más evidente en individuos con menor habilidad. Además en el caso multidimensional esta distorsión era mucho mayor que en caso unidimensional.

Para tratar de corregir esta disfunción con el tamaño de muestra se han propuesto otros indicadores basados en los anteriores dividiendo estos entre el tamaño de la muestra, que se llaman estadísticos de ajuste basados en la media cuadrática. Son estadísticos que calculan la cantidad de distorsión en el sistema de medida cuyo valor esperado es 1. Valores menores que 1 indican o bien un sobreajuste de los datos al modelo o bien la existencia de redundancia de los datos, mientras que valores mayores que 1 indican ruido aleatorio. Tienen el inconveniente de que no se conoce su distribución a diferencia de los indicadores chi-cuadrado.

Hulin y col. [1982] utilizan el indicador Raíz del Error Cuadrático Medio (RM-

SE) en la recuperación de curvas características de ítems con modelos[3] 2-PL y 3-PL. Drasgow y Parsons [1983] utilizaban las diferencias de las raíces de medias cuadráticas para recuperar de forma adeacuada los parámetros de ítems para el modelo 2-PL. En ambos estudios se pudo comprobar cómo el estadístico de ajuste mostraba mínimas distorsiones para tamaños de muestras mayores de 2000 individuos sobre evaluaciones de los mismos entre 15 y 65 variables o ítems.

Zhao y col. [2002] trabajan tanto con el RMSE como con el estimador de la media del error estándar(ASE) y se dan cuenta de que el RMSE para el modelo 3-PL capturaba la peculiaridad para cada dimensión de forma más precisa que el RMSE de los modelos 1-PL(conocido como modelo de Rasch[4], $RMSE_1$) y 2-PL($RMSE_2$). Zhao y col. [2002] econtraron que $RMSE_3$ <$RMSE_1$ <$RMSE_2$, relación que se cumplía utilizando tanto métodos de estimación de máxima verosimilitud, como bayesianos secuenciales y bayesianos de tipo EAP.

Por otra parte, el indicador Raíz del Residuo Cuadrático Medio de Aproximación (RMSEA)[Steiger y Lind, 1980] tiene en cuenta la complejidad del modelo y además el tamaño de la muestra, de tal forma que valores muy cercanos a 0 indican un ajuste excelente, valores menores que 0.05 se consideran correspondientes a un ajuste aceptable [Steiger, 1990] y valores entre 0.08 y 0.1 denotan ajustes mediocres [Brown y Cudeck, 1993]. Esta medida es muy utilizada en el campo de la modelización de ecuaciones estructurales y provee un mecanismo para ajustar

[3]Considerando modelos con 2 factores, el modelo 2-4PL se puede expresar

$$P(x = 1/\theta, \psi) = g + \frac{(u - g)}{1 + e^{-(a_1 * \theta_1 + a_2 * \theta_2 + d)}}$$

y dependiendo del modelo u será o no igual a 1 y g será o no igual a 0.

[4]Considerando $X_{ni} = x$, con $x \in 0, 1, \ldots, m_i$, siendo m_i el número de categorías del ítem i se tiene que en un modelo politómico

$$P(X_{ni} = x, x > 0) = \frac{e^{\sum_{k=1}^{x}(\beta_n + \tau_{ki})}}{1 + \sum_{j=1}^{m_i} e^{\sum_{k=1}^{j}(\beta_n + \tau_{ki})}}$$

,siendo τ_{ki} es el k-ésimo umbral del ítem i y β_n es la posición del individuo n correspondiente a ese ítem.

correctamente cuando se utilizan estadísticos chi-cuadrado.

Orlando y Thissen [2000, 2003] proponen un indicador de ajuste para cada ítem, S-X^2, en modelos de IRT dicotómicos que tiene un comportamiento mejor que los estadísticos tradicionales como el Q_1 de Yen [1981] y el G^2 de McKinley y Mills [1995], siendo extendido al caso de modelos politómicos de IRT, incluidos el modelo de crédito parcial generalizado[Muraki, 1992] y el modelo de crédito parcial[Masters, 1982], por Kang y Chen [2007].

Existen diversas alternativas para intentar conseguir p-valores más precisos; los métodos de remuestreo tales como el método paramétrico bootstrap (Bartholomew y Tzamourani [1999]; Langeheine y col. [1996]; Tollenaar y Mooijaart [2003]) son una alternativa, aunque existe una evidencia fuerte de que no conducen a la precisión deseada(Mavridis y col. [2007]; Tollenaar y Mooijaart [2003]), teniendo el inconveniente además de que son muy pesados computacionalmente.

No obstante, un test de ajuste global es útil como una primera etapa de estudio del ajuste, pero puede ser erróneo para revelar cualquier desviación del modelo respecto a los datos. Podría ocurrir que precisamente algunos de los ítems sean los responsables de un mal ajuste porque dependieran de algún factor específico para ellos únicamente. El efecto de esta situación podría ser muy difuso en los patrones de respuesta y por tanto muy difícil de detectar en las contribuciones individuales a estadísticos como la χ^2 por ejemplo. Por este motivo puede ser recomendable llevar a cabo tests suplementarios para detectar los efectos de esos ítems inapropiados.

Estos tests se basan en los residuos calculados con las frecuencias marginales de varios órdenes($P(x_{j_1}) = 1, P(x_{j_1} = 1, x_{j_2} = 1), P(x_{j_1} = 1, x_{j_2} = 1, x_{j_3} = 1), j_i = 1, \ldots, J$); normalmente hasta el tercero es suficiente. Si denotamos por O la frecuencia observada para cualquier probabilidad marginal y E la correspondiente frecuencia esperada, se define el residuo como $R = \frac{(O-E)^2}{E}$(que no es uno de los términos construidos del test χ^2 global). De esta forma, valores altos de R para los marginales de segundo orden identificarán pares de $x's$ que están mucho más o mucho menos asociados de lo que el modelo predice. En el caso binario Reiser

y VandenBerg [1994] y Bartholomew y col. [2008] utilizaban estos residuos para identificar parejas y tripletas de variables cuyo ajuste era deficiente. Recientemente estadísticos de bondad de ajuste basados en las probabilidades marginales observadas y esperadas del orden más bajo han sido propuestos, y se conocen con el nombre de tests de bondad de ajuste de información limitada(Reiser [1996],Bartholomew y Leung [2000],Cai y col. [2006],Maydeu-Olivares y Joe [2005]). En ellos, como hemos comentado, únicamente la información contenida en estadísticos que resumen los datos(normalmente los marginales de orden bajo de la tabla de contingencia) se utiliza para evaluar el modelo. Esto significa la agrupación de celdas a priori de una forma sistemática, de tal forma que los estadísticos resultantes tienen una distribución nula asintótica conocida. Estos métodos además son mucho más eficientes que los anteriores a efectos de cálculo. Maydeu-Olivares y Joe [2008] presentan un estudio sobre una visión de conjunto de este tipo de medidas en el análisis de datos categóricos con aplicaciones en la modelización de la IRT.

Es interesante el trabajo de Khalid [2009], en el cual se aborda desde diferentes perspectivas el ajuste de modelos de IRT, así como los trabajos de Sinharay y col. [2011] que aplican dos metodologías para evaluar la bondad de ajuste de los modelos de IRT(Análisis residual generalizado[Haberman, 2009] y el Análisis residual para evaluar el ajuste de los ítems[Bock y Haberman, 2009]) sobre diversos conjuntos de datos.

En investigaciones recientes, como las de Maydeu-Olivares [2013]; Maydeu-Olivares y Joe [2014]; Maydeu-Olivares y Montaño [2013] pueden encontrarse tanto una revisión de los métodos de bondad de ajuste de este tipo de modelos como el análisis del funcionamiento de estadísticos de ajuste en modelos de tipo Rasch y nuevas propuestas de indicadores que permitan, cuando el modelo se rechaza, evaluar la bondad de la aproximación, e incluso se proponen indicadores para analizar y detectar partes del modelo que se pueden mejorar.

7.6. Bondad de ajuste en la regresión logística.

La problemática de saber si un modelo de ajuste es razonablemente aceptable está muy extendida en la literatura referente a la regresión logística. No obstante vamos a presentar algunas cuestiones sobre las que es necesaria una reflexión para no incurrir en interpretaciones inconsistentes y erróneas.

Para contestar a esta cuestión podemos trabajar con la utilización de estadísticos que midan cómo de bien uno puede predecir la variable dependiente dadas un conjunto de variables independientes(se suelen llamar "medidas de poder predictivo"). Normalmente suelen estar entre 0 y 1, siendo el mayor valor el que está asociado a una predicción perfecta del fenómeno. El problema que se presenta es que no existe un consenso sobre el umbral que determine si el modelo es aceptable o no en base a esas medidas. Hablaremos de algunas de ellas en este apartado a continuación, como los pseudo-R^2, o el área bajo la curva conocida como Curva Característica de Operación (ROC).

Otro enfoque posible es el cálculo, para el modelo ajustado, de estadísticos de bondad de ajuste, como pueden ser la Deviance, la Chi-cuadrado de Pearson, o el test de Hosmer-Lemeshow. Estas medidas corresponden a test de hipótesis que contrastan que el modelo ajustado es correcto y que se traducen en la obtención de un p-valor, que si es suficientemente alto nos lleva a no rechazar que el ajuste es aceptable con nuestros datos de la muestra. Hay que resaltar que este tipo de enfoque no proporciona lo bien que uno puede predecir la variable dependiente, sino si complicando el modelo el ajuste podria ser mejor incluso que el ajuste actual, especialmente añadiendo términos no lineales, interacciones de las variables independientes o cambiando la función de enlace. Pero en el planteamiento que nos concierne, el espacio reducido a 2 ó 3 dimensiones mediante los métodos de ajuste nos hará disponer de las coordenadas de los individuos en dichas dimensiones y el propósito será construir modelos lineales en dichas componentes lo más sencillos

posibles, por tanto dichas medidas podrán ser utilizadas para nuestros propósitos.

En la regresión lineal se descompone la suma de cuadrados total de la variable dependiente en las partes explicadas por el modelo y residual para calcular el coeficiente de determinación y contrastar el modelo mediante un análisis de la varianza, pero en la regresión logística se utiliza lo que se llama Deviance, que es una medida de la falta de ajuste en modelos logísticos, y que es análoga a la suma de cuadrados de los residuales en la regresión simple. Se calcula comparando un modelo dado con el modelo saturado(el que ajusta de forma perfecta), que es lo que se conoce como test de razón de verosimilitud, es decir,

$$D = -2 \cdot ln \frac{L_{fitted}}{L_{saturated}}$$

Cuando el modelo en estudio contiene p parámetros y es ajustado a n observaciones binomiales, D sigue asintóticamente una distribución χ^2_{n-p-1} [Hosmer y Lemeshow, 2000]. En el caso del modelo logístico simple tenemos $n-2$ grados de libertad. Cuánto menor es el valor de D mejor es el ajuste, puesto que el modelo ajustado se acerca al saturado. Inversamente, un valor significativo de D según la distribución que sigue indica que existe un porcentaje muy alto de varianza que no se explica mediante el modelo elegido.

Se suelen utilizar dos valores de la Deviance en regresión logística, que son la Deviance nula y la Deviance ajustada, que corresponden a modelos con sólo el término independiente(modelo nulo) y el modelo que hemos ajustado, y que indican la diferencia entre ambos y el modelo saturado. De esta forma, el modelo nulo se correspondería con el umbral inferior a la hora de comparar modelos de predicción, y puesto que la Deviance es una medida de la diferencia entre un modelo dado y el saturado, para evaluar la contribución de un predictor a un conjunto de ellos, lo que se hace es sustraer la Deviance ajustada menos la nula, que seguirá una distribución $\chi^2_{s_1-s_2}$ con tantos grados de libertad como la diferencia entre el

número de parámetros estimados en ambos modelos, es decir,

$$\begin{aligned} D_{fitted} - D_{null} &= \left(-2 \cdot ln\frac{L_{fitted}}{L_{saturated}}\right) - \left(-2 \cdot ln\frac{L_{null}}{L_{saturated}}\right) = \\ &= -2 \cdot \left(ln\frac{L_{fitted}}{L_{saturated}} - ln\frac{L_{null}}{L_{saturated}}\right) = \\ &= -2 \cdot ln\left(\frac{L_{fitted}}{L_{null}}\right) \end{aligned} \tag{7.25}$$

por lo que si la deviance del modelo ajustado es mucho más pequeña que la del modelo nulo quiere decir que el conjunto de predictores mejoran el ajuste. Así es posible comparar no sólo el modelo ajustado con el nulo, sino también cualquier par de modelos analizando los cambios en la D al incluir (o excluir) términos en el modelo, comparando $G = D_{m_1} - D_{m_2}$ con los percentiles de la ji-cuadrado. El contraste es similar al que se utiliza en los modelos lineales para comparar las sumas de cuadrados explicadas.

Es posible plantear contrastes individuales de nulidad de los parámetros, teniendo en cuenta que el cociente entre el estimador y su error estándar tiende a una normal estándar. Este es el conocido como estadístico de Wald

$$W = \frac{\widehat{\beta}_i}{\widehat{S}_{\beta_i}} \equiv N(0,1)$$

que permite también calcular intervalos de confianza para los parámetros del modelo $\mathbf{I}_{\beta_i}^{1-\alpha} = \left[\widehat{\beta}_i \pm \widehat{S}_{\beta_i} \cdot z_{\frac{\alpha}{2}}\right]$.

No existe un acuerdo general sobre cuál es el mejor indicador equivalente al R^2 de la regresión lineal ordinaria cuando se calcula sobre regresiones logísticas. En los paquetes estadísticos más usuales aparecen los pseudo-R^2 propuestos por McFadden [1974] y por Cox y Snell [1989], aunque este último fue estudiado anteriormente por Maddala [1983] y por Cragg y Uhler [1970], cuyas expresiones son las siguientes:

$$R^2_{McFadden} = 1 - \frac{Ln(L_{fitted})}{Ln(L_{null})}$$

$$R^2_{CoxSnell} = 1 - \left(\frac{L_{null}}{L_{fitted}}\right)^{\frac{2}{n}}$$

siendo L_{null} el valor de la función de verosimilitud para un modelo que no tiene predictores, es decir, sólo el término independiente, y L_{fitted} dicho valor para el modelo que estamos tratando de estimar. Estos indicadores son índices alternativos de bondad de ajuste que de alguna forma guardan relación con el valor R^2 de la regresión lineal. El de Cox y Snell toma un valor máximo menor que 1, concretamente $1 - L_{null}^{\frac{2}{n}}$, lo cual hace que sea problemático, para lo cual se desarrolló su versión corregida, conocida como indicador de Nagelkerke, que varía entre 0 y 1 [Nagelkerke, 1991].

Existen otros indicadores, como el de Tjur [2009], cuyo umbral superior es 1 y que está relacionado con la definición de R^2 para los modelos lineales y es sencillo de calcular. Lo que se hace es calcular la media de las probabilidades predichas de un suceso para cada par de categorías de la variable dependiente. Entonces se toma el valor absoluto de la diferencia de estas dos medias. El razonamiento de esta medida se basa en que si un modelo produce buenas predicciones los casos asociados a los eventos deberían tener valores predichos elevados y viceversa. Tjur mostró que además su R^2 propuesto, que llamó *coeficiente de discriminación* es igual a la media aritmética de dos R^2 basados en los cuadrados de los residuos e igual a la media geométrica de otros dos R^2 basados en dichos cuadrados. El problema de este coeficiente es que no está basado en la maximización de la función de verosimilitud, y que no se puede generalizar fácilmente a las regresiones nominales u ordinales, cuestión que en el caso de McFadden y Cox-Snell es trivial.

En ocasiones se utiliza un índice relativo a la razón de verosimilitud, calculado como $R_L^2 = \frac{D_{null} - D_{fitted}}{D_{null}}$, que está entre 0 y 1 y tiene analogía con el valor calculado en la regresión lineal.

Es necesario tener en cuenta que la interpretación de los pseudo-R^2 debe ser cuidadosa, puesto que no representan la disminución del error de la misma forma que lo hacen los R^2 en la regresión lineal [Cohen y col., 2002], y de ahí el nombre de "pseudo". La regresión lineal supone homoscedasticidad del modelo, sin embargo la regresión logística siempre será heteroscedástica, puesto que la varianza de los

errores no son constantes. Estas medidas se basan en la proporción de deviance explicada por el modelo y son una extensión directa, como hemos comentado, de los cálculos basados en las sumas de cuadrados residuales en los modelos lineales.

Si suponemos que hemos estimado un modelo y queremos evaluar si proporciona un buen ajuste a la matriz de datos, hay que tener en cuenta que tanto la deviance[5] como la chi-cuadrado de Pearson [Pearson, 1900] tienen buenas propiedades cuando el numéro esperado de eventos para cada perfil es al menos 5. Pero si solo hay un caso en cada perfil ambos estadísticos tienen distribuciones alejadas de una chi-cuadrado, y por tanto arrojarían p-valores demasiado inexactos. De hecho, en este caso, la deviance no depende de los valores observados, lo cual hace que no pueda utilizarse como medida de bondad de ajuste [McCullagh, 1985].

Hosmer y Lemeshow [1980] propusieron agrupar los casos segun sus valores predichos por el modelo de regresión logística. Se reunen dichos valores de menor a mayor y se separan en varios grupos de aproximadamente igual tamaño, siendo 10 la recomendación usual. Para cada uno, se calcula el número observado de eventos y no-eventos e igualmente el número esperado[6] de ambos. Después se calcula la chi-cuadrado de Pearson para comparar los conteos esperados y observados, siendo los grados de libertad el número de grupos menos dos, de forma que si el p-valor es bajo se rechazará el modelo propuesto. No obstante, aunque este test parece ofrecer una solución al problema planteado anteriormente, presenta diversos problemas,

[5] $G^2 = 2\sum_j O_j log\left(\frac{O_j}{E_j}\right)$, donde cada j es una celda de una tabla de contingencia de 2 vías en la que cada fila es un perfil y cada columna es una de las dos categorías de la variable dependiente, O_j es la frecuencia observada y E_j es la frecuencia esperada según el modelo ajustado. Este estadístico, para muestras grandes sigue aproximadamente una distribución chi-cuadrado con $n - p$ grados de libertad, siendo n el número de grupos y p el número de parámetros del modelo incluída la constante, al igual que el estadístico de Pearson $X^2 = \sum_j \frac{(O_j - E_j)^2}{E_j}$

[6] El número esperado de eventos es la suma de las probabilidades predichas para todos los individuos del grupo, y el número esperado de no-eventos es el tamaño del grupo menos el número esperado de eventos.

como por ejemplo el número de grupos que se eligen, no habiendo una teoría consistente de elección de dicho número. Incluso uno podría pensar y esperar que añadiendo un término de interacción en el modelo o un término no-lineal que fueran significativos mejoraría el ajuste, pero este test no lo detecta, e igualmente, añadiendo un término de interacción no significativo mejora el ajuste según el test, lo cual ha hecho que numerosos autores hayan investigado en este campo, como Cox [1958], Tsiatis [1980], Brown [1982], Azzalini y col. [1989], le Cessie y van Houwelingen [1991, 1995], Su y Wei [1991], Osius y Rojek [1992] y Pigeon y Heyse [1999].

La mayoría de los test propuestos se basan en mecanismos diferentes de agrupación de los datos (Tsiatis [1980], Pigeon y Heyse [1991], Pulkstenis y Robinson [2002], Xie y col. [2008], Liu y col. [2012]). Una vez hecha la agrupación se calcula el estadístico de Pearson para evaluar la discrepancia entre los valores predichos y observados en los grupos. El problema principal de estos tests es que resulta muy costoso el proceso de agrupación y requiere una atención especial, la cual no siempre resulta cuidadosa por parte de los analistas, e incluso a veces es arbitraria.

le Cessie y van Houwelingen [1991], viendo estos problemas propusieron una clase de tests basados en residuos alisados, cuya motivación y uso provienen de la regresión no paramétrica, y trabajos posteriores como los de Hosmer y col. [1997] presentan versiones mejoradas de tests de ajuste y comparan el funcionamiento de diversos tests basados en dichos residuos respecto de los clásicos y algunos otros, como los de Royston [1992], que inicialmente se diseñaron para contrastar la desviación de un modelo respecto de la nonotonicidad de la función logit o la detección de una función logit cuadrática.

Con datos relativos de encuestas, en los que en numerosas ocasiones estos se recogen utilizando un muestreo complejo por conglomerados, ocurre que los elementos del mismo cluster suelen ser más homogéneos que los de otros clusters. Esto implica que existe una covarianza positiva entre las unidades del mismo cluster, por tanto, la correlación intra-clase (que mide la homogeneidad dentro de los

clusters) es generalmente positiva, y por tanto los métodos tradicionales de máxima verosimilitud no se pueden utilizar, resultando necesario el uso de lo que se conoce como pseudo-máxima verosimilitud[7] [Skinner y col., 1989]. Además de que existen métodos de estimación de parámetros en regresión logística que tienen en cuenta este tipo de muestreos complejos, trabajos como los de Archer y col. [2007] estudian diseños de test de bondad de ajuste en estos casos.

Hay que destacar también que en situaciones en las que el tamaño de muestra no es grande, o con matrices dispersas o datos muy sesgados el enfoque tradicional asintótico [Hosmer y Lemeshow, 2000] es inapropiado [Mehta y Patel, 1995]. En estas situaciones estos últimos autores recomiendan la estadística inferencial exacta y proporcionan una revisión y discusión de este planteamiento. En el caso de variables binarias Man-Lai [2001] desarrolla un test de bondad de ajuste en esta línea.

Otro enfoque para evaluar la bondad de ajuste se basa en los errores de predicción. Si suponemos que el modelo ajustado se utiliza para predecir el suceso "éxito" si la probabilidad ajustada es mayor que un valor π_0, digamos por ejemplo 0,5 y

[7] Conceptualmente la estimación por pseudo-máxima verosimilitud es como obtener los estimadores máximo verosímiles para el conjunto de datos elevado o expandido, es decir, el modelo logístico estaría siendo ajustado al censo de datos. La contribución de una observación a la pseudo-máxima verosimilitud sería $\pi(\mathbf{x}_{ji})^{w_{ji}\times y_{ji}}[1-\pi(\mathbf{x}_{ji})^{w_{ji}\times(1-j_{ji})}]$ y la función de pseudo-máxima verosimilitud se construye como el producto de cada una de estas contribuciones individuales, pero hay que tener en cuenta el número de clusters y de individuos en cada uno, es decir

$$l_p(\beta) = \prod_{j=1}^{m}\prod_{i=1}^{n_j}\pi(x_{ji})^{w_{ji}\times y_{ji}}[1-\pi(x_{ji})^{w_{ji}\times(1-j_{ji})}]$$

, calculándose los estimadores maximizando dicha función.

"fracaso" en otro caso, podríamos construir una tabla cruzada[8] con las respuestas observadas y predichas como instrumento para averiguar la proporción de clasificaciones correctas. Lo que ocurre es que un modelo puede ajustar razonablemente bien los datos y no predecir igual de bien, puesto que depende de si la variable respuesta es fácil o difícilmente predecible. Si el objetivo del análisis es la predicción este indicador podría ser un criterio ideal para comparar distintos modelos. Existe incluso un instrumento más informativo que la tabla de clasificación, puesto que resume el poder predictivo para cualquier valor posible de π_0, que es lo que

[8]Si la variable respuesta **y** es binaria y sólo puede tomar dos valores, "positivo"($y_i = 1$) o "negativo"($y_i = 0$), se llaman "Positivos verdaderos" a aquellos casos en los que el resultado de una predicción es $\hat{y}_i = 1$(si $\hat{\pi}_i > \pi_0$) y el valor real de la variable respuesta es $y_i = 1$.

$\hat{y}$ \ y	1	0	Total
1	Positivo verdadero	Positivo falso	P
0	Negativo falso	Negativo verdadero	N
Total	P	N	

De la misma forma se definirían los demás casos, derivándose de esta tabla los siguientes indicadores:

$$\text{Tasa de Positivos Verdaderos} = \frac{\text{Positivos correctamente clasificados}}{\text{Total de positivos}}$$

$$\text{Tasa de Falsos Positivos} = \frac{\text{Negativos clasificados incorrectamente}}{\text{Total de negativos}}$$

$$\text{Especificidad} = \frac{\text{Negativos verdaderos}}{\text{Positivos falsos} + \text{Negativos verdaderos}} = P(\hat{y} = 0|y = 0)$$

$$\text{Sensibilidad} = \frac{\text{Positivos verdaderos}}{\text{Negativos falsos} + \text{Positivos verdaderos}} = P(\hat{y} = 1|y = 1)$$

se conoce como curva ROC[9]. Esta curva es la gráfica de la sensibilidad vista como función de la tasa de falsos positivos, y su análisis proporciona herramientas para seleccionar los modelos posiblemente óptimos y descartar modelos subóptimos independientemente del coste de la distribución de las dos clases sobre las que se decide. Uno de sus primeros usos fue durante la segunda guerra mundial para el análisis de señales de radar.

La curva ROC resume graficamente el desempeño potencial de S como índice de riesgo y el área bajo la misma concentra esta información en un valor numérico que mide la calidad de S(Bamber [1975], Hanley y McNeil [1982] y Zweig y Campbell [1993]), de tal forma que valores cercanos a 1 indican un alto potencial discriminante en el índice de riesgo. Si $F_0(t) = F_{1(t)} \forall t \in \mathbb{R}$ entonces S no es informativo y el area bajo la curva es $\frac{1}{2}$.

Como hemos comentado, cuando tratamos de modelar un fenómeno es importante disponer de criterios para decidir qué tipo de modelo elegir, y ser capaces de seleccionar aquel modelo más satisfactorio, para lo cual es necesario disponer de indicadores que nos ayuden en esta labor de minimizar la pérdida de información al estimar. Es interesante comentar dos índices cuyo uso ha crecido significativamente, que son el Criterio de información de Akaike (AIC) y el Criterio de información bayesiano (BIC). El AIC fue propuesto por Akaike [1974] como un estimador

[9]Consideremos una población $\Omega = \Omega_0 \cup \Omega_1$, con $\Omega_0 \cap \Omega_1 = \varnothing$, y un vector $\mathbf{V}$: $\Omega \to \mathbb{R}^p$ mediante el cual observamos los elementos de Ω. Se define S: $\mathbb{R}^p \to \mathbb{R}$ como un indicador de riesgo, que resume la información de $\mathbf{v}$. De esta forma clasificamos ω en Ω_1 u Ω_0 si $s > \pi_0$ ó $s \geq \pi_0$ respectivamente, siendo π_0 un umbral de decisión calibrado por el usuario(pueden consultarse trabajos como los de López-Ratón y col. [2014], Vuk y Curk [2006] y Hajian-Tilaki [2013] para estudiar la problemática de esta elección). A partir de la sensibilidad y la especificidad se evalua la calidad de las reglas de clasificación [Hand, 1994]. Dado el indicador S se define la curva ROC como el conjunto $\{(u, v)|u = 1 - F_0(t)$ y $v = 1 - F_1(t); t \in \mathbb{R}\}$, donde $F_i(t) = Prob[S \leq t|\omega \in \Omega_i]$, de tal forma que el poder de discriminación de S depende de todas las combinaciones de sensibilidad y especificidad posibles.

insesgado asintótico de la información de Kullback-Leibler esperada [Kullback y Leibler, 1951], entre un modelo candidato ajustado y el verdadero modelo (pueden consultarse las relaciones entre ambos conceptos en Montesinos López [2011]). Este criterio se define como:

$$\text{AIC} = -2 \cdot (\text{log-verosimilitud}) + 2K = -2l(\hat{\beta}_K) + 2K$$

siendo K el número de parámetros estimados en el modelo, aunque pueden encontrarse formulaciones equivalentes o alternativas según el tipo de análisis o el tipo de muestras. En sí mismo, el valor del AIC no tiene significado y es cuando comparamos estos valores para un conjunto de modelos especificados a priori cuando resulta interesante, de manera que aquel con menor AIC será el mejor de ellos para el conjunto de datos de que disponemos.

No obstante la minimización del criterio de Akaike tiene algunos inconvenientes, puesto que no proporciona estimadores asintóticamente consistentes del modelo correcto. Si denotamos por $Modelo(K^*)$ al modelo correcto, entonces para cualquier $K > K^*$ tenemos que

$$Pr[AIC(K) < AIC(K^*)] = Pr[2\{l(\hat{\beta}_K) - l(\hat{\beta}_{K^*})\} > 2(K - K*)] \qquad (7.26)$$

siendo en este caso la variable aleatoria $2\{l(\hat{\beta}_K) - l(\hat{\beta}_{K^*})\}$ es el logaritmo de la razón de verosimilitud de dos posibles modelos que , bajo ciertas condiciones de regularidad se conoce que sigue una distribución $\chi^2_{K-K^*}$, y por tanto la probabilidad dada por 7.26 no es 0 asintóticamente. Para solventar este problema algunos autores sugieren multiplicar el término de penalización en el AIC por una función creciente de n, $a(n)$, de tal forma que la probabilidad

$$Pr[2\{l(\hat{\beta}_K) - l(\hat{\beta}_{K^*})\} > 2a(n)(K - K^*)] \underset{n}{\to} 0$$

Schwarz [1978] y Kashyap [1982] sugieren utilizar un enfoque bayesiano para el problema de la selección de modelos, que en el caso de muestras independientes e idénticamente distribuidas, resulta en un criterio que es similar al AIC el cual

está basado también en la utilización de la función logaritmo de la verosimilitud penalizada evaluada en el estimador máximo verosímil para el modelo en cuestión. El término de penalización en el BIC obtenido por Schwarz [1978] es el del AIC, K, multiplicado por la función $a(n) = \frac{1}{2}log(n)$, resultando que $BIC(K) = -2l(\hat{\beta}_K) + Klog(n)$.

Por último, decir que puesto que es necesario contemplar el problema de separación en regresión logística que hemos presentado anteriormente, al utilizar la regresión logística con un término de penalización(regresión ridge[10]) el cálculo de ambos criterios varía ligeramente, de forma que

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

siendo RSS la suma de cuadrados de los residuos del modelo y df los grados de libertad en un modelo de regresión ridge, que son $df_{\text{ridge}} = \sum \frac{\lambda_i}{\lambda_i+\lambda}$, con λ_i los valores propios de $\mathbf{X'X}$ (df es un función decreciente de λ, siendo $df = K$ con $\lambda = 0$ y $df = 0$ con $\lambda = \infty$).

[10]El estimador de regresión ridge $\hat{\beta}$ se define como el valor de β que minimiza

$$\sum_i (y_i - \mathbf{x}_i'\beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

Puede demostrarse que la solución a este problema es $\hat{\beta} = (\mathbf{X'X}+\lambda\mathbf{I})^{-1}\mathbf{X'y}$ y que si $\lambda \to 0$, $\hat{\beta}^{\text{ridge}} \to \hat{\beta}^{\text{OLS}}$ y si $\lambda \to \infty$, $\hat{\beta}^{\text{ridge}} \to \mathbf{0}$. $Var(\hat{\beta}) = \sigma^2\mathbf{WX'XW}$, con $\mathbf{W} = (\mathbf{X'X} + \lambda\mathbf{I})^{-1}$, y $Sesgo(\hat{\beta}) = -\lambda\mathbf{W}\beta$. También se cumple que la regresión ridge es un estimador lineal ($\hat{\mathbf{y}} = \mathbf{Hy}$), con $\mathbf{H}_{\text{ridge}} = \mathbf{X}(\mathbf{X'X} + \lambda\mathbf{I})^{-1}\mathbf{X'}$, luego $\text{tr}(\mathbf{H}_{\text{ridge}}) = \text{df} =$ "grados de libertad"). Pueden consultarse más detalles en Hoerl y Kennard [1970].

7.7. Cálculo de los parámetros para variables de tipo nominal y ordinal

El algoritmo alternado descrito en Vicente-Villardón y col. [2006] se puede utilizar y extender reemplazando las regresiones logísticas binarias por regresiones logísticas nominales u ordinales en la etapa M, teniendo en cuenta que es necesario penalizarlas, de acuerdo con lo comentado en la sección 7.4 para resolver el posible problema de la separación y atendiendo a las particularidades que ya comentábamos al inicio de la sección 5.1.5.

La notación que se ha seguido a lo largo del trabajo es equivalente en ambos casos, diferenciandose en la estructura de la matriz de parámetros de las variables. Vamos a detallar las etapas del proceso de estimación siguiendo la notación del capítulo 6, aunque para el caso nominal es completamente similar.

Tenemos que la función de verosimilitud es:

$$M\left(\mathbf{P}|\,\mathbf{d},\mathbf{A},\mathbf{B}\right)=\prod_{i=1}^{I}\prod_{j=1}^{J}\prod_{k=1}^{K_j}\pi_{ij(k)}^{p_{ij(k)}},$$

donde $p_{ij(k)}=1$ si el individuo i elige la categoría k de la variable j y $p_{ij(k)}=0$ en otro caso. Por tanto su logaritmo se escribe:

$$L\left(\mathbf{P}|\,\mathbf{d},\mathbf{A},\mathbf{B}\right)=\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K_j}p_{ij(k)}\log\left(\pi_{ij(k)}\right). \tag{7.27}$$

Si los parámetros $\mathbf{A}$ de los individuos fueran conocidos, el logaritmo de la verosimilitud se podría separar en J partes, una para cada variable:

$$L\left(\mathbf{P}|\,\mathbf{d},\mathbf{B}\right)=\sum_{j=1}^{J}L_j(\mathbf{P}|\,\mathbf{d}_j,\mathbf{b}_j)=\sum_{j=1}^{J}\left(\sum_{i=1}^{I}\sum_{k=1}^{K_j}p_{ij(k)}\log\left(\pi_{ij(k)}\right)\right), \tag{7.28}$$

siendo $\mathbf{d}_j$ y $\mathbf{b}_j$ las submatrices de parámetros para la j-ésima variable. Maximizar dicho logaritmo de la verosimilitud es equivalente a maximizar cada parte, es decir, obtener los parámetros para cada variable de forma separada. Maximizar cada L_j

es equivalente a llevar a cabo una regresión logística ordinal utilizando la j-ésima columna de $\mathbf{X}$ como variable respuesta y las columnas de $\mathbf{A}$ como variables predictoras. No describimos este tipo de regresión logística porque es suficientemente conocida. Las apreciaciones que hacíamos en cuanto al problema de la separación en el caso nominal siguen siendo válidas en este caso, por lo que en lugar de maximizar $L_j(\mathbf{P}|\,\mathbf{d}_j, \mathbf{b}_j)$ maximizamos:

$$L_j(\mathbf{P}|\,\mathbf{d}_j, \mathbf{b}_j) - \lambda \left(\|\mathbf{d}_j\|^2 + \|\mathbf{b}_j\|^2 \right). \tag{7.29}$$

con lo que cambiando los valores de λ obtenemos soluciones ligeramente distintas que no están afectadas por problemas de separación.

Por otra parte, si los parámetros de las variables fueran conocidos, el logaritmo de la verosimilitud se podría separar en I partes, una para cada individuo:

$$L\left(\mathbf{P}|\,\mathbf{A}\right) = \sum_{i=1}^{I} L_i(\mathbf{P}|\,\mathbf{a}_i) = \sum_{i=1}^{I} \left(\sum_{j=1}^{J} \sum_{k=1}^{K_j} p_{ij(k)} \log\left(\pi_{ij(k)}\right) \right).$$

La maximización de cada parte se podría resolver mediante el algoritmo de Newton-Raphson, pero en lugar de esto utilizaremos estimadores a posteriori esperados para los marcadores de los individuos. Para cada individuo(o patrón de respuesta) $\mathbf{p}_i$, la verosimilitud es:

$$M\left(\mathbf{P} = \mathbf{p}_i|\,\mathbf{d}, \mathbf{A}, \mathbf{B}\right) = \prod_{j=1}^{J} \prod_{k=1}^{K_j} \pi_{ij(k)}^{p_{ij(k)}}.$$

Si suponemos una distribución $g(\mathbf{a})$ (por ejemplo, una normal multivariante), la distribución marginal es:

$$P(\mathbf{P} = \mathbf{p}_i|\mathbf{d}, \mathbf{B}) = \int M\left(\mathbf{P} = \mathbf{p}_i|\,\mathbf{d}, \mathbf{A} = \mathbf{a}, \mathbf{B}\right) g(\mathbf{a}) d\mathbf{a},$$

y la verosimilitud observada:

$$M(\mathbf{P}|\mathbf{d}, \mathbf{B}) = \prod_{i=1}^{I} \left[\int M(\mathbf{P} = \mathbf{p}_i|\mathbf{d}, \mathbf{A} = \mathbf{a}, \mathbf{B}) g(\mathbf{a}) d\mathbf{a} \right].$$

Aproximaremos la integral con una cuadratura de Gauss-Hermite S-dimensional:

$$\tilde{P}(\mathbf{P} = \mathbf{p}_i|\mathbf{d}, \mathbf{B}) = \sum_{qS=1}^{Q} \ldots \sum_{q1=1}^{Q} M\left(\mathbf{P} = \mathbf{p}_i|\,\mathbf{d}, \mathbf{Y} = y, \mathbf{B}\right) \psi(y_{q1}) \ldots \psi(y_{qS}). \tag{7.30}$$

La cuadratura multivariante S-dimensional(figura 7.2), $\mathbf{Y}$, se ha obtenido como

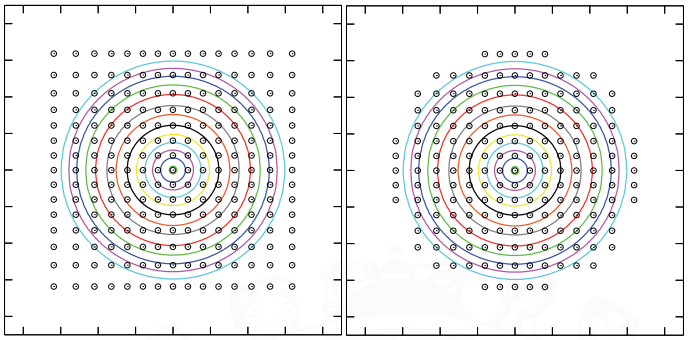


Figura 7.2: Cuadratura Gaussiana multivariante(2 dimensiones) podada[Jäckel, 2005] utilizada para optimizar el tiempo de cálculo.

el producto de S cuadraturas unidimensionales $(y_1, \ldots, y_Q)$ con Q nodos cada una; $\{\psi(y_q) : q = 1, \cdots, Q\}$ son los pesos asociados en la cuadratura e $y_{q1\cdots qS} \overset{not}{=} y$ representa cada punto de la cuadratura multidimensional. De esta forma, la puntuación a posteriori esperada para cada individuo se puede aproximar por:

$$E(\mathbf{a}|\mathbf{p}_i) \cong \frac{\sum_{qS=1}^{Q} \cdots \sum_{q1=1}^{Q} y_{q1\cdots qS} P(\mathbf{P} = \mathbf{p}_i|\mathbf{d}, \mathbf{Y} = y_{q1\cdots qS}, \mathbf{B})\psi(y_{q1}) \ldots \psi(y_{qS})}{\tilde{P}(\mathbf{P} = \mathbf{p}_i|\mathbf{d}, \mathbf{B})}, \tag{7.31}$$

siendo $y_{q1\cdots qS}$ cada punto de la cuadratura, y $\tilde{P}(\mathbf{P} = \mathbf{p}_i|\mathbf{d}, \mathbf{B})$ dado por (7.30). La habilidad del individuo i tiene S componentes (tantas como dimensiones del espacio reducido, es decir, $(\mathbf{a}_i = (a_{i1}, \cdots, a_{iS}))$, y cada una de ellas $\{a_{is}, s = 1, \cdots, S\}$ se aproximará por la expresión 7.31.

De esta forma el procedimiento de estimación se resume en un algoritmo iterativo que se repite hasta que el cambio en la verosimilitud es menor que una cantidad prefijada, el cual comienza con valores iniciales para los nodos de las cuadraturas, así como para los parámetros de los individuos(habilidades), con los

cuales se estiman los parámetros de las variables en una primera etapa. Después, dentro del procedimiento de iteraciones, en la etapa "E", utilizando las cuadraturas y los parámetros de las variables, se computan de nuevo las habilidades mediante la fórmula 7.31, finalizando cada iteración con la etapa "M", en la cual se estiman para cada variable de nuevo los parámetros mediante regresiones logísticas penalizadas nominales u ordinales según la matriz de datos inicial.

Capítulo 8

El paquete de R NominalLogisticBiplot para conjuntos de datos nominales.

When you run the marathon, you run against the distance, not against the other runners and not against the time.
– Haile Gebreselassie

8.1. Visión general del paquete

El paquete de R NominalLogisticBiplot tiene como principal objetivo el cálculo y representación de un biplot para un conjunto de variables categóricas de tipo nominal siguiendo la metodología descrita en el ca-

pítulo 5.1. Esto se lleva a cabo principalmente en dos etapas: primeramente es necesario calcular un objeto con la información de la estimación de los parámetros del modelo así como indicadores de la bondad de ajuste para posteriormente dibujar el biplot teniendo en cuenta las teselaciones calculadas a partir del objeto anterior.

La tabla 8.1 resume el conjunto de funciones que han sido programadas en el paquete, y que son públicas, junto con los conjuntos de datos disponibles para testear dichas funciones. Una de estas matrices de datos, que tiene el nombre de PhD_nomCyL, se utilizará posteriormente(ver sección 8.4) para demostrar cómo utilizar las funcionalidades de la herramienta software.

Cuadro 8.1: Resumen de rutinas del paquete NominalLogisticBiplot.

Funciones principales	Funciones auxiliares	Matrices de datos
NominalLogisticBiplot	polylogist	PhD_nomCyL
summary	RidgeMultinomialRegression	HairColor
plot	hermquad	Env
plotNominalVariable	multiquad	
	Nominal2Binary	
	NominalMatrix2Binary	
	NominalLogBiplotEM	
	NominalDistances	
	PCoA	
	afc	
	Generators	

8.2. Descripción de las principales funciones, clases y métodos

El paquete NominalLogisticBiplot se ha construido en torno a la rutina principal llamada NominalLogisticBiplot() y se ha estructurado en métodos, funciones y clases. Desde el punto de vista del usuario estos elementos se organizan de la siguiente forma:

Paso 1: Como primer paso, una función con el mismo nombre que el paquete, NominalLogisticBiplot construye el objeto con los cálculos, que incluye las coordenadas de las filas, y para cada variable del conjunto de datos los indicadores básicos que indican las bondades del ajuste(fundamentalmente varios pseudo-R^2 y porcentaje de clasificaciones correctas), que es una clase del tipo "nominal.logistic.biplot".

La sintaxis de la llamada a la misma es:

```
NominalLogisticBiplot(datanom,sFormula=NULL,numFactors=2,
        method="EM",rotation="varimax",metfsco="EAP",
        nnodos = 10,tol = 1e-04, maxiter = 100,
        penalization = 0.1,cte=TRUE,initial=1,alfa=1,
        show=FALSE)
```

A esta función se le pasa el conjunto de datos como primer parámetro datanom, y utiliza por defecto en el argumento method un algoritmo EM para calcular las coordenadas de las filas y columnas, que como hemos detallado, se ha modificado convenientemente para tener en cuenta el problema de la separación en regresión logística. Por este motivo es posible configurar algunos argumentos de la función para el algoritmo, como nnodos(número de nodos de la cuadratura de Gauss), tol(tolerancia que detendrá el proceso iterativo), maxiter(número máximo de iteraciones), cte(el modelo de regresión incluye una constan-

te. Por defecto el valor es TRUE) y penalization(valor utilizado en la matriz diagonal para evitar las singularidades). Los valores válidos para el parámetro method son "EM", "MIRT"(se utilizará el paquete mirt en los cálculos de las coordenadas de los individuos [Chalmers, 2012]), "ACM"(Análisis de Correspondencias Simples. El argumento alfa determina el peso de las filas y las columnas) y "PCOA"(Análisis de Coordenadas Principales). Por otro lado, los parámetros rotation y metfsco se deberían establecer si se ha elegido la opción MIRT para el parámetro method. El usuario puede elegir el procedimiento de cálculo de las habilidades iniciales para los individuos mediante el parámetro initial, el cual es necesario si se eligió como método de cálculo la opción EM, como describíamos en el procedimiento de estimación de parámetros, en el cual era necesario iniciar el mismo con unas habilidades iniciales. La información sobre la salida puede ser visualizada en la consola de R cambiando el valor del argumento show a verdadero, es decir, TRUE.

Paso 2: En un segundo paso el usuario podría utilizar el objeto de la clase mencionada anteriormente, que ha sido calculado para comprobar la información principal que en él hay almacenada, por medio del método summary, que mostrará los indicadores de bondad de ajuste, los porcentajes de clasificaciones correctas para cada regresión, el logaritmo de la verosimilitud, los coeficientes estimados, los p-valores, etc . . . cuya sintaxis es la siguiente:

```
summary(nlbo,completeEstim = FALSE,coorInd = FALSE, ...)
```

Dependiendo del conjunto de datos podría ser práctico tener la opción de mostrar las coordenadas de las filas o individuos o los coeficientes estimados, lo cual se puede configurar mediante los argumentos completeEstim y coorInd.

Paso 3: Finalmente se ha desarrollado un método para dibujar y mostrar el biplot con varias opciones de configuración que faciliten de una forma sencilla la customización y personalización de gráfico.

```
plot(nlbo,planex=1,planey=2,QuitNotPredicted=TRUE,
        ReestimateInFocusPlane=TRUE,proofMode=FALSE,
        AtLeastR2 = 0.01,xlimi=-1.5,xlimu=1.5,ylimi=-1.5,
        ylimu=1.5,linesVoronoi = FALSE,ShowAxis = TRUE,
        PlotVars = TRUE, PlotInd = TRUE, LabelVar = TRUE,
        LabelInd = TRUE,CexInd = NULL, CexVar = NULL,
        ColorInd = NULL,ColorVar = NULL,SmartLabels = FALSE,
        PchInd = NULL,PchVar = NULL,LabelValuesVar=NULL,
        ShowResults=FALSE,...)
```

Su utilización requiere pasar un objeto de la clase "nominal.logistic.biplot" (se almacenaría en el argumento nlbo), siendo el resto de los parámetros configurables y con valores por defecto. El usuario puede decidir el plano de representación(si la solución tiene 3 factores, uno podría estar interesado en estudiar el plano $\mathcal{P}_{12}$, es decir, el plano formado por el primer y segundo factor, el plano $\mathcal{P}_{13}$ o el plano $\mathcal{P}_{23}$), si las regiones(categorías) que no se predicen se deberían representar (QuitNotPredicted), si cada variable debería dibujarse en una ventana separada o todas conjuntamente(proofMode), los límites del gráfico, si las teselaciones para cada variable se deberían dibujar(linesVoronoi, que tiene valor por defecto FALSE y sólo se pintarán los CLP para una lectura más sencilla del gráfico), configurar el tamaño, color, posición de las etiquetas, etc .... El argumento ReestimateInFocusPlane permite reestimar los parámetros de las variables utilizando sólo las dimensiones elegidas en la representación del gráfico, o bien utilizar aquellos almacenados en el objeto que se pasa como primer parámetro

de la función.

Aunque este es el proceso usual para utilizar el paquete, el usuario podría añadir a un biplot ya construido una nueva variable categórica mediante la función plotNominalVariable. Esta función es muy útil para superponer variables categóricas en representaciones biplot clásicas con variables numéricas.

```
plotNominalVariable(nameVar,nominalVar,estimRows,planex = 1,
        planey = 2,xi=-3.5,xu=3.5,yi=-3.5,yu=3.5,CexVar=0.7,
        ColorVar="blue",PchVar=0.7,addToPlot=FALSE,
        QuitNotPredicted=TRUE,ShowResults=FALSE,
        linesVoronoi=TRUE,LabelVar=TRUE,tol = 1e-04,
        maxiter = 100,penalization = 0.1,showIter = FALSE)
```

El primer argumento (nameVar) guarda el nombre de la variable categórica, mientras que el segundo (nominalVar) almacena los valores de dicha variable y debe ser un factor. La estimación de las coordenadas de los individuos en el espacio reducido se pasa a la función mediante el parámetro estimRows. Si el usuario quiere añadir la representación a un gráfico ya existente puede hacerlo estableciendo a TRUE el argumento addToPlot. Esta función utiliza la regresión logística con penalización ridge y por eso aparece el argumento relativo a la penalización, que trata el problema de la separación. Por este motivo algunos de los parámetros son similares a los de la primera función descrita. El paquete proporciona otra función, similar a esta última, llamada plotNominalFittedVariable(). La diferencia principal es que esta rutina necesita la matriz de coeficientes estimados(mediante algún procedimiento externo), mientras que en plotNominalVariable() la estimación se hace dentro de la propia función mediante funciones auxiliares que describiremos más tarde.

8.3. Funciones auxiliares y otros detalles

Durante la implementación del paquete se ha considerado útil declarar como públicas algunas funciones que se utilizan en realidad de forma auxiliar. De esta forma se permite al usuario la flexibilidad de usar sus características en otros contextos puesto que los cálculos y procedimientos programados en ellas son generales.

Describimos de forma resumida estas rutinas: La función polylogist() ajusta una regresión logística politómica con corrección ridge de la variable nominal frente a las variables independientes, devolviendo un objeto de la clase "polylogist" que contiene los principales índices de bondad de ajuste. Por su parte, RidgeMultinomialRegression() calcula un objeto resultado del ajuste de una regresión logística multinomial de una variable nominal comparando este modelo con el modelo nulo, de tal forma que seremos capaces de decidir cuál de ellos ajusta mejor la variable dependiente. En esta función se utiliza polylogist() para estimar tanto el modelo completo como el nulo.

Las funciones hermquad() y multiquad() calculan los pesos de la cuadratura de Gauss-Hermite para un conjunto de puntos de una rejilla en una o más de una dimensión respectivamente. Devuelven un objeto de las clases "GaussQuadrature" y "MultiGaussQuadrature" con las coordenadas de los nodos y de los pesos asociados a cada nodo.

Nominal2Binary() transforma una variable nominal en una matriz binaria con tantas columnas como categorías(cada fila de la matriz tiene un valor de 1 para el nivel correspondiente de la categoría y 0 en otro caso) y NominalMatrix2Binary() construye la matriz indicadora para una matriz de variables nominales.

NominalLogBiplotEM() calcula, mediante un algoritmo alternado, los parámetros de las filas y columnas de un biplot logístico nominal de datos politómicos. Las coordenadas de las filas (E-step) se calculan utilizando cuadraturas de Gauss-Hermite multidimensionales y puntuaciones esperadas *a posteriori* (EAP), y los parámetros para cada variable (M-step) utilizando regresiones logísticas nominales

con corrección ridge de manera que cuando los puntos de diferentes categorías de una variable están completamente separados en el plano de representación y los métodos usuales no convergen este problema se solvente. Este problema de separación está presente en casi todos los conjuntos de datos para los que la bondad de ajuste de las variables es alta. Esta función utiliza una matriz de números con la información de las variables nominales como primero de sus parámetros y algunos argumentos ya descritos anteriormente y utiliza las funciones ya comentadas que trabajan tanto con las cuadraturas como con las regresiones.

PCoA() calcula el Análisis de Coordenadas Principales utilizando una matriz de distancias entre un conjunto de objetos devueltos por la función NominalDistances(), que computa las distancias hamming(o similaridades) entre individuos de una matriz de datos nominales. La función PCoA() se ejecutará si el argumento method de la función NominalLogisticBiplot() se fijó al valor "PCOA" indicando que este análisis sería el elegido para calcular las coordenadas de las filas.

afc() calcula para una matriz de datos el análisis de correspondencias simples. Se usa para computar las habilidades iniciales de los individuos cuando el usuario elige el valor 1 para el parámetro initial en la función NominalLogBiplotEM.

Finalmente, la función Generators(), con la matriz de parámetros estimados, resultantes del ajuste de un modelo logístico nominal sobre las coordenadas de las filas de una variable dada, calcula toda la información necesaria para dibujar la teselación resultante del ajuste. El argumento beta tiene tantas filas como número de categorías menos una y tres columnas(una para la constante y otras dos para las coordenadas x-y del plano en el que queremos representar la teselación). Esta función es llamada por las funciones de dibujo y devuelve un objeto de la clase "voronoiprob", el cual almacena las coordenadas de los puntos reales y virtuales, los vecinos de cada punto real, el número de puntos virtuales, los centros que resultan de invertir la teselación e información sobre las categorías ocultas de la variable. Se puede consultar la sección 5.1 para recordar los detalles computacionales del proceso.

El conjunto de datos llamado HairColor se ha extraído de Gower y col. [2011], y se corresponde con las características demográficas de 7 individuos almacenadas en cinco variables categóricas. Pueden resumirse de la siguiente forma:

- Sex(Sexo): Dos niveles o categorías (M=Hombre,F=Mujer)
- HairColor(Color de pelo): Cuatro categorías (Dark(Oscuro), Grey(Gris), Fair(Claro) y Brown(Marrón))
- Region: (E = England(Inglaterra),S = Scotland(Escocia), W = Wales(Gales))
- Work(Trabajo): (Manual,Clerical,Professional)
- Education(Educación): School(Escuela), Univ(Universidad), Postgrad(Postgraduado)

Por otra parte, la matriz de datos Env ha sido tomada de Gower y Hand [1996] y guarda las observaciones de cuatro variables observadas en 20 granjas de la isla holandesa de Terschelling. El conjunto de datos se presenta en Jongman y col. [1987] como parte de una encuesta más grande. Está relacionada con factores medioambientales y diferentes formas de gestión de las explotaciones. Hemos elegido estos datos porque han sido analizados previamente en la literatura y pueden servir para establecer comparaciones con los métodos que proponemos.

Las variables son las siguientes:

- Moisture class(Clase de Humedad), con 5 categorías, aunque la tercera de ellas no está presente en los datos. Los niveles se etiquetan con M1, M2, M4 y M5.
- Grassland management type(Tipo de Gestión de los pastizales), con 4 niveles (standard farming (SF), biological farming (BF), hobby farming (HF) y nature conservation management(NM))
- Grassland use(Utilización de los pastizales), que tiene 3 categorías: (production(U1), intermediate(U2) y grazing(U3))

- Manure class(Clase de estiércol), con 5 niveles etiquetados con C0, C1, C2, C3 y C4. Esta variable probablemente es ordinal puesto que dichos niveles asumen un nivel creciente de estiércol, pero la trataremos como variable categórica nominal.

8.4. Utilizando el paquete. Una aplicación a las personas doctoradas de Castilla-León.

Esta aplicación utilizará como fuente de datos la que se describió en el capítulo 5.2.2.3, que medía una serie de variables en individuos, residentes en Castilla y León, que habían obtenido el doctorado entre 1990 y 2006, y que eran menores de 70 años. La matriz de datos nominales con la que vamos a trabajar tiene finalmente 681 filas y 7 columnas.

Mostramos pues el proceso de cómo utilizar el paquete mediante la sintaxis en línea de comandos de la consola de R y en el orden lógico ante un conjunto de datos:

Etapa 1: Cargar el paquete NominalLogisticBiplot y el conjunto de datos de ejemplo PhD_nomCyL que hemos descrito anteriormente.

```
> library(NominalLogisticBiplot)
> data(PhD_nomCyL)
```

Etapa 2: Se construye el objeto de la clase "nominal.logistic.biplot" y elegimos una solución con 2 factores y que utilizará el algoritmo-EM modificado para calcular las coordenadas de las filas y las columnas. Establecemos el argumento penalization al valor 0.2 y el parámetro initial al valor 2 para comenzar el algoritmo con los valores proporcionados por el paquete mirt, con el objetivo de cambiar algún parámetro respecto de los valores por defecto.

```
> nlboPhD = NominalLogisticBiplot(PhD_nomCyL,numFactors=2,
+           method="EM",penalization=0.2,initial=2)
```

Etapa 3: Una vez que el objeto se ha creado es conveniente visualizar sus principales características para darnos una idea de la bondad de ajuste de cada variable. Esta información es crucial para no extraer conclusiones erróneas de algunas de ellas.

```
> summary(nlboPhD)

Nominal Logistic Biplot Estimation with Ridge Penalization 0.2
and logit link
n:  681
AIC:  6695.489
BIC:  7007.614

Goodness-of-Fit Statistics for the variables:
```

	MS	SECT	MIN	DES	PJREL	FOSAT	SOF
loglikelihood	-546.96	-254.45	-299.30	-238.93	-341.67	-753.09	-844.30
Deviance	1093.93	508.90	598.61	477.87	683.35	1506.19	1688.61
AIC	1111.93	526.90	616.61	495.87	695.35	1536.19	1712.61
BIC	1152.64	567.61	657.32	536.58	722.49	1604.04	1766.89
CoxSnell	0.03	0.74	0.70	0.77	0.64	0.69	0.34
Nagelkerke	0.04	0.85	0.80	0.78	0.74	0.69	0.34
McFaden	0.02	0.65	0.58	0.68	0.50	0.35	0.14
PCC	64.02	85.46	83.70	87.96	76.21	57.85	51.54

De acuerdo con la salida proporcionada por el método summary puede apreciarse que es aceptable el ajuste de las variables SECT, MIN, DES, PJREL y FOSAT, y al mismo tiempo el porcentaje de clasificaciones correctas es bastante alto, excepto para la última de las citadas. Por otra parte, las variables relacionadas con la fuente de financiación y el estado civil tienen una bondad de ajuste deficiente, por lo que lo que podamos decir de ellas deberá tomarse de una forma conservadora puesto

que existe variabilidad en ellas que no se está recogiendo convenientemente con esta solución bidimensional. Finalmente, respecto a la variable estado civil, a pesar de que su ajuste es muy pobre, puede observarse que el porcentaje de clasificaciones correctas es del 64 %, pese a lo cual esta variable no es válida para entender su contenido y relación con las demás en el espacio reducido propuesto.

Etapa 4: Por último, dibujamos el biplot. Elegimos en la opción proofMode TRUE, de tal forma que cada variable se dibujará en una ventana diferente, y linesVoronoi también TRUE para dibujar las teselaciones para las variables(ver Figura 5.16).

```
> plot(nlboPhD,proofMode=TRUE,LabelInd=FALSE,linesVoronoi = TRUE,
+           SmartLabels = FALSE,
+           PlotInd=TRUE,
+           CexInd = c(0.4),
+           PchInd = c(1),
+           ColorInd="azure3",
+           PlotVars=TRUE,LabelVar = TRUE,
+           PchVar = c(1,2,3,4,5,6,7,8,9),
+           ColorVar = c("red","black","maroon","blue","green",
+           "chocolate4","coral3","brown","brown2"))
```

Si nuestra elección para los argumentos proofMode y linesVoronoi hubiera sido FALSE, obtendríamos una representación gráfica similar al MCA porque se fusionarían todos los gráficos de las variables en uno sólo y se eliminarían las líneas divisorias de las teselaciones, es decir, sólo aparecerán los CLP, como puede comprobarse en la figura 8.2. Como se ha comentado, la interpretación ha de hacerse teniendo en cuenta el ajuste de cada variable.

Describimos ahora la última función principal del paquete, que es plotNominalVariable. Para ello, dado que la utilidad de la misma es añadir información proporcionada por una variable categórica nominal a un gráfico existente, haremos uso de una función del paquete OrdinalLogisticBiplot(ver sección 9). Aunque

la describiremos con más detalle posteriormente, la función BiplotDensity() dibuja para un conjunto de puntos un gráfico de densidad con las líneas de contorno, calculándose las densidades para cada uno de los grupos definidos por la variable nominal. En el ejemplo elegimos las variables SECT y DES de forma separada. Encima de estas representaciones se dibujarán las teselaciones para dichas variables utilizando la función plotNominalVariable.

```
> library(OrdinalLogisticBiplot)
> for(i in c(2,4)){
>         dev.new()
>         par(mai=c(0.2,0.2,0.2,0.2))
>         plot(nlboPhD$RowsCoords[, 1], nlboPhD$RowsCoords[, 2],
+           cex = 0,xlim=c(-1.5,1.5),ylim=c(-1.5,1.5),
+           axes=FALSE,xlab=NULL,ylab=NULL)
>         groupcolsA = c("white","black","red","black","brown","green")
>         BiplotDensity(as.matrix(nlboPhD$RowsCoords),
+           as.matrix(as.numeric(nlboPhD$dataSet$datanom[,i])),
+           img=TRUE,separate = FALSE,ColorType=4,
+           groupcols = groupcolsA[1:length(nlboPhD$dataSet$LevelNames[[i]])],
+           xliml=-2,xlimu=2,yliml=-2,ylimu=2)
>         nominalVar = as.factor(nlboPhD$dataSet$datanom[,i])
>         nameVar = dimnames(nlboPhD$dataSet$ColumNames)[i]
>         levels(nominalVar)<-nlboPhD$dataSet$LevelNames[[i]]
>         estimRows = nlboPhD$RowsCoords
>         plotNominalVariable(nameVar,nominalVar,estimRows,
+            planex = 1,planey = 2,xi=-3.5,xu=3.5,yi=-3.5,
+            yu=3.5,CexVar=1,ColorVar="white",PchVar=i,
+            addToPlot=TRUE,QuitNotPredicted=TRUE,ShowResults=TRUE,
+            linesVoronoi=TRUE,LabelVar=TRUE,tol = 1e-04,
+            maxiter = 100, penalization = 0.3,showIter = FALSE)
```

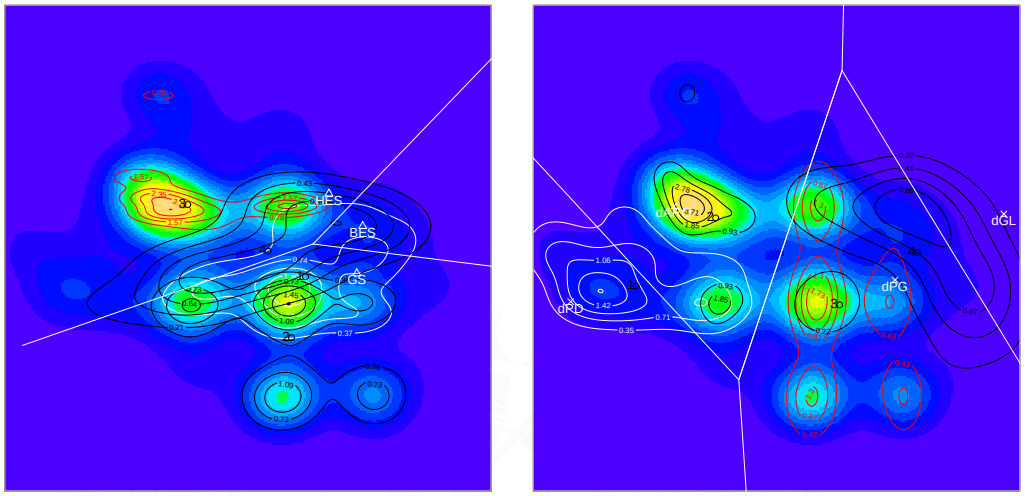
```
+}
>
```

La superposición de ambas representaciones(figura 8.1(a)) revela para la variable sector de empleo(SECT) que la categoría(PNP) aparece completamente mezclada e incluso oculta en la teselación dada por el método NLB, lo cual quiere decir que no se predice nunca.

Si analizamos el mismo tipo de gráfico para otra variable con una buena calidad de representación, por ejemplo, la variable "Nivel de formación deseado para el empleo principal el 1 de Diciembre de 2006", puede comprobarse que la teselación obtenida es congruente con lo que aparece en las configuraciones de densidad(Figura 8.1(b)). O dicho de otra forma, se puede ver cómo en cada región de la teselación están presentes las líneas de contorno de la categoría correspondiente a esa región. La superposición es inevitable cuando el tamaño de muestra es grande, en general, pero esta asociación aparece claramente. Cuando una categoría no se predice nunca en el plano considerado, las líneas de contorno asociadas a ella no se organizan de acuerdo con algun patrón concreto.

Es importante volver a resaltar que el NLB construye regiones de predicción de una forma diferente a como lo hacían Gower y Hand [1996]. Ellos comenzaban con un conjunto de puntos que llamaron "category points"($\mathcal{C}_k$) que se calculaban de un MCA con algunas modificaciones, y después se construían las regiones(teselaciones) a partir de estos puntos utilizando las distancias. Nosotros no tenemos los CLP y utilizamos las probabilidades en lugar de las distancias. Calculamos las regiones, no los puntos, y con la rama del conocimiento conocida como "Geometría Computacional" invertimos las teselaciones para obtener el conjunto de "puntos categoría" que nos permiten interpretar el biplot en términos de distancias.

Analizando el gráfico de la figura 8.2 puede constatarse que el sector Educación Superior(HES) engloba doctorados de las ciencias naturales, sociales e ingeniería, mientras que es menos probable la presencia de individuos doctorados de ciencias médicas. Dicho sector también aglutina empleos que se pueden desarrollar por



(a) Sector de Empleo

(b) Nivel de formación recomendable para el actual empleo.

Figura 8.1: Gráficos de densidad y teselaciones superpuestas obtenidas con la técnica del Biplot Logístico Nominal

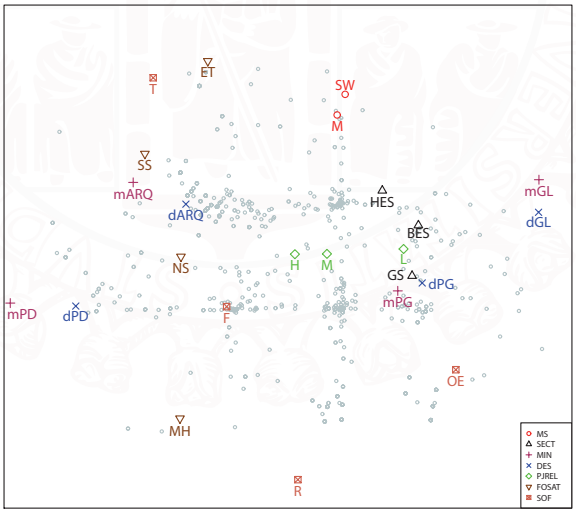


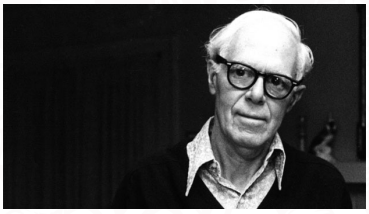
Figura 8.2: Biplot Logístico Nominal de los doctorados en Castilla-León

personal menos cualificado que al nivel de doctorado, aunque los doctorados más cualificados es más probable que acaben en el ámbito de la Educación Superior que

en cualquier otro. Los doctorados de sector empresarial se caracterizan porque su trabajo principal está poco relacionado con su grado de cualificación investigadora y con su rama de estudios de doctorado y para desempeñar su puesto de trabajo sería deseable ser postgraduado únicamente. Del sector de las "Instituciones Privadas sin Fines de Lucro", que está oculto en el biplot, no podemos decir nada con esta solución. Los doctorados que están empleados en el sector público pertenecen principalmente al campo de ciencias médicas y ciencias naturales y financian sus estudios utilizando otros tipos de empleos, lo cual hay que tomarlo con cautela puesto que la financiación no está correctamente representada.

Capítulo 9

Programación de una herramienta software en R para variables ordinales. El paquete OrdinalLogisticBiplot.



Mathemagical mathematics combines the beauty of mathematical structure with the entertainment value of a trick.

– Martin Gardner

Se ha desarrollado e implementado un paquete de R que contiene los procedimientos descritos en las secciones anteriores por Hernández y Vicente-Villardón [2014].

El Biplot Logístico Ordinal (OLB) extendía de alguna forma el Biplot Logístico Binario a los datos politómicos ordenados. Los individuos se representan como puntos sobre el plano y las variables como líneas en lugar de vectores, que es lo que aparece en los biplots clásicos o binarios, teniendo la geometría que hemos desarrollado en la sección 6.1.

La organización y la filosofía del paquete es similar a la del paquete NominalLogisticBiplot, en el sentido de que es necesario seguir los mismos pasos o etapas para construir el objeto con la información sobre la estimación, visualización del resumen de la misma y la representación gráfica del biplot, así como la función que permite añadir una variable ordinal a un biplot ya creado. Además, hay que decir que para aprovechar algunas de las funciones codificadas en el paquete relativo a variables nominales este nuevo paquete depende de aquel.

9.1. Rutinas del paquete

La estructura de las rutinas disponibles del paquete se presenta en la tabla 9.1. El paquete tiene en cuenta el problema de la separación en regresión logística de

Cuadro 9.1: Resumen del contenido del paquete OrdinalLogisticBiplot.

Funciones principales	Funciones auxiliares	Otras funciones	Datos
OrdinalLogisticBiplot	pordlogist	BiplotDensity	LevelSatPhD
summary	summary.pordlogist	plotOrdinalFittedVariable	
plot	CheckDataSet	PlotClusters	
plotOrdinalVariable	OrdinalLogBiplotEM		

tal forma que los algoritmos se han adaptado para tratar esta cuestión mediante un argumento que permite penalizar la función de verosimilitud.

La función principal de este paquete es OrdinalLogisticBiplot, la cual devuelve un objeto de la clase "ordinal.logistic.biplot" con varios componentes, como los

coeficientes estimados, los umbrales, las coordenadas para las filas en el espacio reducido, la verosimilitud, las puntuaciones factoriales de las variables en los ejes, las comunalidades, el ajuste de cada variable, las funciones de información de los ítems, etc ...

El método summary presenta la siguiente sintaxis:

```
summary(olbo,data = FALSE,rowCoords = FALSE,coefs = FALSE,
                loadCommun = FALSE,...)
```

Si el usuario desea visualizar los coeficientes estimados, las puntuaciones factoriales, las comunalidades, los datos o las coordenadas de las filas debe elegir los argumentos convenientes y ponerlos el valor de TRUE. Por otra parte, el método plot aparece de la siguiente forma:

```
plot(olbo,planex=1,planey=2,AtLeastR2 = 0.01,
      xlimi=-1.5,xlimu=1.5,ylimi=-1.5,ylimu=1.5,margin = 0,
      ShowAxis = TRUE, PlotVars = TRUE, PlotInd = TRUE,
      LabelVar = TRUE, LabelInd = TRUE,CexInd = NULL,
      CexVar = NULL, ColorInd = NULL, ColorVar = NULL,
      PchInd = NULL, PchVar = NULL,showIIC=FALSE,iicxi=-1.5,
      iicxu=1.5,legendPlot = FALSE,PlotClus = FALSE,
      Clusters=NULL,chulls = TRUE,centers = TRUE,
      colorCluster = NULL,ConfidentLevel=NULL,
      addToExistingPlot=FALSE,...)
```

Hay varios argumentos que podemos comentar. Primeramente, showIIC proporciona al usuario la posibilidad de dibujar las curvas de respuesta del ítem en ventanas separadas, utilizando los límites superior e inferior especificados en los parámetros iicxi y iicxu. Si se quiere utilizar una variable de clasificación para cada fila(calculada quizá por un procedimiento externo) , se puede especificar en el argumento Clusters. Relacionadas con esta opción están los parámetros PlotClus, chulls y centers. El primero de ellos indica si los clusters o grupos deben ser dibujados, el segundo es

un valor booleano para decidir si las convex hulls se construirán para cada cluster y el tercero dibuja el centro de cada cluster si su valor es TRUE. Para evitar valores extremos en la representación de las convex hulls, se puede fijar el porcentaje de los datos a tener en cuenta para ello mediante el parámetro ConfidentLevel. Si el valor es, por ejemplo, 0.95, sólo el 95 % de los datos se utilizará para calcular tanto los centros como las convex hulls.

Ambos métodos reciben un objeto del tipo "ordinal.logistic.biplot" que debe haber sido creado anteriormente.

CheckDataSet() comprueba si el conjunto de datos(que es su único argumento) es de tipo "data frame" y almacena los datos como una matriz de números enteros, con los nombres de las filas, columnas y niveles de las categorías en diferentes atributos formando un objeto de la clase "data.ordinal". Además elimina las filas que tengan algún valor NA y recodifica las variables con más categorías que valores diferentes presentan, manteniendo las etiquetas originales.

En uno de los ejemplos del biplot logístico nominal se utilizó la función BiplotDensity(), cuyo principal propósito es poner a disposición del usuario una utilidad que dibuje una estimación de la densidad de una muestra en 2D mediante la rutina kde2d del paquete MASS. La sintaxis es:

```
BiplotDensity(X, y = NULL, nlevels = max(y), grouplabels = 1:nlevels,
        ncontours = 6,groupcols = 1:nlevels, img = TRUE,
        separate = FALSE, ncolors=20, ColorType=4, xliml=-1,
        xlimu=1,yliml=-1,ylimu=1,plotInd = FALSE)
```

En X se especifican las coordenadas de los individuos en el plano mientras que y puede ser una variable categórica. Si no se especifica otro valor, la densidad se calcula para el conjunto completo de puntos sin importar las características de la población. Es posible configurar el número de líneas de contorno, los colores para las categorías de las variables, si la densidad debe dibujarse de forma separada de las líneas de contorno, etc. Puede ser útil en algunos casos complementar la interpretación con las teselaciones calculadas en los biplots categóricos.

pordlogist() lleva a cabo una regresión logística ordinal con penalización ridge, devolviendo un objeto que puede ser visualizado con la función summary.pordlogist().

Finalmente, la función PlotClusters() utiliza una variable categórica para representar grupos o clusters de individuos. Se pueden representar los centroides y las convex hulls para cada cluster. La estructura es:

```
PlotClusters(A, Groups = ones(c(nrow(A), 1)),colors = NULL,
        chulls = TRUE,centers = TRUE,ConfidentLevel=0.95)
```

El primer argumento almacena las coordenadas de cada punto y debe ser una matriz con dos columnas. El parámetro Groups es la variable de clasificación. También es posible elegir los colores para cada grupo y decidir si los centroides y las convex hulls deben dibujarse o no. El último parámetro tiene el mismo sentido del que ya se explicó anteriormente y evita representaciones en los que los valores extremos distorsionan las mismas.

9.2. Utilización del paquete con un conjunto de datos

El conjunto de datos proviene de la operación estadística que se citó en la sección 6.5 y el paquete OrdinalLogisticBiplot contiene una matriz en la que se han almacenado las contestaciones de 100 doctorados a nivel nacional.

Cada ítem se considerará como una variable ordinal, y aunque inicialmente teníamos 11 variables en total, por simplicidad en el manejo del paquete, el conjunto de datos LevelSatPhd sólo almacena cinco de ellas, que son las cinco primeras.

Veamos cómo utilizar el paquete:

Paso 1: Cargar el paquete OrdinalLogisticBiplot y el conjunto de datos LevelSatPhd.

```
> library(OrdinalLogisticBiplot)
> data(LevelSatPhd)
```

Paso 2: Se crea el objeto con la información del biplot logístico ordinal.

```
> olbo = OrdinalLogisticBiplot(LevelSatPhd,penalization=0.2)
```

Paso 3: Se puede consultar si se desea la principal información de dicho objeto.

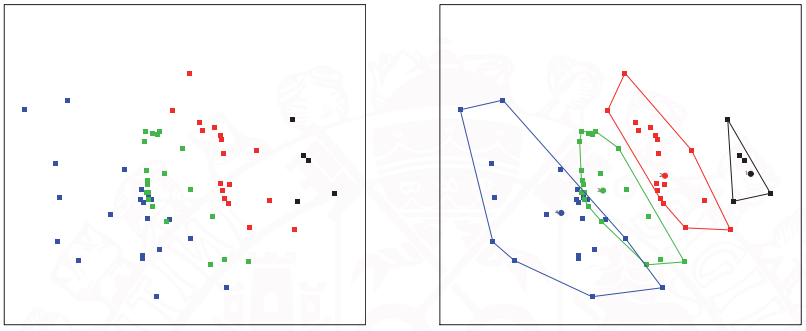
```
> summary(olbo,coefs=TRUE,loadCommun=TRUE)
```

```
 Ordinal Logistic Biplot Estimation  with Ridge Penalization : 0.2 ,
 EM algorithm and logit link

 Percentage of correct classifications,Pseudo R-squared measures and
 other indicators:
                    logLik  Deviance df p-value  PCC  CoxSnell
Salary           -55.08723 110.17446  2       0 0.84 0.7627785
Benefits         -39.68121  79.36241  2       0 0.91 0.9049595
Job Security     -88.26691 176.53383  2       0 0.66 0.5454943
Job Location     -39.50326  79.00652  2       0 0.87 0.8812088
Working conditions -48.99264  97.98528  2       0 0.80 0.8252842
                                        Macfaden Nagelkerke
                                       0.5663286  0.8280509
                                       0.7478218  0.9455991
                                       0.3087631  0.5915006
                                       0.7294717  0.9314210
                                       0.6403481  0.8832071

 Number of factors for the reduced solution:2

 Number of categories of the variables:
           Salary           Benefits       Job Security
                4                  4                  4
           Job Location         Working conditions
```

```
                4                          4

 Coefficients:
                       Dim 1      Dim 2
Salary             -4.575850 -0.6662892
Benefits           -5.934768 -3.0785921
Job Security       -2.502405  2.4957284
Job Location       -2.402148  6.2743167
Working conditions -3.798115  3.3327126

 Thresholds:
                          1          2        3
Salary             -3.3362796  0.8869201 4.477108
Benefits           -4.9387308 -0.1111747 3.261917
Job Security       -0.6953381  0.4770276 2.453456
Job Location       -1.7158768  1.8370239 4.214996
Working conditions -3.0710447  1.2625928 4.662199

 Factor Loadings and Communalities:
                         F_1        F_2 Communalities
Salary             -0.9672060 -0.1408348     0.9553219
Benefits           -0.8779089 -0.4554051     0.9781178
Job Security       -0.6813033  0.6794856     0.9258748
Job Location       -0.3536498  0.9237193     0.9783256
Working conditions -0.7373569  0.6470049     0.9623106
```

Utilizando el algoritmo alternado para estimar los parámetros de un modelo en dos dimensiones hemos obtenido los indicadores y las puntuaciones factoriales y comunalidades mostradas en la salida del método summary mencionado anteriormente. El porcentaje de clasificaciones correctas es muy alto para las variables

"Benefits(Beneficios)" y "Job Location(Ubicación laboral)", presentando valores del indicador *pseudo-*R^2 de Nagelkerke próximos a uno. Esto nos hace pensar que podría existir un problema de cuasi-separación entre las categorías de las variables, especialmente en la primera de ellas (ver figura 9.1), problema que ya sabemos que el método de estimación considera y trata correctamente.



(a) Coloración de los individuos según el valor que toman en la variable Beneficios

(b) Convex hulls de cada categoría de respuesta de la variable Beneficios

Figura 9.1: Problema de la cuasi-separación en la variable "Beneficios"

Si analizamos la interpretación de los factores utilizando las cargas factoriales, el primero presenta pesos más elevados para las variables "Salario", "Beneficios", "Seguridad laboral" y "Condiciones laborales", que en definitiva son características relacionadas con asuntos económicos, sociales y propios del puesto de trabajo. El segundo factor presenta altos valores para "Ubicación laboral" y tiene una importancia reseñable la "Seguridad laboral" y las "Condiciones laborales".

Paso 4: Dibujar el biplot logístico ordinal, utilizando el objeto calculado anteriormente.

```
> plot(olbo,margin=0.2,ColorVar=c("red","green","black","blue",
                "brown"),CexVar=c(0.7))
```

Si hubiéramos elegido el valor TRUE para el parámetro showIIC, podríamos ver el aspecto de las curvas de información del ítem y guardarlas una a una(figura 9.2).

En ella aparece lo que sería una parte de la salida en este caso.

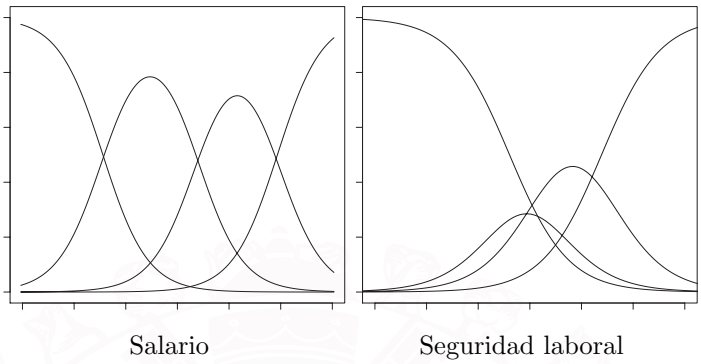


Figura 9.2: Curvas de respuesta de los ítems proyectadas sobre la dirección que mejor explica la variable para dos de ellas.

Hay que observar que en la variable relativa a la seguridad laboral la segunda categoría está oculta, al igual que ocurría en el ejemplo descrito en la sección 6.5, lo cual corrobora las impresiones que comentábamos en aquel caso para el total nacional.

El biplot de datos ordinales puede verse en la figura 9.3, en el cual los puntos de corte, para cada variable, y de cada una de las curvas y sus proyecciones en el espacio reducido, se corresponden con los puntos marcados en cada eje del biplot. En esta representación, el ángulo entre el eje principal y algunas variables, como por ejemplo, el salario y los beneficios es pequeño y por otra parte, para la ubicación laboral, es cercano a 90°, apareciendo una región en el cuarto cuadrante, alejada del origen, en la que los doctorados están muy satisfechos en todos los aspectos estudiados, mientras que en el segundo cuadrante existe una percepción negativa de las variables consideradas. Se podrían estudiar variables adicionales para detectar qué tipo de doctorados están en uno u otro grupo e incluso superponer variables nominales, por ejemplo, el sector de empleo, utilizando la función plotNominalVariable() si tuvieramos los valores medidos en estos doctorados para caracterizar al máximo estos grupos.

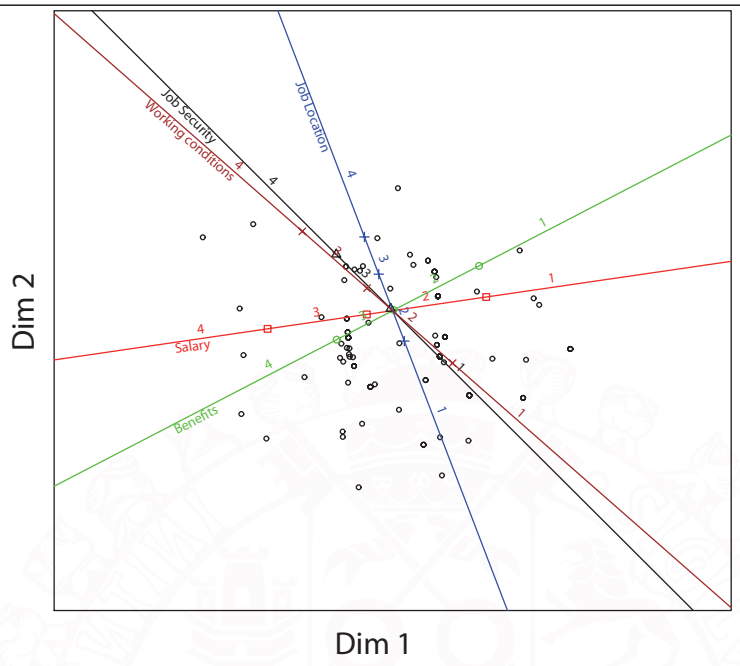


Figura 9.3: Biplot Logístico Ordinal para el conjunto de datos LevelSatPhd.

Algunas de las variables tienen un comportamiento similar, como el "Salario" o los "Beneficios", en las cuales los puntos que definen la posición de cada categoría se distribuyen de una forma similar y los ejes del biplot en ellas tienen una pendiente parecida. No obstante, aunque existen grupos de variables cuyas direcciones son similares, la posición de las categorías es muy diferente, como ocurre por ejemplo con la "Seguridad Laboral" y las "Condiciones del trabajo".

El gráfico de la figura 9.1 se puede obtener con la siguiente instrucción:

```
plot(olbo,PlotInd=TRUE,LabelInd = FALSE,
        xlimi=-2,xlimu=2,ylimi=-2,ylimu=2,margin = 0.1,
        CexInd=0.7,ColorInd=LevelSatPhd[,2],CexVar = c(0.7),
        iicxi=-7,iicxu=7,PchInd = 8,PlotClus = TRUE,
        Clusters=LevelSatPhd[,2],chulls = TRUE,
        centers = TRUE,colorCluster=LevelSatPhd[,2])
```

o también utilizando la función PlotClusters(), aunque tenemos que dibujar prime-

ro los puntos, después agruparlos de acuerdo con la variable que pasamos en el argumento Groups, y por último superponer el biplot logístico ordinal:

```
plot(olbo$RowCoords[, 1],olbo$RowCoords[, 2], col=LevelSatPhd[,2],
                cex = 0.7, pch=8, xlim=c(-2,2),ylim=c(-2,2))
PlotClusters(olbo$RowCoords, Groups = as.factor(LevelSatPhd[,2]),
                colors = LevelSatPhd[,2],chulls = TRUE,centers = TRUE,
                ConfidentLevel=NULL)
plot(olbo,margin = 0.2,ColorVar = c("red","green","black","blue","brown"),
                CexVar = c(0.7),addToExistingPlot=TRUE)
```

Capítulo 10

Biplot Logístico de Variables Categóricas

10.1. Modelos de IRT para variables observables de tipo mixto

Art doesn't transform. It just plain forms.
– Roy Lichtenstein

Hasta ahora habíamos supuesto que todas las variables observadas eran del mismo tipo, o bien nominales u ordinales con la misma distribución, Bernoulli para respuestas binarias o multinomiales para respuestas politómicas. En muchos campos de aplicación, especialmente cuando estudiamos datos provenientes de encuestas, ambos tipos de variables aparecen simultáneamente.

Incluso es muy frecuente que convivan también variables numéricas continuas, que no se pueden considerar como categóricas, cuya distribución condicional suele asumirse que es normal.

La forma tradicional de trabajar con problemas en los que están presentes variables de distintos tipos ha sido convirtiendo los mismos en problemas en los que las variables son del mismo tipo. Esto se consigue, por ejemplo, categorizando las variables continuas, de forma que el problema puede ser abordado desde el punto de vista de las metodologías de variables categóricas, aunque se tiene que pagar un cierto precio debido a la pérdida de información fruto de la agrupación de los elementos. Otro enfoque es la introducción de variables subyacentes, de tal forma que todas las variables categóricas se consideran como variables continuas observadas incompletas para abordar el problema como un modelo factorial lineal usual. Según el tipo de variable se estiman las correlaciones tetracóricas o policóricas para utilizarlas en estos modelos. Esta forma de proceder es debida a Muthén [1984] y se ha implementado en el programa Mplus(Muthén y Muthén [1984]). Una aproximación similar está disponible en el paquete LISREL(Jöreskog y Sörbom [2006]) dentro de un marco más amplio de modelización de ecuaciones estructurales.

Bartholomew [1987] analizó la reducción de la dimensionalidad mediante lo que se conoce como "principio de suficiencia[1]", mediante el cual toda la información que se necesita conocer sobre las variables latentes está contenida en un estadístico

[1]Si queremos construir un modelo con S variables latentes, lo mejor que podemos esperar es encontrar S funciones de $\mathbf{x}, X_1, X_2, \ldots, X_S$, de tal forma que la distribución condicional dada $\mathbf{x}$ no dependa de las variables latentes $\mathbf{y}$, situación en la que $\mathbf{x}$ sería suficiente minimal. La pregunta entonces es si hay una clase de distribuciones condicionadas $g_j(x_j|\mathbf{y})(j = 1, 2, \ldots, J)$ para las que existe un conjunto suficiente minimal. La respuesta es debida a Barankin y Maitra [1963], que establecieron las condiciones necesarias y suficientes para ello, las cuales requerían, bajo ciertas hipótesis de regularidad débiles, que al menos $J - S$ de las g_j debían pertenecer a la familia exponencial, es decir,

$$g_j(x_j|\mathbf{y}) = F_j(x_j)G_j(\mathbf{y})\sum_{i=1}^{S} u_{ji}(x_j)\phi_i(\mathbf{y})$$

suficiente, que en el caso de las variables binarias y continuas es lineal en las x's y se puede utilizar para construir medidas de escalamiento de individuos en el espacio reducido permitiendo asignar pesos diferentes a cada ítem[Bartholomew, 1984a,b]. Si denotamos por $\mathbf{y}$ al conjunto de variables latentes, cualquier modelo de IRT implica la especificación de $h(\mathbf{y})$, es decir, la distribución de dicho conjunto, así como la distribución condicionada $g(\mathbf{x}/\mathbf{y})$. En la estimación del modelo factorial el propósito es obtener S variables latentes a partir de las variables observadas, lo cual no es en absoluto sencillo en el caso general de que ambos conjuntos tengan distribuciones arbitrarias, pero Bartholomew y Tzamourani [1999] mostraron que existían estadísticos suficientes si las variables observables seguían una distribución perteneciente a la familia exponencial. Esta familia aglutina un conjunto de distribuciones diversas para variables continuas como la normal, la exponencial, la Erlang o algunas gammas, así como para discretas (Poisson o binomial) y tiene un extenso espectro de aplicaciones.

Dentro del marco de la IRT han sido muchos los autores que han estudiado modelos para variables binarias, politómicas y métricas, como Lawley y Maxwell [1971], Bock y Aitkin [1981] y Bartholomew y Tzamourani [1999]. El hecho de que las variables se puedan medir con escalas binarias, nominales, ordinales o escalas basadas en intervalos hace necesaria la especificación de diferentes distribuciones para ellas. Bartholomew y Tzamourani [1999] propusieron una estimación unifica-

, siendo $f(\mathbf{x}) = \int_{R_y} h(\mathbf{y})g(\mathbf{x}|\mathbf{y})d\mathbf{y}$ la función de densidad conjunta de las variables observables $\mathbf{x}$, $h(\mathbf{y})$ la distribución a priori de las variables latentes, R_y el espacio de variación de $\mathbf{y}$. Nuestro interés se centra en determinar qué conocemos sobre $\mathbf{y}$ una vez que hemos observado $\mathbf{x}$, que sería, suponiendo independencia condicional, según el teorema de Bayes:

$$h(\mathbf{y}|\mathbf{x}) = \frac{h(\mathbf{y})g(\mathbf{x}|\mathbf{y})}{f(\mathbf{x})} = \frac{h(\mathbf{y})\prod_{j=1}^{J} g_j(x_j|\mathbf{y})F_j(x_j)G_j(\mathbf{y})\sum_{i=1}^{S} X_j(x_j)\phi_i(\mathbf{y})}{\int h(\mathbf{y})\prod_{j=1}^{J} g_j(x_j|\mathbf{y}) = F_j(x_j)G_j(\mathbf{y})\sum_{i=1}^{S} X_j(x_j)\phi_i(\mathbf{y})d\mathbf{y}}$$

Se utiliza h para la distribución a priori y para la condicionada, aunque evidentemente son diferentes y es una cuestión de notación. Para encontrar $h(\mathbf{y}|\mathbf{x})$ es necesario conocer tanto h como g, pero lo que podemos estimar es f.

da para variables categóricas y métricas, que fue extendida por Moustaki [2000] dentro del marco de un modelo lineal generalizado que permitía simultáneamente analizar diversos tipos de variables. Además este trabajo tiene un enfoque diferente al de las variables subyacentes y va más allá del de Moustaki [1996] en el que se presentan modelos con dos variables latentes(cuyas distribuciones se supone que son normales) sobre variables observables de tipo mixto(binarias y métricas) estimados mediante el algoritmo EM.

Mellenbergh [1992] estudió la posibilidad de englobar los modelos de la IRT dentro del marco del Modelo Lineal General (GLIM), en el cual una función monótona de la respuesta esperada de un ítem se expresara como una función lineal de los factores latentes y de las variables observables explicativas. Pero no trató del problema de tener diversos tipos de distribuciones asociadas a las variables. Green [1996] analiza datos de tipo mixto dentro del marco del GLIM para modelos de variables latentes.

Existen un gran número de paquetes informáticos para modelar distintas situaciones. LISCOMP[Muthén, 1987] estima modelos desde la perspectiva de análisis estudiada por Muthén [1978, 1984] basada en tres etapas; MULTILOG[Thissen, 1991] y PARSCALE[Muraki y Bock, 1991], que pueden ajustar modelos para respuestas nominales, ordinales, o ítems de elección múltiple, pero sólo trabajan en una dimensión, o el ya comentado LISREL[Jöreskog y Sörbom, 1993], que se basa en el análisis de las correlaciones tetracóricas, policóricas y poliseriales y en el trabajo de Jöreskog [1990] en el que propone un método de mínimos cuadrados ponderados para la estimación de los parámetros estructurales que no requiere que las variables respuesta sean normales.

Autores como Sammel y col. [1997] trabajaron con modelos de respuesta latente con variables categóricas y métricas con efectos de covarianza tanto en las variables observadas como en las latentes dentro del marco del GLIM. No obstante, en él se trataba con variables binarias y normales en modelos unidimensionales. En el estudio de Moustaki [2000] se proponen modelos de medida que muestran el efecto

de un conjunto de variables latentes sobre variables observables cuya distribución pertenece a la familia exponencial, extendiendo por tanto lo anterior a distribuciones de variables observables politómicas, Poisson y gamma. Además el software LATENT [Moustaki, 1999] ajusta los modelos que estudian estos trabajos. Wedel y Wagner [2001] desarrollan y generalizan un poco más las aproximaciones anteriores puesto que proponen modelos multidimensionales con más de dos factores en los que las variables observadas y latentes pueden tener distribuciones diversas, pudiendo estas últimas ser distintas a la normal, para lo cual utilizan los avances en la teoría de la estimación de la verosimilitud simulada [Gourieroux y Monfort, 1997; McFadden, 1989; Stern, 1997].

El análisis estadístico de los modelos de IRT presentan una dificultad, consistente en que puesto que las variables latentes no se pueden observar, deben ser integradas fuera de la función de verosimilitud. Moustaki [1996] y Moustaki [2000] propusieron la utilización de una cuadratura de Gauss-Hermite simple como método de aproximación numérica, opción que cuando el número de dimensiones es mayor que 2 se hace a veces inviable. Una mejora sería utilizar una cuadratura adaptativa que centre y reescale apropiadamente los nodos para asegurar que se consigue el máximo, incluso con un número bastante menor de puntos [Rabe-Hesketh y col., 2002]. Esta técnica está disponible en STATA en la función *gllamm* que ajusta modelos mixtos de variables latentes generalizados. En trabajos como los de Huber y col. [2004] se propone como aproximación de la función de verosimilitud la aproximación de Laplace, que tiene la ventaja de que nos permite estimar modelos más complejos y modelos cuyas variables latentes están correlacionadas. Además es posible estimar las puntuaciones de los individuos en el espacio reducido de un modo directo así como la realización de contrastes de hipótesis sobre la base de las propiedades estadísticas del estimador obtenido.

Los modelos más populares de IRT son unidimensionales, y han predominado en el campo de las ciencias sociales y académico, dado que son más simples y tienen unas propiedades interesantes, como los modelos de Rasch, existiendo

paquetes de R[R Development Core Team, 2012] que pueden estimar modelos de Rasch, de respuesta latente general, 3-PL, modelos de respuesta graduada, como el ltm[Rizopoulos, 2006], u otros que estiman modelos de crédito parcial, como eRm[Mair y Hatzinger, 2007], pero sólo trabajan en el plano unidimensional.

Pero muchos constructos, sobre todo en el ámbito de la psicología son multidimensionales y es necesario un enfoque de este tipo. Esto implica una complicación evidente en el cálculo de los modelos, puesto que la estimación de los parámetros de las variables cuando aumenta el número de factores es dificil abordarlo con las técnicas de integración numéricas. No obstante, la investigación en la teoría de la estimación y la evolución tecnológica están haciendo que el análisis estadístico con esta herramienta metodológica sea factible[Edwards, 2010; Reckase, 2009].

Existen métodos de estimación basados en aproximaciones estocásticas que trabajan con restricciones lineales y datos faltantes, como la estimación bayesiana MCMC(Markov chain Monte Carlo), que utiliza el muestreo de Metropolis-Hastings[Hastings, 1970; Metropolis y col., 1953] o el de Gibbs[Casella y George, 1992], el algoritmo MH-MR(Metropolis-Hastings-Robbins-Monro)[Cai, 2010b] o el procedimiento de Monte Carlo EM [Yau y McGilchrist, 1996], aunque tienen inconvenientes como el tiempo de cálculo y las reglas de parada de los mismos. Recientemente se han estudiado soluciones a problemas en los que la modelización era multidimensional con un número de dimensiones alto mediante estos métodos de estimación, tanto para análisis exploratorio como confirmatorio[Edwards, 2010; Sheng, 2010], en el caso de modelos de respuesta dicotómica y politómica.

Dentro del marco multidimensional existen algunas opciones software relativamente actuales, como el paquete de R MCMCpack[Martin y col., 2011], que utiliza la teoría estocástica MCMC para estimar modelos multidimensionales de IRT con dos parámetros de forma robusta, pero requiere un conocimiento de la teoría bayesiana y computacionalmente es muy exigente, estando disponible solo para conjuntos de datos dicotómicos. Otra opción ya comentada es el paquete mirt[Chalmers, 2012], que implementa la estimación de una gran variedad de modelos de respuesta lo-

gística para datos tanto dicotómicos como politómicos, cuyas categorías pueden o no estar ordenadas y en un entorno de estimación multidimensional. Ofrece dos procedimientos de estimación, que son el algoritmo EM con cuadraturas fijas o el método bi-factor[Gibbons y col., 2007; Gibbons y Hedeker, 1992] para análisis exploratorio, y el método MH-RM para análisis confirmatorio de modelos politómicos[Cai, 2010a,b]. Por tanto, con este último paquete, por ejemplo, se podrían estimar modelos multidimensionales con variables categóricas de diversos tipos, pero no parece tener en cuenta el problema de la separación, al igual que otras herramientas software.

10.2. Estimación máximo verosímil.

Hemos detallado los procedimientos de estimación máximo verosímil cuando todas las variables eran del mismo tipo, por tanto ahora hay que presentar un marco de estimación válido para modelos de variables latentes lineales cuando las variables que se observan tienen diferentes distribuciones, y lo haremos en el caso de que estén dentro de la familia exponencial.

Suponemos que

$$g_j(x_j|\theta_j) = F_j(x_j)G_j(\theta_j)e^{\theta_j u_j(x_j)} \tag{10.1}$$

con $\theta_j = \alpha_{j0}y_0 + \alpha_{j1}y_1 + \cdots + \alpha_{jS}y_S$ $(j = 1, \ldots, J)$ e $y_0 \equiv 1$. Veamos como estimar el conjunto de parámetros αs, para lo cual si disponemos de una muestra de tamaño I, el logaritmo de la verosimilitud es

$$\begin{aligned} L &= \sum_{\ell=1}^{I} ln f(\mathbf{x}_\ell) \\ &= \sum_{\ell=1}^{I} ln \int h(\mathbf{y}) \prod_{j=1}^{J} F_j(x_{jl})G_j(\theta_j)e^{\sum_{j=1}^{J}\theta_j u_j(x_{j\ell})}d\mathbf{y} \end{aligned} \tag{10.2}$$

Las derivadas parciales de L respecto de αs vienen dadas por

$$\frac{\partial L}{\partial \alpha_{js}} = \sum_{\ell=1}^{I} \frac{1}{f(\mathbf{x}_\ell)} \int h(\mathbf{y}) \prod_{j=1}^{J} F_j(x_{j\ell}) \frac{\partial}{\partial \alpha_{js}} \left\{ e^{\left(\sum_{j=1}^{J}\theta_j u_j(x_{j\ell}) + lnG_j(\theta_j)\right)} \right\} d\mathbf{y}$$

con $(j = 1, \ldots, J; s = 1, \ldots, S)$. Las derivadas parciales que están en la integral se pueden calcular diferenciando respecto a θ_j de tal forma que

$$\left\{ u_j(x_{j\ell}) + \frac{dlnG_j(\theta_j)}{d\theta_j} \right\} e^{\left(\sum_{j=1}^{J} \theta_j u_j(x_{j\ell}) + lnG_j(\theta_j)\right)} \frac{\partial \theta_j}{\partial \alpha_{js}}$$

Por tanto

$$\frac{\partial L}{\partial \alpha_{js}} = \sum_{\ell=1}^{I} \int h(\mathbf{y}|\mathbf{x}_\ell) \left\{ u_j(x_{j\ell}) + \frac{dlnG_j(\theta_j)}{d\theta_j} \right\} y_j d\mathbf{y} \tag{10.3}$$

para $(j = 1, \ldots, J; s = 1, \ldots, S)$, donde $h(\mathbf{y}|\mathbf{x}_\ell)$ es la distribución a posteriori de $\mathbf{y}$ dado $\mathbf{x}_\ell$, $y_0 = 1$.

Si operamos intercambiando el sumatorio y la integral e igualando las parciales a cero se obtienen las ecuaciones de estimación básicas siguientes:

$$\int y_s \sum_{\ell=1}^{I} u_j(x_{j\ell}) h(\mathbf{y}|\mathbf{x}_\ell) d\mathbf{y} = \int y_s \frac{dlnG_j(\theta_j)}{d\theta_j} \sum_{\ell=1}^{I} h(\mathbf{y}|\mathbf{x}_\ell) d\mathbf{y} \tag{10.4}$$

para $(j = 1, \ldots, J; s = 1, \ldots, S)$, y siendo $h(\mathbf{y}|\mathbf{x}_\ell)$ la distribución a posteriori de $\mathbf{y}$ dado $\mathbf{x}_\ell$. Por tanto si existe el estimador máximo verosímil debe satisfacer estas ecuaciones.

10.3. Algoritmo EM genérico para resolver las ecuaciones máximo verosímiles.

Es posible reescribir las ecuaciones 10.4 y adaptarlas para resolverlas mediante el algoritmo EM. Denotemos por

$$n_{\mathbf{y}} = \sum_{\ell=1}^{I} h(\mathbf{y}|\mathbf{x}_\ell)$$

y por

$$r_{j\mathbf{y}} = \sum_{\ell=1}^{I} u_j(x_{j\ell}) h(\mathbf{y}|\mathbf{x}_\ell)$$

Si $\mathbf{y}$ tuviera una distribución discreta, $n_{\mathbf{y}}$ se podría interpretar como el número esperado de elementos de la muestra con valor $\mathbf{y}$, dados los datos. Y si $\mathbf{y}$ fuera

continua, sería la función de densidad de la esperanza, y en cualquier caso se podría pensar que especifica la distribución de los individuos en el espacio reducido.

Si $u_j(x_{j\ell})$ se considera como una puntuación propia del individuo ℓ sobre la variable j, entonces $\sum_{\ell=1}^{I} u_j(x_{j\ell})$ será la puntuación total asociada a la variable j. La cantidad $r_{j\mathbf{y}}$ representa cómo se distribuye esa puntuación total en el espacio reducido. De esta forma podemos escribir las ecuaciones como

$$\int y_s r_{j\mathbf{y}} d\mathbf{y} = -\int y_s n_{\mathbf{y}} \frac{d\ lnG_j(\theta_j)}{d\theta_j} d\mathbf{y} \quad (j = 1, \ldots, J; s = 1, \ldots, S) \tag{10.5}$$

El algoritmo EM por tanto sería así:

1. Elegir valores iniciales para $r_{j\mathbf{y}}$ y $n_{\mathbf{y}}$.
2. Resolver las ecuaciones 10.5 para α_{js}
3. Utilizar las estimaciones obtenidas para actualizar los valores de $r_{j\mathbf{y}}$ y $n_{\mathbf{y}}$
4. Repetir estos pasos 2 y 3 hasta que se consiga el criterio de parada elegido.

En la práctica, las integrales de la expresión 10.5 se aproximan con sumas en una rejilla de puntos sobre el espacio reducido. Y los valores iniciales para $r_{j\mathbf{y}}$ y $n_{\mathbf{y}}$ se obtienen especificando una versión discreta de $h(\mathbf{y}|\mathbf{x}_\ell)$.

10.4. Interpretación de las variables latentes y bondad de ajuste.

La interpretación está íntimamente relacionada con la capacidad de asignar nombres a las dimensiones del espacio reducido, es decir, de unirlas con conceptos teóricos del campo de estudio del cual disponemos datos. Una forma de conseguir este objetivo es por medio de la teoría clásica del análisis factorial utilizando las puntuaciones factoriales. Valores elevados de α_{js} significan que el factor s-ésimo tiene una influencia considerable sobre la variable j-ésima. De esta forma, identificando el conjunto de variables observables para las cuales y_j es importante

podríamos disponer de una interpretación a los ejes. A veces ocurre que una variable latente tiene influencia sobre todas las observables, en cuyo caso se denomina factor general. La idea, por tanto, de "buscar lo que es común" en un conjunto de variables es clave para poder interpretar. En la práctica, es extraño encontrar una estructura factorial inicial que determine perfectamente la interpretación de las dimensiones, pero puesto que las soluciones se pueden rotar[2] se podría buscar una configuración que fuera interpretable. Existen otras alternativas de interpretación en el modelo factorial lineal que son generales y aplicables en este contexto de modelos de variables latentes, y que pueden consultarse en Bartholomew [1987]. Cuando las variables de nuestro conjunto de datos son de diferentes tipos hay que tener cuidado puesto que la interpretación depende de la escala de las x's. Una variable con distribución de Bernoulli sólo toma valores 0 y 1, mientras que una normal varía de forma continua con una desviación estándar desconocida que hay que estimar. Por tanto, para "etiquetar" las variables latentes habrá que asegurar que las puntuaciones α's están calibradas para que puedan ser comparadas con sentido a través de las diferentes variables. Albanese [1990] sugirió una reparametrización de los parámetros de discriminación α_{js} que proporcionan resultados útiles, puesto que mejoran el comportamiento de la función de verosimilitud, que son

$$\alpha^*_{js} = \frac{\alpha_{js}}{\sqrt{\sum_{s=1}^{S} \alpha^2_{js} + 1}} \tag{10.6}$$

Pueden encontrarse ejemplos en los que se estudia esta problemática en Moustaki [1996]; Moustaki y Knott [2000]; Moustaki y Papageorgiou [2004].

[2]Si tenemos una matrix arbitraria $\mathbf{A}$ sabemos que la SVD proporciona la descomposición $\mathbf{A} = \mathbf{MDN}'$, siendo $\mathbf{D}$ diagonal con los valores propios. Por tanto la matriz $\tilde{\mathbf{A}} = \mathbf{MD}$ es una rotación ortogonal de $\mathbf{A}$, puesto que $\mathbf{A}$ se obtiene multiplicando por detrás de $\tilde{\mathbf{A}}$ la matriz ortogonal $\mathbf{N}'$. Además esta matriz cumple la propiedad de diagonalización porque $\tilde{\mathbf{A}}'\tilde{\mathbf{A}} = \mathbf{DM}'\mathbf{MD} = \mathbf{D}^2$, puesto que $\mathbf{M}'\mathbf{M} = \mathbf{I}$. Por tanto siempre se puede encontrar una rotación que hace ese producto diagonal para obtener una matriz de puntuaciones con una estructura simple.

En cuanto al estudio de la bondad de ajuste de estos modelos nos encontramos con el problema relativo a la exactitud de las aproximaciones calculadas si algunos de los parámetros toman valores extremos. En estos casos parece recomendable utilizar cálculos basados en la teoría bootstrap en lugar de las teorías asintóticas.

Hay que tener en cuenta que no existen métodos globales para medir la bondad de ajuste de los modelos mixtos. El estadístico razón de verosimilitud se utiliza de forma frecuente para contrastar la bondad de ajuste del modelo factorial continuo suponiendo que se fijan a priori el número de dimensiones latentes. En el caso de modelos latentes de datos binarios se usan tanto éste como el X^2 de Pearson, aunque es conocido que cuando los patrones de respuesta de las frecuencias esperadas son bajos, la aproximación χ^2 de la distribución de ambas medidas puede ser errónea [Reiser y VandenBerg, 1994]. Si existen variables continuas que sean normales, la cercanía de las distribuciones observadas y esperadas hay que evaluarla en base a las covarianzas. Para otro tipo de variables continuas habría que estudiar las distribuciones conjuntas para encontrar parámetros relevantes. Pero si las variables observables son de cualquier tipo(continuas y discretas) no es para nada obvio cómo medir esa bondad de ajuste y es necesaria una investigación más profunda, puesto que ninguno de estos test serían válidos. Se podrían comparar las frecuencias observadas y esperadas de las marginales de orden 2 ó 3 en el caso de variables categóricas y de forma similar comparar las matrices de correlación observadas y esperadas para cualesquiera variables normales continuas.

Si nuestra principal preocupación es la selección del modelo, entonces un indicador de bondad de ajuste no es el instrumento adecuado. En este caso, sería necesario un criterio de selección, como podrían ser el AIC, que tiene en cuenta el valor máximo de la verosimilitud y el número de parámetros estimados, o bien el BIC que además tiene en cuenta el tamaño de muestra. Ambos criterios se pueden aplicar sin importar el tipo de las variables, de tal forma que aquel modelo cuyo valor de estas medidas sea más pequeño será el más apropiado. En el trabajo de Sclove [1987] se analizan ambos indicadores con una comparativa.

10.5. Modelo conjunto de variables categóricas

Uno de los objetivos principales de esta investigación era la construcción de un biplot adecuado para variables categóricas sean éstas del tipo que sean. Se han analizado las geometrías y planteamientos de todos los tipos de variables categóricas, comenzando por las binarias, nominales y ordinales, adaptando los algoritmos de estimación convenientemente y sólo queda combinarlas todas para ser capaces de trabajar con conjuntos de datos en los que están presentes diferentes tipos de variables categóricas. Hasta ahora, en cada capítulo nos restringíamos a conjuntos de datos de un sólo tipo, y los paquetes de R respondían a estas situaciones concretas, pero ahora lo que haremos será extender la filosofía de trabajo a conjuntos genéricos.

La solución pasa por utilizar toda la infraestructura que hemos creado para adaptar el algoritmo de estimación y que sea capaz de estimar los parámetros correspondientes a cada variable y modelo según el tipo de cada una. Pero esto es sencillo utilizando un parámetro en el software que indique a las funciones pertinentes de qué tipo es cada variable, puesto que esa información se tiene a priori. De esta forma, el tratamiento diferenciado de cada una viene dado por las particularidades estudiadas anteriormente y que el algoritmo EM combinado de estimación considerará convenientemente en la etapa de maximización.

En la página web del departamento de estadística de la Universidad de Salamanca(http://biplot.usal.es) está colgado un paquete de R con el nombre BiplotForCategoricalVariables que permite estudiar conjuntos de datos categóricos con variables nominales y ordinales. Dicha herramienta utiliza los dos paquetes que se han detallado en las anteriores secciones, aprovechando el conocimiento de la geometría de los biplots para cada tipo de variable, de forma que en una misma representación puedan convivir ambos tipos de variables.

El paquete tiene públicas tres funciones para obtener la información sobre la

estimación y representación del biplot, que son las siguientes:

```
BiplotForCategoricalVariables(x,itemtype=NULL, dim = 2,
        nnodos = 10, tol = 1e-04, maxiter = 100,
        penalization=0.2, initial=1, alfa=1,
        Plot=FALSE, showResults=FALSE)
```

A esta función se le pasa el conjunto de datos en el primer parámetro, y en el segundo un vector que especifica de qué tipo es cada variable de dicho conjunto, de tal forma que si es una variable con valores desordenados la posición del vector correspondiente contendría el valor "nominal", y si los valores tienen un orden, "ordinal". Con este vector, el algoritmo EM que hemos implementado, para cada variable, es capaz de elegir, segun su tipo, el ajuste correspondiente. Si la variable es nominal utilizará la función polylogist() y si no pordlogist(). Esta función devuelve un objeto de la clase "mixed.logistic.biplot.EM" que contiene el resultado de la estimación de cada variable como resultado de la aplicación del algoritmo EM. Las variables nominales se agrupan en una lista y las ordinales en otra. También almacena la habilidad de los individuos e información adicional que contiene lo que se le pasa a la función como parámetros.

La función summary es similar a las de los otros dos paquetes, y permite de forma resumida analizar el resultado de la estimación global.

```
summary(object, summFitting = FALSE, coorInd = FALSE,
                nominalsFitting = FALSE, ordinalsFitting = FALSE)
```

El parámetro summFitting, si es TRUE, ordena a la función extraer un resumen de los indicadores de ajuste, tanto de variables nominales como ordinales, que nos orienta sobre la calidad del ajuste en términos generales.

Por último, la función plot() tiene una filosofía ya conocida en cuanto a la customización de las opciones del gráfico y el detalle de los diferentes parámetros puede consultarse en el apéndice I.

```
plot(x,planex=1,planey=2,QuitNotPredicted=TRUE,
        sepVarDifWindow=TRUE,xlimi=-1.5,xlimu=1.5,
        ylimi=-1.5,ylimu=1.5,margin=0.1,linesVoronoi = TRUE,
        ShowAxis = TRUE, PlotVars = TRUE, PlotInd = TRUE,
        LabelVar = TRUE, LabelInd = TRUE,CexInd = NULL,
        CexVar = NULL, ColorInd = NULL, ColorVar = NULL,
        SmartLabels = FALSE, PchInd = NULL, PchVar = NULL,
        ShowResults=TRUE, showIIC = FALSE, penalOrd=0.1,
        penalNom=0.1, tol = 1e-04, maxiter = 100,
        iicxi=-1,iicxu=1,addToPlot=TRUE,legendBR=FALSE)
```

10.6. Uso del paquete con una matriz de datos.

La matriz de datos sobre la que vamos a trabajar(tabla 10.1) está extraída del capítulo dedicado al Análisis de Correspondencias Múltiples del libro "Understanding Biplots" [Gower y col., 2011] y resume la medición de 5 variables sobre 7 individuos.

El cálculo del objeto que contiene la estimación con este nuevo paquete, el resumen de dicho objeto y el dibujo del biplot se obtienen mediante las siguientes sentencias:

```
library(BiplotForCategoricalVariables)
HairGower = data(HairColor)
xHair = BiplotForCategoricalVariables(HairGower,
        itemtype=c('nominal','nominal','nominal','nominal',
        'ordinal'), dim = 2, nnodos = 10, tol = 0.0001,
        maxiter = 500,penalization = 0.05,initial=1,
        showResults=FALSE)
```

Cuadro 10.1: Matriz de datos con información de 7 individuos sobre 5 variables

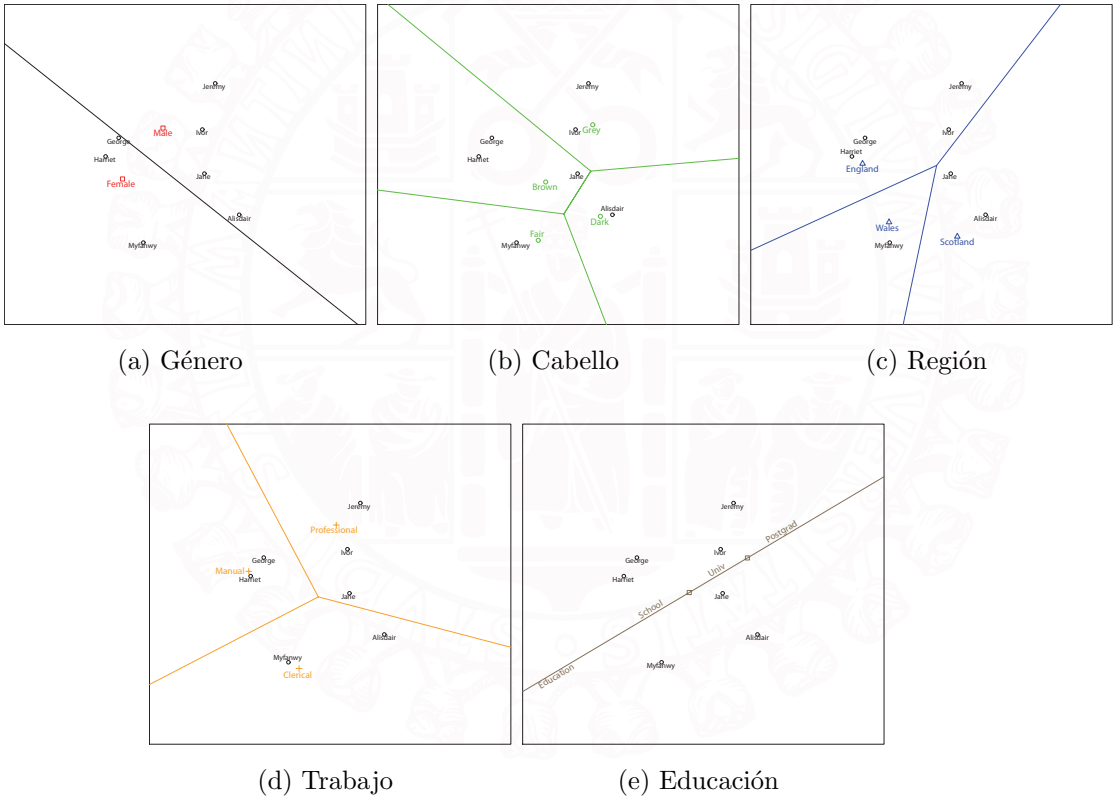
Caso	Sexo	Color de pelo	Región	Trabajo	Educación
1 George	M	Brown	England	Manual	School
2 Alisdair	M	Dark	Scotland	Clerical	University
3 Jane	F	Brown	Scotland	Professional	University
4 Ivor	M	Grey	Wales	Professional	University
5 Myfanwy	F	Fair	Wales	Clerical	School
6 Harriet	F	Brown	England	Manual	School
7 Jeremy	M	Grey	England	Professional	Postgrad

```
summary(xHair, summFitting = TRUE, coorInd = FALSE,
        nominalsFitting = FALSE, ordinalsFitting = FALSE)

ColorVar = c("red","green","blue","orange","burlywood4")

plot.mixed.logistic.biplot.EM(xHair,xlimi=-2,xlimu=2,
        ylimi=-2,ylimu=2,sepVarDifWindow=TRUE,
        PlotVars=TRUE,PlotInd=TRUE,LabelInd=TRUE,
        SmartLabels=FALSE,ColorInd="black",
        linesVoronoi = TRUE, QuitNotPredicted=TRUE,
        addToPlot=TRUE,ColorVar = ColorVar,ShowResults=FALSE)
```

El resultado en la consola de R tras ejecutar estas sentencias es:

```
Mixed Logistic Biplot Estimation with Ridge Penalization
0.05  and logit link
n:  7
```

```
AIC:  81.36529

BIC:  79.85078

                PCC Nagelkerke
Sex         71.42857  0.5201964
HairColor  100.00000  0.9043964
Region      85.71429  0.7643156
Work       100.00000  0.9214671
Education  100.00000  0.9282545
```

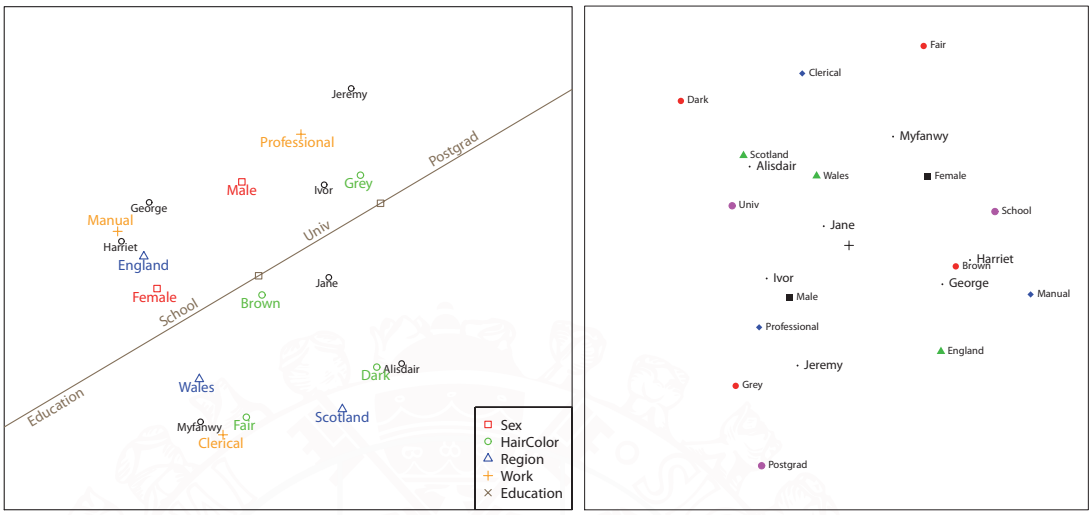


(a) Género (b) Cabello (c) Región

(d) Trabajo (e) Educación

Figura 10.1: Regiones de predicción del Biplot Logístico de Variables Categóricas. La variable Educación se ha considerado ordinal por lo que las regiones vienen determinadas por los puntos en la recta.

Las regiones de predicción resultantes, junto con los puntos característicos de



(a) Biplot Logístico Categórico (b) MCA biplot

Figura 10.2: Representaciones biplot del CLB y MCA(basado en la matriz de Burt) en un espacio reducido de dimensión 2.

cada categoría que se predicen, de este análisis aparecen en la figura 10.1. La variable Educación se representa por una recta con dos puntos característicos que delimitan las franjas de predicción entre las tres categorías. Si utilizamos la función plot con los argumentos linesVoronoi y sepVarDifWindow a FALSE, de la misma forma que hemos descrito anteriormente, obtenemos el gráfico simplificado de la figura 10.2(a).

Si representamos un MCA biplot basado en la matriz de Burt, como puede verse en la figura 10.2(b), la disposición de los puntos de las variables nominales que verdaderamente aportan información es congruente a la del Biplot Logístico de Variables categóricas. (CLB), pero llama la atención el tratamiento de la variable ordinal "Educación(Education)" y las regiones que presenta en este último si lo comparamos con la posición de los 3 niveles educativos en el gráfico del MCA. Las franjas de predicción en el primer caso engloban a los individuos que en el segundo están más cercanos a cada categoría. Como es conocido, todo lo que es

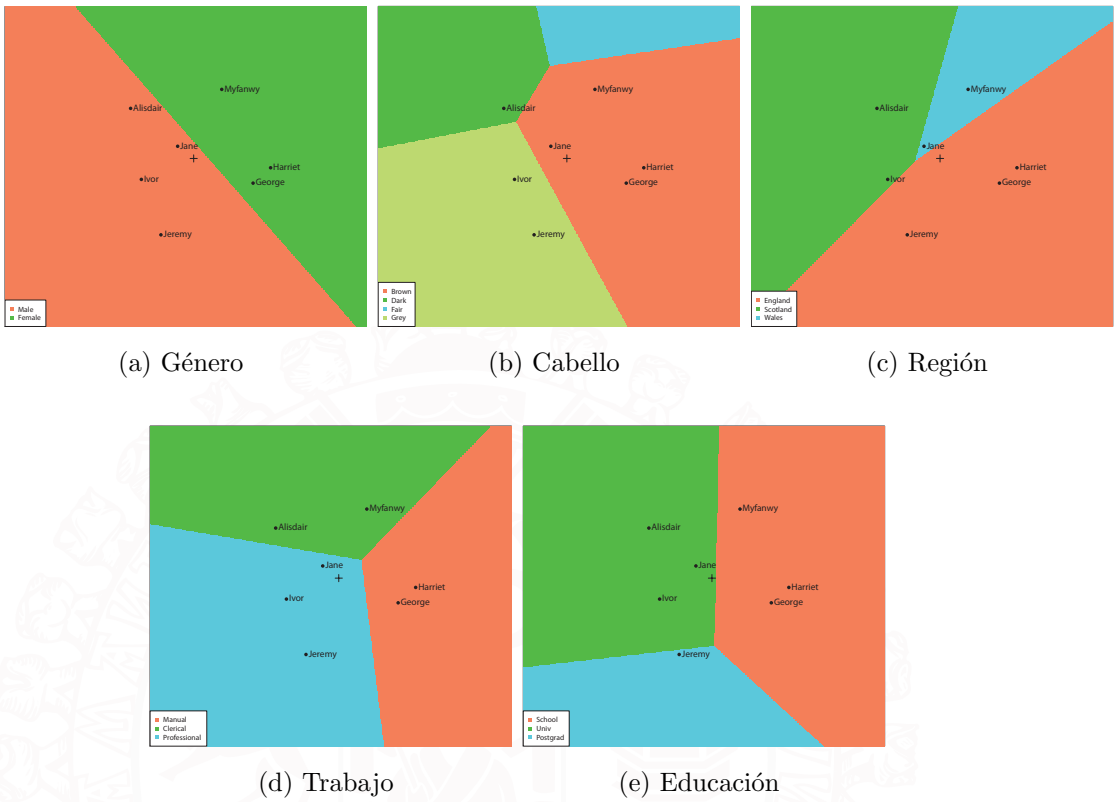
poco frecuente el MCA lo "sitúa" lejos del origen, como por ejemplo los puntos Dark, Fair, Clerical, Postgrad, siendo esta consideración totalmente ajena al CLB. Si analizamos el porcentaje de clasificaciones correctas en ambos casos tenemos que es ligeramente superior en el Biplot de Variables Categóricas, como muestra la tabla 10.2. Por otra parte, las regiones de predicción del MCA biplot aparecen calculadas en la figura 10.3, cuyo aspecto para las variables nominales es similar a las del CLB, aunque ligeramente menos precisas que las de esta técnica. Nuevamente, incluso en un caso sencillo como este, por ejemplo para la variable Región, los tres puntos característicos de las categorías no están situados en sus respectivas regiones, como ocurre con "Wales" que caería en la región de "Scotland"(viendo la figura 10.2(b) conjuntamente con la configuración relativa a la variable Región en la figura 10.3(c)).

Cuadro 10.2: Porcentaje de clasificaciones correctas con CLB y MCA biplot para el conjunto de datos HairColor

Variable	Género	Color de Pelo	Región	Trabajo	Formación
CLB	71.43	100	85.71	100	100
MCA biplot	71.43	85.71	71.43	100	100

10.7. Ilustración de las diferencias salariales por Género en la Universidad de Stellenbosch(Sudáfrica)

Vamos a estudiar el comportamiento del Biplot Logístico de variables categóricas(CLB) utilizando un conjunto de datos conocido que pretendía medir las desigualdades entre la remuneración de los hombres y las mujeres. Este campo ha sido estudiado por numerosos autores como Barbezat y Hughes [2005], Ward [2001]



(a) Género (b) Cabello (c) Región

(d) Trabajo (e) Educación

Figura 10.3: Regiones de predicción del MCA Biplot basado en la matriz de Burt normalizada.

y Warman y col. [2006].

Algunos investigadores atribuían las diferencias por género a factores como el historial de publicaciones [Ward, 2001], la distribución del tiempo [Toutkoushian, 1999], la categoría y antigüedad [McNabb y Wass, 1997], el tipo de institución [Barbezat y Hughes, 2005], la edad y la formación [Bayer y Astin, 1975], pero seguía existiendo una parte no explicada(ver Toutkoushian [1998];Barbezat y Hughes [2005]). Esta diferencia desconocida se conoce como "Brecha salarial en la remuneración por genero".

En lo que se refiere a la situación de Sudáfrica, tras 1994 las instituciones de educación superior se enfrentaron con la realidad de una transformación hacia una

cultura organizativa apropiada para lo que se llamó la nueva Sudáfrica. En este ejemplo, ya estudiado por Gower y col. [2011], el objetivo es analizar las diferencias de remuneración entre los géneros para el personal académico de la Universidad de Stellenbosch durante los años 2002-2005. Dicho personal examinado es personal fijo a tiempo completo.

El conjunto de datos está disponible en R con el nombre de Remuneration.data y tiene las siguientes columnas:

- **ID** Número de identificación.
- **Remun** Coste total del empleo sin deducciones en Diciembre de 2002 y 2005(en unidades de R10000). La inflación no se tuvo en cuenta porque afectaba de forma igual a todo el personal durante el periodo de estudio.
- **Resrch** Resultados de la investigación expresados en una escala numérica.
- **Rank** Posición académica o rango.
- **Age** En años.
- **Gender** Género(hombre o mujer).
- **AQual** Calificación académica expresada en una escala continua.
- **Faclty** Valor entero que representa una de las nueve aptitudes incluidas en el estudio.

Para trabajar con el ejemplo, puesto que las variables en las que estamos interesados deben ser categóricas, utilizaremos los datos del año 2002 tratándolas como nominales para un estudio del Análisis de Correspondencias Múltiples y como ordinales(excepto Gender y Faclty) para ilustrar brevemente cómo sería la representación de un Análisis de Componentes Principales Categórico. Por tanto, debemos agrupar los valores de las variables en intervalos para poder de alguna forma categorizarlas, quedando como se muestra en la tabla 10.3.

Cuadro 10.3: Variables categóricas del conjunto de datos Remuneration.cat.data.2002

Variable	Valores
Remun	Las categorías son los deciles de la variable original
	R1,R2,...,R10(valor más alto)
Resrch	Res0,Res1,...,Res7(más alto)
Rank	Rank1,Rank2,...,Rank5(más alto)
Age	A1,A2,...,A7(más viejo)
Gender	Male(Hombre),Female(Mujer)
AQual	Q1,Q2,...,Q9(más alto)
Faclty	F1,F2,...,F9(escala nominal)

Mediante la utilización del método propuesto para construir Biplots Logísticos de Variables Categóricas, en la figura 10.4 pueden observarse las regiones de predicción. Se han considerado como nominales "Gender" y "Faclty", siendo el resto ordinales, de modo que para dichas dos variables obtenemos teselaciones en las que se han situado los puntos resultado de la inversión de las mismas, y para las ordinales se presentan las rectas y sus puntos característicos. Se ha considerado una solución en dos dimensiones para analizar la información que cada método es capaz de captar. Las curvas características de los ítems de la figura 10.5 muestran que existen muchas categorías ocultas que no aportan información de los individuos en este plano, como por ejemplo con la variable Resrch, que sólo se predice en su valor máximo y mínimo. Igualmente ocurre con Rank, que presenta 3 de sus 5 valores en las predicciones. Respecto de las variables nominales, en Faclty, de las 9 categorías iniciales solo aparecen 5 en la teselación.

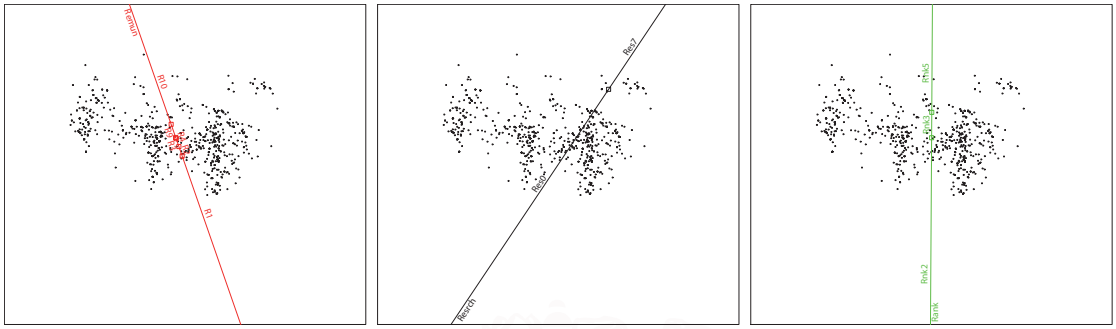
En la figura 10.6(d) la categoría de las mujeres(Female) correspondiente a la variable Género está completamente separada de la de los hombres(Male). Además, las mujeres aparecen más cercanas a las categorías de personal en los deciles

de retribuciones más bajas, el personal más joven, el rango más bajo, y con los resultados de investigación más pequeños. En contraposición a esta situación está la de los hombres claramente viendo la situación del centroide. Si se hubiera considerado una solución en 3 dimensiones se podría contrastar que esta tercera dimensión no aporta más información a las diferencias de género de la que aquí tenemos. Este planteamiento también se puede confirmar mediante el CLB, en el que las mujeres aparecen en el gráfico en una situación descrita de igual forma que antes(ver figura 10.6(a,b)).

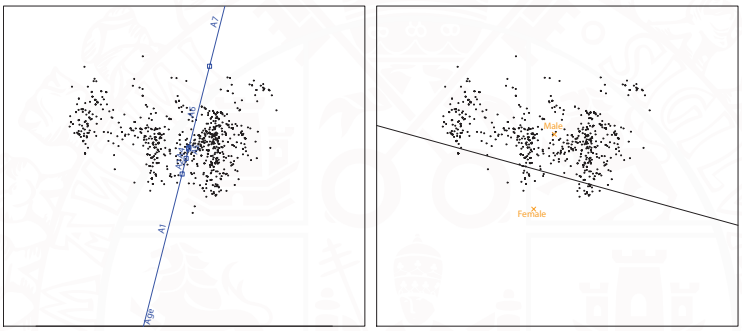
Por último, podemos mostrar en la figura 10.7 la representación obtenida mediante la técnica CATPCA, que se ha obtenido con el paquete de R Bbip disponible en el libro Gower y col. [2011]. A veces se le conoce con el nombre de Análisis de Componentes Principales no lineal, puesto que es un método de escalamiento óptimo, y digamos que es un método cuyo homólogo para variables continuas es el PCA. En cierto modo se podría entender como una técnica exploratoria de reducción de las dimensiones dada una matriz de datos en la que aparecen variables categóricas, nominales u ordinales. Esta técnica es capaz de mostrar las relaciones que se presentan entre las variables de partida, entre las filas y entre ambos. Lo que intenta este método es capturar las posibles no linealidades de las relaciones que están presentes en las variables estudiadas en las transformaciones de las mismas. A cada categoría de las variables originales le corresponde un valor que se llama cuantificación, y a cada variable transformada le corresponde un coeficiente con el que interviene en cada dimensión, de manera que se minimiza una función de pérdida en el sentido de mínimos cuadrados, encontrándose dichos valores mediante un algoritmo llamado mínimos cuadrados alternados [Gifi, 1990]. En este caso lo que interesa es analizar las variables de forma individual y no conjuntamente como en el análisis de homogeneidad. Pueden consultarse más detalles de este método en Gower y col. [2011]. En dicha figura, en la que la calibración de los ejes se hace según los CLP, puede verse, analizando las puntuaciones z que muchas categorías podrían juntarse, cuestión que ya veíamos con el biplot logístico, en términos de

categorías ocultas o no predichas, y además sabíamos cuales eran.

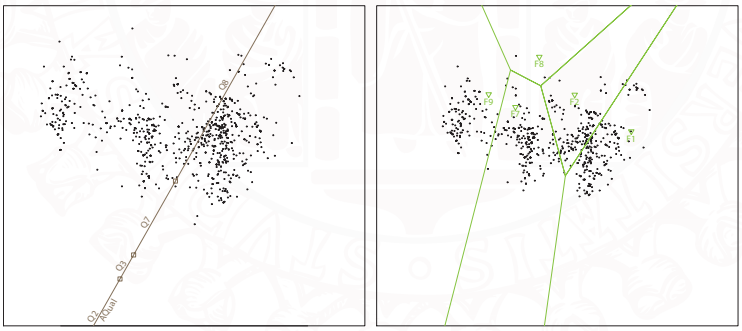
Puede comprobarse fácilmente en el gráfico que los hombres se sitúan en posiciones en las cuales el rango es más alto en comparación al de las mujeres, al mismo tiempo que presentan cualificaciones más altas, una mayor edad y sueldo, lo cual permite de alguna forma evidenciar lo que se podría llamar una brecha de género. Del gráfico del biplot logístico categórico se llegan a las mismas conclusiones (figura 10.6(a)), confirmando la validez de nuestra propuesta.



(a) Remun (b) Resrch (c) Rank

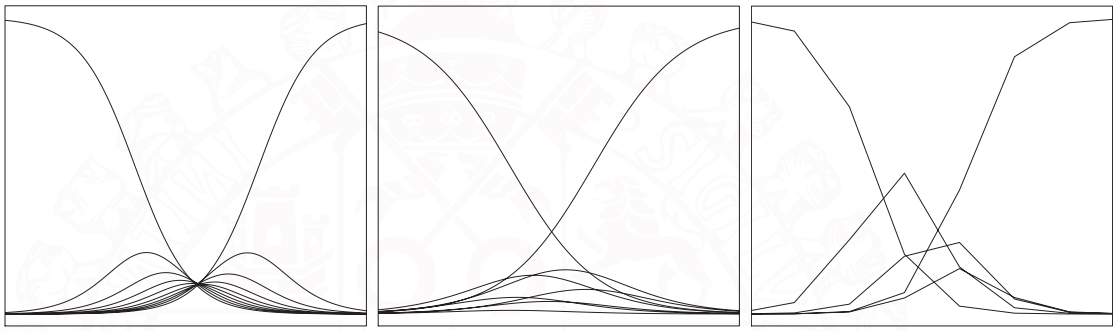


(d) Age (e) Gender

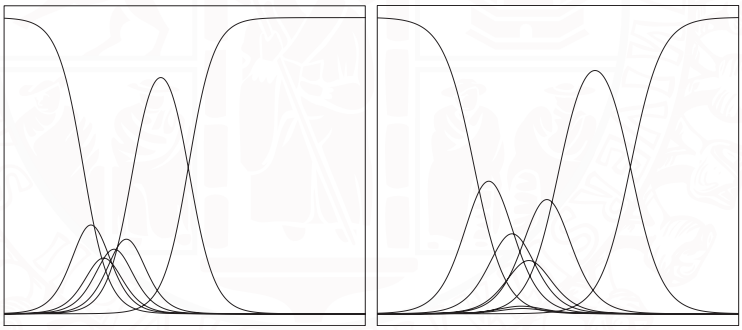


(d)AQual (e) Faclty

Figura 10.4: Regiones de predicción para las variables del conjunto de datos Remuneration.data realizando un Biplot Logístico Categórico en 2 dimensiones.

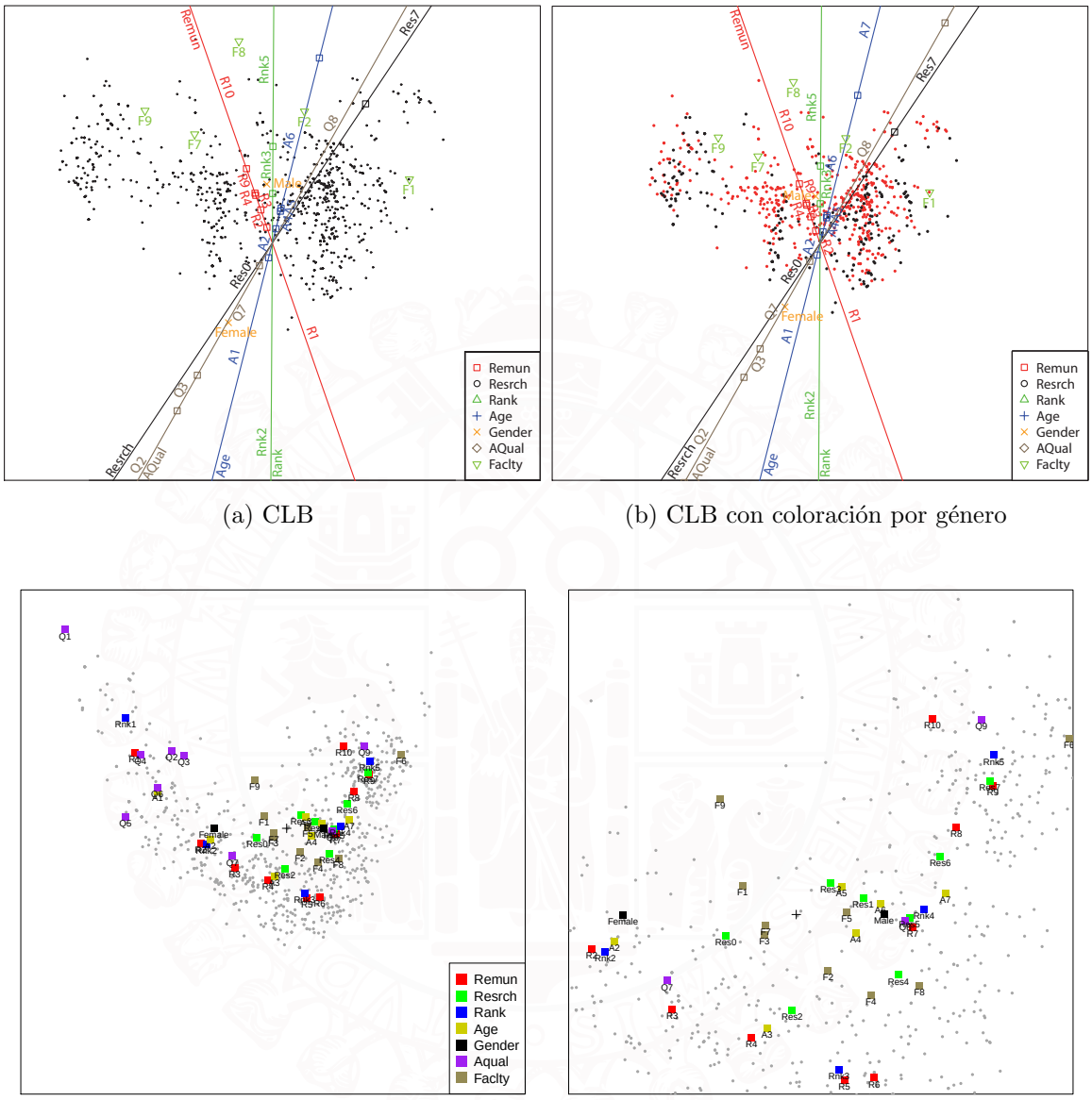


(a) Remun (b) Resrch (c) Rank



(d) Age (e) AQual

Figura 10.5: Curvas características de los ítems de variables ordinales.



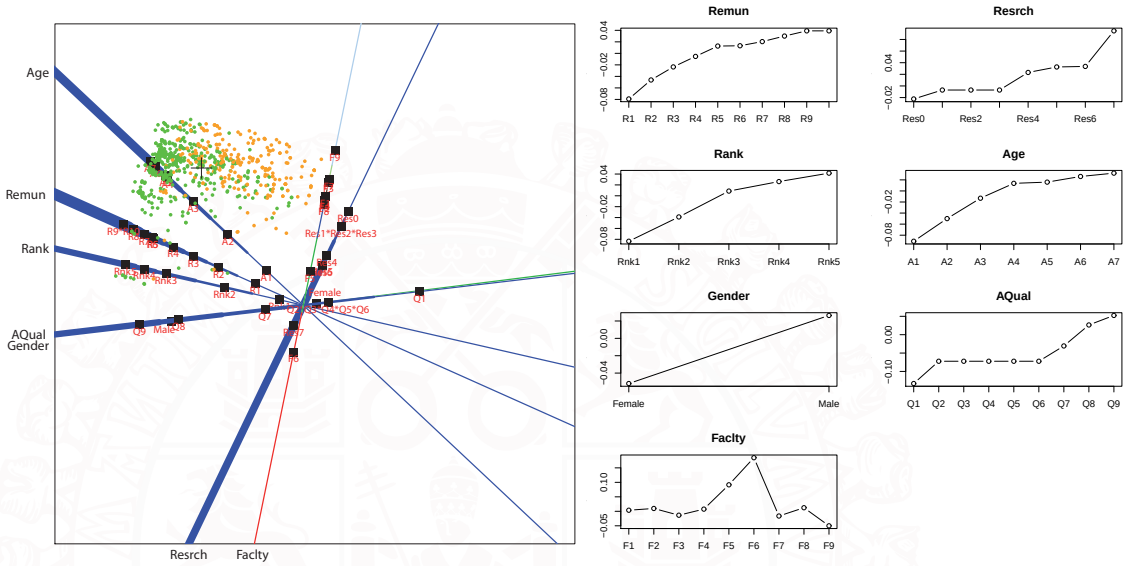
(a) CLB

(b) CLB con coloración por género

(c) MCA biplot

(d) zoom del MCA biplot

Figura 10.6: Biplot Logístico Categórico en 2 dimensiones para el conjunto de datos Remuneration.cat.data.2002(a,b) y Análisis de Correspondencias Múltiples(c,d) basado en la matriz indicadora. Los cuadrados coloreados corresponden a los centroides de cada categoría y los puntos grises son los individuos de la muestra.

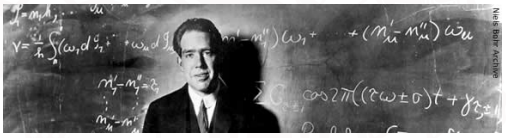


(a) CATPCA (b) Puntuaciones z para cada variable.

Figura 10.7: Análisis de Componentes Principales Categórico en el que se representan líneas que conectan los CLPs asociados a cada variable categórica. Todas las variables se considera que tienen valores ordenados excepto Faclty y Gender. Dichas líneas tienen una mayor anchura en el sentido correspondiente a su ordenación. Los puntos de la muestra se han coloreado según el Género(verdes para los hombres y naranja para las mujeres). Además se han movido los ejes para evitar la interferencia con los puntos de la muestra. Estos puntos se podrían representar como densidades, con elipses de concentración, etc ...

Capítulo 11

Discusión



Prediction is very difficult, especially about the future.

– Niels Bohr

El recorrido descrito hasta ahora nos ha conducido desde la revisión de los biplots clásicos lineales, pasando por el estudio de los biplots logísticos de variables binarias, con destino a proponer un método para construir un biplot cuando el conjunto de datos contiene variables nominales, en el cuál cada individuo se representa como un punto en un espacio de baja dimensión y las variables se representan como "regiones de predicción" para cada una de las categorías de dichas variables. Dichas regiones son polígonos convexos que dividen el espacio de representación en tantas regiones como categorías tenga la variable, excepto si existen categorías ocultas, por tanto, lo que tenemos es una teselación del espacio en cualquier caso, que puede aproximarse por un diagrama de Voronoi, de tal forma que los generadores de dicho diagrama se pueden considerar como los "puntos categoría". Este tipo de representación se interpreta en términos de

distancias euclídeas, en el sentido de que la categoría predicha para cada individuo es aquella definida por el "punto categoría" más cercano a él.

La estimación de los parámetros del modelo propuesto es una adaptación simple del algoritmo EM. El clásico algoritmo EM alternado se ha tenido que modificar para incluir una estimación ridge penalizada de los parámetros del modelo logístico para evitar el conocido problema de la separación en regresión logística que hace que los estimadores no converjan y queden indeterminados. Existen otros métodos de penalización, como el Método Lasso para regresión logística [Meier y col., 1984] o el método de Firth [1993], aplicado a modelos multinomiales por Bull y col. [2002]. Los estimadores obtenidos del paquete de R mirt [Chalmers, 2012] se pueden usar como punto de partida para construir el biplot, utilizando las puntuaciones factoriales, pero con un paso adicional para reajustar el modelo logístico nominal para los parámetros de las variables. Esto es debido a que mirt está designado para la IRT y las puntuaciones están siempre calculadas con una rotación, pero los parámetros no parecen estar rotados como cabría esperar.

Mediante algunos ejemplos hemos comprobado que los valores numéricos que proporcionaba mirt en algunos casos eran extraños, posiblemente debido al problema de la separación y que este paquete no siempre es capaz de tenerlo en cuenta. Hemos detectado que tanto el método propuesto para variables nominales como mirt funcionan mejor cuando el número de individuos es bastante más grande que el de variables, pero hay muchos casos prácticos que se encuentran en la realidad en los cuales esto no es así, por ejemplo, tratando de clasificar un conjunto de individuos con los genotipos que resultan de miles de polimorfismos nucleóticos simples [Demey y col., 2008]. Para estos casos probablemente es más eficiente estimar los marcadores de los individuos mediante las coordenadas principales de la matriz $\mathbf{G}$ de indicadores definida anteriormente y después ajustar los modelos nominales sobre dichas coordenadas. Esta solución no es una solución de máxima verosimilitud, pero es una buena aproximación cuando el resto de métodos son inestables. La principal ventaja de usar máxima verosimilitud es que es posible

formular hipótesis que permiten comparar diferentes modelos, por ejemplo para conocer el número de dimensiones a retener. El método propuesto para variables nominales comparte las características "formales" de los modelos procedentes de la IRT o LTA y "descriptivas" de modelos como los del MCA y se podría considerar como una representación gráfica del modelo formal. Hay que indicar que el funcionamiento del algoritmo para aproximar e invertir la teselación depende en gran medida de la bondad de ajuste de la regresión nominal. Sólo aquellas variables con un ajuste razonablemente bueno se deberían representar en el gráfico. No obstante, sería necesario un estudio pormenorizado del método en casos en los que las variables tienen un gran número de categorías con un número también elevado de individuos, así como una investigación del método de inversión en esas situaciones para que sea más eficiente. Es necesaria una profundización mayor en el estudio de los métodos de inversión de las teselaciones, que pueden ser muy diversas y que en determinados casos quizá no tendrían por qué tener un diagrama de Voronoi "próximo" a ellas, así como una adaptación de algunos de los algoritmos existentes que permitieran comparar cómo afectan a las interpretaciones del biplot.

Hemos proporcionado una visión de cómo el "Biplot Logístico Nominal" puede contribuir a aportar información interesante contenida en una encuesta llevada a cabo a nivel nacional, comparándolo con el tradicional Análisis de Correspondencias Múltiples para variables categóricas. NLB es capaz de seleccionar la información relavante sobre las variables para la solución seleccionada y para cada plano de la misma, determinando aquellas categorías ocultas que nunca se predicen. En este sentido el gráfico es más limpio y facilita su lectura porque además el NLB se basa en distancias euclídeas, por lo que para cada fila, sus principales características son aquellos niveles de las variables más cercanas a ella. En esta línea, se ha tratado de plasmar las dificultades que se presentan a la hora de interpretar los gráficos en las técnicas más utilizadas con datos categóricos, intentando ofrecer alternativas más sencillas y que funcionan mejor a la hora de clasificar a los individuos e inferir sus características.

De Leeuw [2005] comenzó a trabajar en las técnicas de PCA no lineales para datos binarios utilizando ecuaciones de verosimilitud con funciones de enlace logit o probit y utilizando los algoritmos de mayorización iterativa de la escuela Gifi. Definía la función de pérdida mediante la matriz indicadora del conjunto de variables categóricas, y asumiendo residuos independientes, mediante la teoría de la mayorización (Blasius y Greenacre [2014]) el problema se reduce a problemas de componentes principales o MS según la función ϕ (la probabilidad de que el individuo i elija la categoría k de la variable j es proporcional a $\beta_{jk}e^{\phi(\mathbf{x}_i,\mathbf{y}_{jk})}$), que puede ser el producto interno, la distancia euclídea con signo menos, o la distancia euclídea al cuadrado con signo menos. El objetivo era minimizar menos la función de verosimilitud

$$\mathcal{L} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{k_j} g_{ijk} log\left\{\frac{\beta_{jk}e^{\phi(\mathbf{x}_i,\mathbf{y}_{jk})}}{\sum_{v=1}^{k_j}\beta_{jv}e^{\phi(\mathbf{x}_i,\mathbf{y}_{jv})}}\right\}$$

sobre las puntuaciones de los objetos ($\mathbf{x}_i$) y las cuantificaciones de las categorías ($\mathbf{y}_{jk}$). Esta formulación permitía, utilizando las restricciones sobre la cuantificación de las categorías usuales del entorno Gifi, reemplazar mínimos cuadrados por máxima verosimilitud y mínimos cuadrados alternados por mayorización. Este planteamiento es bastante diferente al que hemos desarrollado, puesto que en ningún momento se utiliza técnica alguna para cuantificar las variables categóricas en nuestra propuesta, aunque existe una interpretación subyacente por medio de la verosimilitud en ambos casos.

En el contexto del unfolding[1] multidimensional, Evans [2014] desarrolla un modelo logístico de asociación de distancias para el análisis exploratorio de datos categóricos ("Logistic Gifi"), en el cuál se utiliza una función de pérdida probabilística para obtener representaciones geométricas de datos en dimensiones reducidas, de forma que las distancias se corresponden de una manera directa con la estruc-

[1]El Unfolding es una técnica que se deriva del escalamiendo multidimensional, cuyo objetivo es situar en un mismo espacio de percepción a un conjunto de individuos y de estímulos en función de una serie de valoraciones de preferencia.

tura probabilística de los datos.

De igual forma, el desarrollo del "Biplot Logístico Ordinal", sus características y geometría ofrecen una alternativa multivariante al usuario para representar conjuntos de datos de una forma diferente a como se hacía hasta ahora. En dichas representaciones sólo la información relevante, en cuanto a las categorías predichas, de las variables aparece, al igual que en el caso nominal, haciendo que su lectura enfoque a lo que verdaderamente es importante e influyente en el plano de representación.

La principal contribución de este trabajo es que se ha extendido la construcción de biplots al caso de variables nominales, ordinales y conjuntos de datos con cualquier tipo de variables categóricas. En este sentido y para esta propuesta se han detallado nuevas metodologías para su construcción que han puesto en relación tanto al Análisis Multivariante como a la Geometría Computacional.

Algunos métodos estadísticos relacionados con variables binarias comentadas en este trabajo se han usado en algunos trabajos, como en Vicente-Galindo y col. [2011] en el sector empresarial con resultados satisfactorios, pero es necesario testar estas metodologías para variables categóricas en otros dominios. De la misma manera, los datos proporcionados por la "Encuesta de recursos humanos en ciencia y tecnología" del INE se han utilizado en recientes artículos como en Canal y Rodríguez [2012], aunque para unos propósitos diferentes utilizando variables en escala de intervalos. En cualquier caso, los estudios focalizados en este colectivo en España son muy escasos, sin embargo, estas técnicas permiten trabajar con este tipo de conjuntos grandes de datos de una forma visual y efectiva. Como hemos visto, el estudio de la satisfacción laboral por parte de los trabajadores españoles que poseen el título de doctor mediante el "Biplot Logístico Ordinal" ilustra de una forma gráfica algunas de las características de este colectivo ya estudiadas por otros autores, como Canal Domínguez [2013], con la flexibilidad además de incorporar al mismo variables de tipo nominal que refuerzan la interpretación de los resultados y amplían el abanico de posibilidades que ofrecen estas técnicas.

En nuestra investigación no hemos estudiado modelos relacionados con la matriz de Burt, que contiene las clasificaciones cruzadas de cada par de variables. En este ámbito Gabriel y col. [1998] utilizaban un biplot alternado con los logaritmos del recuento de las tablas de contingencia resultantes de cruzar parte de las variables en filas y el resto de variables en columnas; esto sería una submatriz de la matriz de Burt. Dichos autores utilizan una tabla de 3 vías con sólo 3 variables, pero el procedimiento se puede generalizar fácilmente a cualquier número. En este contexto la respuesta se puede aislar en las columnas (o filas) de la matriz o tabla. El biplot se utiliza posteriormente para hacer diagnosis de modelos logarítmico-lineales (o podrían ser modelos logit) para la matriz completa. En este caso el biplot se ajustaba con el logaritmo de los recuentos usando una aproximación lineal.

Utilizando modelos bilineales generalizados con la matriz de Burt de una forma similar a como se ha hecho en la modelizaciones descritas conduciría a una propuesta muy relacionada con la que hemos presentado y sería interesante analizar las diferencias que presentan, al igual que podría ser interesante en investigaciones futuras efectuar una comparación en profundidad de los métodos de análisis de variables categóricas en diferentes ámbitos de aplicación que pueda ayudar a identificar potencialidades adicionales de las técnicas que hemos desarrollado.

Capítulo 12

Conclusiones

Progress is impossible without change, and those who cannot change their minds cannot change anything.

George Bernard Shaw

1. La principal aportación de este trabajo es el estudio y la construcción de métodos biplot adecuados para conjuntos de datos formados por cualquier tipo de variables categóricas, de forma que el tratamiento de las mismas no requiere ninguna modificación de su naturaleza intrínseca, y cuya interpretación está basada en distancias euclídeas.

2. Se han analizado, como paso previo al estudio de las variables categóricas, los biplots clásicos desde el punto de vista de las regresiones alternadas, que permiten la estimación de las coordenadas de los individuos y las variables, y cuyo procedimiento converge al mismo subespacio generado por la SVD de la matriz de datos centrada. Además, se ha detallado la geometría de

estos biplots para comprender el funcionamiento de las escalas de medida de las variables, así como la relación de los mismos con los biplots predictivo e interpolativo de Gower.

3. Como primer paso en el estudio de las variables categóricas, en esta investigación se ha presentado una revisión de la teoría existente relativa a la construcción de biplots logísticos de variables binarias o dicotómicas, según la cual la respuesta a lo largo de las dimensiones retenidas es logística y cuyo ajuste se basa en regresiones alternadas. Los elementos de este modelo, formulación, estimación, geometría e interpretación han sido descritos y van a constituir la base sobre la que descansarán las construcciones posteriores.

4. En el caso de que las variables tengan más de dos categorías y que estas no estén ordenadas, es decir, en situaciones con conjuntos de datos nominales, se ha propuesto una metodología novedosa de construcción de un biplot adecuado para los mismos, que hemos denominado "Biplot Logístico Nominal". Este procedimiento estima modelos multidimensionales de respuesta latente para variables politómicas, cuya geometría se basa en la determinación de las regiones de predicción de cada variable mediante un algoritmo que es capaz de detectar los nodos de la teselación resultante. La inversión de dicha teselación mediante el cálculo de sus generadores nos ha conducido a una representación gráfica sencilla, en la cual, sólo aquellas categorías que se predicen probabilísticamente son visibles, cuya interpretación está basada en distancias euclídeas entre puntos y en la que no son necesarias las proyecciones.

5. El método multivariante "Biplot Logístico Nominal" ha sido puesto a prueba con varios conjuntos de datos, poniendo de manifiesto su funcionamiento y ha sido comparado con el Análisis de Correspondencias Múltiple para uno de dichos conjuntos, puesto que esta última es una de las técnicas multivariantes más extendidas para el tratamiento de datos categóricos. Dicho estudio ha

mostrado como la nueva técnica es capaz de proporcionar información tan valiosa como la del MCA, pero eliminando información superflua del gráfico, situando los puntos categoría en las regiones de predicción, clasificando mejor a los individuos según el perfil que presentan en los datos y abstrayéndose de interpretaciones basadas en perfiles medios, puesto que la lectura del gráfico está basada en la proximidad entre los elementos del mismo.

6. Avanzando un paso más en el tratamiento de variables categóricas, se ha propuesto un método multivariante para la construcción de biplots cuando existe una ordenación en los valores de las categorías de los ítems, llamado "Biplot Logístico Ordinal". La obtención de la representación biplot se ha llevado a cabo atendiendo a la particular geometría de las curvas, que no son sigmoides, aunque el conjunto de puntos que preciden un valor de la probabilidad se sitúan en una línea recta y diferentes probabilidades para todas las categorías de una variables se disponen en rectas paralelas, siendo el eje del biplot perpendicular a todas ellas. Además, este procedimiento tiene en cuenta las categorías que no se predicen nunca en el plano de representación, mostrando sólo aquellas que son relevantes.

7. Un amplio conjunto de datos correspondientes a una encuesta del INE se ha utilizado para testear el método ordinal, arrojando tanto la potencia gráfica del mismo en situaciones con un gran número de invididuos como las posibilidades de análisis individuales para cada una de las variables en el espacio reducido, e incluso combinando esta información con variables externas de otros tipos.

8. Ambos métodos utilizan como método de estimación una versión específica del algoritmo EM preparada para detectar el problema de la separación en regresión logística que sea capaz de proporcionar estimaciones consistentes. Dicho algoritmo se ha descrito convenientemente, así como una revisión de las posibles alternativas de estimación de este tipo de modelos. También se

ha abordado la problemática de la bondad de ajuste de los métodos, tanto global como particular de cada variable, puesto que interesan modelos que sean capaces de clasificar correctamente a los individuos según sus elecciones.

9. Hemos preparado dos paquetes disponibles para la comunidad científica de forma gratuita en el CRAN de R, llamados NominalLogisticBiplot y OrdinalLogisticBiplot que permiten efectuar la estimación, análisis y obtención de los gráficos en situaciones en las que los conjuntos de datos son nominales y ordinales respectivamente, siguiendo las metodologías descritas. Dichas herramientas han sido descritas en cuanto a su funcionamiento y disponen de manuales de uso en la plataforma web citada.

10. Cuando en el conjunto de datos aparecen conjuntamente variables nominales y ordinales se ha analizado la literatura disponible sobre modelos de respuesta latente en un contexto más general, en el cual además nos podríamos encontrar con variables continuas. Fruto de esta revisión se ha modificado el algoritmo de estimación de los modelos anteriores para que puedan convivir todo tipo de variables categóricas, de forma que la construcción del biplot en estos casos queda perfectamente descrita. Además, se ha creado un paquete similar a los anteriores, llamado CategoricalLogisticBiplot, que permite trabajar con datos categóricos mixtos y el mismo se ha puesto a prueba con un conjunto de datos ya estudiado anteriormente para hacer un análisis comparativo de los resultados.

Bibliografía

Adamatzky, A. (1993). Massively parallel algorithm for inverting voronoi diagram. *Neural Network World*, 4:385–392.

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

Albanese, M. (1990). Latent variable models for binary response. Master's thesis, PhD thesis. University of London.

Albanese, M. y Knott, M. (1994). Bootstrapping latent variable models for binary response. *British Journal of Mathematical and Statistical Psychology.*, 47:235–245.

Albert, A. y Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71(1):1–10.

Alonso Ferrero, M. (2011). Voronoi diagram: The generator recognition problem. *arXiv:1105.4246v1 [cs.CG]*.

Aloupis, G., Pérez-Roses, H., Pineda-Villavicencio, G., Taslakian, P., y Trinchet-Almaguer, D. (2013). Fitting voronoi diagrams to planar tesselations. *Conference Paper. IWOCA*.

Alves, M., Cunha, S., Amaral, J., Pereira, J., y Oliveira, M. (2005). Classification of pdo olive oils on the basis of their sterol composition by multivariate analysis. *Analytica Chimica Acta*, 549:166–178.

Andersen, E. B. (1980). *Discrete statistical models with social science applications.*

Anderson, J. (1972). Separate sample logistic discrimination. *Biometrika*, 59:19–35.

Anderson, J. y Philips, P. (1981). Regression, discrimination and measurement models for ordered categorical variables. *Applied Statistics*, 30:22–31.

Archer, K., Lemeshow, S., y Hosmer, D. (2007). Goodness-of-fit tests for logistic regression models when data are collected using a complex sampling design. *Computational Statistica & Data Analysis*, 51:4450–4464.

Aurenhammer, F. (1987). Recognizing polytopical cell complexes and constructing projection polyhedra. *J. Symbolic Comput.*, 3:249–255.

Azzalini, A., Bowman, A., y Härdle, W. (1989). On the use of nonparametric regression for model checking. *Biometrika*, 76(1):1–11.

Baker, F. (1992). *Item Response Theory. Parameter Estimation Techniques.* Marcel Dekker.

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical and Statistical Psicology*, 12(4):387–415.

Banerjee, S., Bhattacharya, B., Das, S., Karmakar, A., Maheshwari, A., y Roy, S. (2012). On the construction of a generalized voronoi inverse of a rectangular tesselation. *In: Procs. 9th Int. IEEE Symp. on Voronoi Diagrams in Science and Engineering. IEEE, New Brunswick, NJ.*, pages 132–137.

Barankin, E. y Maitra, E. (1963). Generalisation of the fisher-darmois-koopman-pitman theorem on sufficient statistics. *Sankhya A.*, 25:153–169.

Barbezat, R. y Hughes, J. (2005). Salary structure effects and the gender pay gap in academia. *Research in Higher Education*, 46:621–640.

BIBLIOGRAFÍA

Bartholomew, D. (1984a). The foundations of factor analysis. *Biometrika.*, 71:221–232.

Bartholomew, D. (1984b). Scaling binary data using a factor model. *Journal of the Royal Statistical Society, Series B.*, 46:120–123.

Bartholomew, D. (1987). Latent variable models and factor analysis. *Griffin's Statistical Monographs, vol.40. Charles Griffin, London.*

Bartholomew, D. y Leung, S. (2000). A goodness of fit test for sparse 2^p contingency tables. *British Journal of Mathematical and Statistical Psychology*, 55:1–15.

Bartholomew, D., Steele, F., Moustaki, I., y Galbraith, J. (2008). *Analysis of Multivariate Social Science Data, 2nd edn.* CRC Press, Boca Raton, FL.

Bartholomew, D. J. y Tzamourani, P. (1999). The goodness-of-fit of latent trait models in attitude measurement. *Sociolological Methods and Research*, 27:525–546.

Bayer, A. y Astin, H. (1975). Sex differentials in the academic reward system. *Science*, 188:796–802.

Benzecri, J. (1973). *L'Analyse des Données.* Tome II:L'Analyse des correspondances. Paris:Dunod., 619 pp.

Benzécri, J. (1973). *L'analyse des données: L'analyse des correspondances.* L'analyse des données: leçons sur l'analyse factorielle et la reconnaissance des formes et travaux. Tome II:L'Analyse des correspondances. Paris:Dunod., 619 pp.

Benzecri, J. (1979). Sur le calcul des taux d'inertie dans l'analyse d'un questionnaire. *Cahiers de l'Analyse des Donneés*, 4:377,378.

Blasius, J. y Greenacre, M. (1998). *Visualization of Categorical Data. Blasius, J.; Greenacre, M. (Eds).* Academic Press. San Diego.

Blasius, J. y Greenacre, M. (2014). *Visualization and Verbalization of Data. Blasius, J.; Greenacre, M. (Eds).* CRC Press. Taylor & Francis Group. Chapman & Hall.

Blázquez, A. (1998). Análisis biplot basado en modelos lineales generalizados. tesis doctoral. universidad de salamanca. españa. page 240p.

Bock, R. y Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an em algorithm. *Psychometrika*, 46(4):443–459.

Bock, R. y Haberman, S. (2009). Confident bands for examining goodness-of-fit of estimated item response functions. *Paper presented at the annual meeting of the Psychometric Society, Cambridge, UK.*

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37:29–51.

Bock, R. D., Gibbons, R., y Muraki, E. (1988). Full-information item factor analysis. *Psychological Measurement*, 12(3):261–280.

Bock, R. D. y Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, 35:179–197.

Bock, R. D. y Shilling, S. G. (1997). High-dimensional maximum marginal likelihood item factor analysis. *Manuscript in preparation.*

Boyles, R. (1983). On the convergence of the em algorithm. *Journal of Royal Statistical Society*, B45(1):47–50.

Brand, H. (2003). PCA and CVA biplots : a study of their underlying theory and quality measures. Master's thesis, Stellenbosch University.

Brown, C. (1982). On a goodness-of-fit test for the logistic model based on score statistics. *Comm. Statist. Theory Meth.*, 11(10):1087–1105.

Brown, M. y Cudeck, R. (1993). Alternative ways of assessing model fit. in k.a. bollen, & j.s. long,(eds). *Testing Structural Equation Models. Newbury Park, CA:Sage.*, pages 136–162.

Browne, R. P. y McNicholas, P. D. (2013). Estimating common principal components in high dimensions. *Advances in Data Analysis and Classification*, pages 1–10.

Bull, S. B., Mak, C., y Greenwood, C. M. (2002). A modified score function for multinomial logistic regression. *Computational Statistics and data Analysis*, 39:57–74.

Byrne, B. (2001). *Structural Equation Modeling with Amos.* Mahwah, New Jersey: Lawrence Erlbaum.

Cai, L. (2010a). High-dimensional exploratory item factor analysis by a metropolis-hastings robbins-monro algorithm. *Psychometrika*, 75(1):33–57.

Cai, L. (2010b). Metropolis-hastings robbins-monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics.*, 35(3):307–335.

Cai, L., Maydeu-Olivares, A., Coffman, D., y Thissen, D. (2006). Limited information goodness-of-fit testing of item response theory models for sparse 2^p tables. *British Journal of Mathematical and Statistical Psychology*, 59:173–194.

Canal, J. y Rodríguez, C. (2012). Wage differences among PhDs by area of Knowledge: are science areas better paid than humanities and social ones?. The spanish case. *Journal of Education and Work*, 1(32).

Canal, J. F. y Muñiz, M. A. (2012). Professional doctorates and the careers: Present and future. the spanish case. *European journal of education.*, 47(1):153–171.

Canal Domínguez, J. (2013). Earnings and job satisfaction of employed spanish doctoral graduates. *Revista Española de Investigaciones Sociológicas*, 144:49–72.

Casella, G. y George, E. (1992). Explaining the gibbs sampler. *The American Statistician.*, 46:167–174.

Cecere, S., Groenen, P. J. F., y Lesaffre, E. (2013). The Interval-Censored Biplot. *Journal of computational and graphical statistics*, 22(1):123–134.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48(6):1–29.

Cohen, J., Cohen, P., West, S., y Aiken, L. (2002). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences(3rd edition)*. Routledge.

Cox, D. (1958). Two further applications of a model for binary responses. *Biometrika*, 45:562–565.

Cox, D. (1970). Analysis of binary data. *Methuen, London*, pages 26–105.

Cox, D. y Snell, E. (1989). *Analysis of Binary Data(2nd edition)*. Chapman & Hall.

Cox, T. y Cox, M. (2000). *Multidimensional Scaling(2nd edition)*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability (Book 88).

Cragg, J. y Uhler, R. (1970). The demand for automobiles. *The Canadian Journal of Economics*, 3:386–406.

Cuadras, C. M., C. D. y Greenacre, M. (2012). Comparison of different methods for representing categorical data. *Comm. Stat. Simul. and Comp.*, 2(35):447–459.

Cuadras, C. M. y Cuadras, D. (2006). A parametric approach to correspondence analysis. *Linear Algebra and its Applications*, 417:64–74.

Cuadras, C. M. y Cuadras, D. (2011). *Partitioning the geometric variability in multivariate analysis and contingency tables.* B. Fichet, D. Piccolo, R. Verde, M. Vichi, Eds., Classification and Multivariate Analysis for Complex Data Structures, pp. 237-244. Springer, Berlin.

De Leeuw, J. (2005). Gifi goes logistic. *SCASA Keynote.*

De Leeuw, J. (2006). Principal component analysis of binary data by iterated singular value decomposition. *Computational Statistics and Data Analysis*, 50(1):21–39.

Demey, J., Vicente-Villardón, J. L., Galindo, M. P., y Zambrano, A. (2008). Identifying molecular markers associated with classification of genotypes using external logistic biplots. *Bioinformatics*, 24(24):2832–2838.

Dempster, A. P., Laird, N. M., y Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38.

Drasgow, F. y Parsons, C. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement*, 7:189–199.

Edwards, M. (2010). A markov chain monte carlo approach to confirmatory item factor analysis. *Psychometrika*, 75(3):474–497.

Escofier, B. (1978). Analyse factorielle et distances répondant au principe d'équivalence distributionnelle. *Revue de Statistiques Appliquées*, 26:29–37.

Evans, D. G. y Jones, S. M. (1987). Detecting voronoi (area of influence) polygons. *Mathematical Geology*, 19(6):523–537.

Evans, G. (2014). Logistic gifi: A logistic distance association model for exploratory analysis of categorical data. Master's thesis, Thesis. University of California, Los Angeles.

Fabra, M. E. y Camisón, C. (2009). Direct and indirect effects of education on job satisfaction: A structural equation model for the spanish case. *Economics of Education Review*, 28:600–610.

Falguerolles, A. d. (1998). Log-bilinear biplots in action. *Visualization of Categorical Data. Blasius, J.; Greenacre, M. (Eds). Academic Press. San Diego.*, page 594p.

Falguerolles, A. d. y Francis, B. (1994). An algorithmic approach to bilinear models for two-way contingency tables. *In New Approaches in Classification and Data Analysis. Springer Berlin Heidelberg.*, pages 518–524.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1):27–38.

Frutos, E., Galindo, M., y Leiva, V. (2013). An interactive biplot implementation in r for modeling genotype-by-environment interaction. *Stochastic Environmental Research and Risk Assessment.*

Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58(3):453–467.

Gabriel, K. R. (1998). Generalised bilinear regresión. *Biometrika*, 85(3):689–700.

Gabriel, K. R. (2002). Goodness of fit of biplots and correspondence analysis. *Biometrika*, 89(2):423–436.

Gabriel, K. R., Galindo, M. P., y Vicente-Villardón, J. L. (1998). *Use of Biplots to diagnose independence models in three way contingency tables.*, pages 391–404. Academic Press.

BIBLIOGRAFÍA

Gabriel, K. R. y Zamir, S. (1979). Lower rank approximation of matrices by least squares with any choice of weights. *Technometrics*, 21(4):489–498.

Galindo, M. P. (1986). Una alternativa de representacion simultanea: Hj-biplot. *Questiio*, 10(1):13–23.

Gallego-Álvarez, I. y Vicente-Villardón, J. L. (2012). Analysis of environmental indicators in international companies by applying the logistic biplot. *Ecological Indicators*, 23(0):250–261.

Gardner-Lubbe, S., Le Roux, N., y Gower, J. (2008). Measures of fit in principal component and canonical variate analyses. *Journal of Applied Statistics*, 35(9):947–965.

Gibbons, R., Darrel, R., Hedeker, D., Weiss, D., Segawa, E., Bhaumik, D., Kupfer, D., Frank, E., Grochocinski, V., y Stover, A. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement*, 31(1):4–19.

Gibbons, R. y Hedeker, D. (1992). Full-information item bi-factor analysis. *Psychometrika*, 57:423–436.

Gifi, A. (1990). *Nonlinear multivariate analysis*. Wiley.

Goodman, L. A. (1991). Measures, models, and graphical displays in the analysis of cross-classified data. *Journal of the American Statistical association.*, 86(416):1085–1111.

Gourieroux, C. y Monfort, A. (1997). *Simulation based econometric methods.* New York, NY: Oxford University Press.

Gower, J., Gardner-Lubbe, S., y le Roux, N. (2011). *Understanding Biplots*. Wiley, N. York.

Gower, J., Groenen, P., y Van De Velden, M. (2010). Area biplots. *Journal of Computational and Graphical Statistics*, 19:46–61.

Gower, J. y Hand, D. (1996). *Biplots.* Monographs on statistics and applied probability. 54. London: Chapman and Hall., 277 pp.

Green, M. (1996). Generalized factor analysis. *Proceedings of the 11th International Workshop on Statistical Modelling. Orvieto, Italy.*

Greenacre, M. J., editor (1984). *Theory and Applications of Correspondence Analysis.* London. Academic Press.

Greenacre, M. J. (1988). Correspondence analysis of multivariate categorical data by weighted least squares. *Biometrika*, 75:457–467.

Greenacre, M. J., editor (1993). *Correspondence analysis in practice.* London. Academic Press.

Greenacre, M. J., editor (2004). *Geometric Data Analysis: From Correspondence Analysis to Structured Data.* Dordrecht: Kluwer.

Greenacre, M. J., editor (2008). *La Práctica del Análisis de Correspondencias.* Fundación BBVA - Rubes Ed., Barcelona.

Greenacre, M. J. (2010). *Biplots in Practice.* Books. Fundacion BBVA / BBVA Foundation.

Greenacre, M. J. (2012). Biplots: the joy of singular value decomposition. *WIREs Comp Stat*, 4:399–406.

Greenacre, M. J. (2013). Contribution Biplots. *Journal of computational and graphical statistics*, 22(1):107–122.

Greenacre, M. J. y Blasius, J., editors (2006). *Multiple correspondence analysis and related methods.* Statistics in the social and behavioral sciences series. Chapman & Hall/CRC, Boca Raton.

Greenacre, M. J. y Groenen, P. J. F. (2013). Weighted euclidean biplots. Economics Working Papers 1380, Department of Economics and Business, Universitat Pompeu Fabra.

Guttman, L. (1941). The quantification of a class of attibutes: a theory and method of scale construction. *The Prediction of Personal Adjustment, (eds P. Horst et al.).The Prediction of Personal Adjustment, New York: Social Science*, 48:319–348.

Haberman, S. (1974). The analysis of frequency data. *University of Chicago Press.*

Haberman, S. (2009). Use of generalized residuals to examine goodness of fit of item response models. *ETS Research Report. No.RR-09-15.*

Hajian-Tilaki, K. (2013). Receiver operating characteristic (roc) curve analysis for medical diagnostic test evaluation . *Caspian Journal of Internal Medicine*, 4(2):627–635.

Hambleton, K. (2000). Introduction to item response theory. *Session presented at the Sylvan Prometric Results 2000 Conference, Tucson AZ.*

Hand, D. (1994). Assessing classification rules. *Journal of Applied Statistics*, 21:3–16.

Hanley, J. y McNeil, B. (1982). The meaning and use of the area under a receiving operating characteristic (roc) curve. *Radiology*, 143:29–36.

Harrel, F. (2001). Regression modeling strategies: With applications to linear models. logistic regression and survival analysis. *(Springer Series in Statistics). Springer, New York.*

Hartvigsen, D. (1992). Recognizing voronoi diagrams with linear programming. *ORSA Journal on Computing*, 4:369–374.

Hastings, W. (1970). Monte carlo simulation methods using markov chains and their applications. *Biometrika*, 57:97–109.

Heinze, G. y Schemper, M. (2002). A solution to the problem of separation in logistic regresion. *Statistics in Medicine*, 21:2409–2419.

Hernández, J. C. y Vicente-Villardón, J. L. (2013). *Nominal Logistic Biplot: Biplot representations of categorical data*. R package version 0.1.

Hernández, J. C. y Vicente-Villardón, J. L. (2014). *Ordinal Logistic Biplot: Biplot representations of ordinal variables*. Universidad de Salamanca.Department of Statistics. R package version 0.3.

Hernández Sanchez, J. C. y Vicente-Villardón, J. L. (2013). Logistic biplot for nominal data. *ArXiv e-prints*.

Hoerl, A. y Kennard, R. (1970). Ridge regression: Biased estimation for non ort-hogonal problems. *Technometrics*, 12:55–67.

Holland, P. W. (1990). On the sampling theory foundations of item response theory models. *Psychometrika*, 55:577–601.

Hosmer, D., Hosmer, T., le Cessie, S., y Lemeshow, S. (1997). A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine*, 16(9):965–980.

Hosmer, D. y Lemeshow, S. (1980). A goodness-of-fit test for the multiple logistic regression model. *Communications in Statistics*, A10:1043–1069.

Hosmer, D. y Lemeshow, S. (2000). *Applied Logistic Regression(2nd edition)*. Wiley.

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

Hotelling, H. (1935). The most predictable criterion. *Journal of Educational Psychology*, 26:139–142.

Hsu, Y., Ackerman, T., y Fan, M. (1999). The relationship between the bock-aitkin procedure and the em algorithm for irt model estimation. *Iowa City, Iowa, ACT, Inc.*

Huber, P., Ronchetti, E., y M.P., V.-F. (2004). Estimation of generalized linear latent variable models. *J. R. Statis. Soc, B.*, 66(4):893–908.

Hulin, C., Lissak, R., y Drasgow, F. (1982). Recovery of two and three parameter logistic item characteristic curves: A monte carlo study. *Applied Psychological Measurement*, 6:249–260.

Jäckel, P. (2005). A note on multivariate gauss-hermite quadrature.

Jongman, R. H. G., Braak, C. J. F. T., y Tongeren, O. F. R. V. (1987). *Data Analysis in Community and Landscape Ecology.* Cambridge University Press.

Jöreskog, K. G. (1990). New developments in lisrel: Analysis of ordinal variables using polychoric correlations and weighted least squares. *Quality and Quantity*, 24:387–404.

Jöreskog, K. G. y Sörbom, D. (1993). Structural equation modeling with the simplis command language. *Chicago: Scientific Software.*

Jöreskog, K. G. y Sörbom, D. (2006). Lisrel 8.8 for windows.. *Scientific Software International. Lincolnwood, IL. Computer software.*

Kang, T. y Chen, T. (2007). An investigation of the performance of the generalized s-x^2 item-fit index for polytomous irt models. *ACT Reseach Report Series.*

Kashyap, R. (1982). Optimal choice of ar and ma parts in autoregresive moving average models. *IEEE Trans. Pattern Anal Mach Intell*, 4(2):99–104.

Khalid, M. (2009). Irt model fit from different perspectives. Master's thesis, Thesis. University of Twente.

Kullback, S. y Leibler, R. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

la Grange, A., le Roux, N., y Gardner-Lubbe, S. (2009). BiplotGUI: Interactive Biplots in R. *Journal of Statistical Software*, 30(12).

Langeheine, R., Pannekoek, J., y van de Pol, F. (1996). Bootstrapping goodness-of-fit measures in categorical data analysis. *Sociological Methods and Research*, 24:492–516.

Lawley, D. y Maxwell, A. (1971). *Factor analysis as a statistical method.* London: Butterworth.

le Cessie, S. y van Houwelingen, J. (1991). A goodness-of-fit test for binary regression models, based on smoothing methods. *Biometrics*, 47:1267–1282.

Le Cessie, S. y Van Houwelingen, J. (1992). Ridge estimators in logistic regression. *Applied Statistics*, 41(1):191–201.

le Cessie, S. y van Houwelingen, J. (1995). Testing the fit of a regression model via score tests in random effects models. *Biometrics*, 51:600–614.

Le Roux, N. y Gardner, S. (2005). Analysing your multivariate data as a pictorial: A case for applying biplot methodology? *International Statistical Review*, 73(3):365–387.

Lee, S., Huand, J., y Hu, J. (2010). Sparse logistic principal component analysis for binary data. *Annals of Applied Statistics*, 4(3):21–39.

Liu, Y., Nelson, P., y Yang, S. (2012). An omnibus lack of fit test in logistic regression with sparse data. *Statistical Methods and Applications*, 21:437–452.

Locke, E. A. (1976). *The nature of causes of job satisfaction.* En: M.D. Dunnette(ed.), Handbook of Industrial and Organizational Psychology, Chicago: Rand-McNally.

López-Ratón, M., Rodríguez-Álvarez, M., Cadarso-Suarez, C., y Gude-Sampedro, F. (2014). Optimalcutpoints: An r package for selecting optimal cutpoints in diagnostic tests. *Journal of Statistical Software*, 61(8):1–36.

Maddala, G. (1983). *Limited Dependent and Qualitative Variables in Econometrics.* Cambridge University Press.

Mair, P. y Hatzinger, R. (2007). Extended rasch modeling: The eRM package for the aplication of irt models in R. *Journal of Statistical Software*, 20(9):1–20.

Mair, P., Reise, S. P., y Bentler, P. (2008). Irt goodness-of-fit using approaches from logistic regression. *UCLA Statistics Preprint Series*, 540.

Malo, N., Libiger, O., y Schork, N. (2008). Accommodating linkage desequilibrium in genetic-association analysis via ridge regression. *Am. J. Hum. Genet.*, 82(2):375–385.

Man-Lai, T. (2001). Exact goodness-of-fit test for binary logistic model. *Statistica Sinica*, 11:199–211.

Martin, A., Quinn, K., y Park, J. (2011). MCMCpack: Markov chain monte carlo in R. *Journal of Statistical Software*, 42(9):1–21.

Masters, G. (1982). A rasch model for partial credit scoring. *Psychometrika*, 47:149–174.

Mavridis, D., Moustaki, I., y Knott, M. (2007). Goodness-of-fit measures for latent variable models for binary data. in s.-y. lee (ed.). *Handbook of Latent Variable and Related Models.*, pages pp. 135–162.

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement: Interdisciplinary Research and Perspectives.*, 11(3):71–101.

Maydeu-Olivares, A. y Joe, H. (2005). Limited and full-information estimation and goodness-of-fit testing in 2^n contingency tables: A unified framework. *Journal of the American Statistical Association.*, 100(471):713–732.

Maydeu-Olivares, A. y Joe, H. (2008). An overview of limited information goodness-of-fit testing in multidimensional contingency tables. in k. shigemasu, a. okada, t. imaizumi, & t. hoshino (eds.). *New Trends in Psychometrics*, pages pp.253–262.

Maydeu-Olivares, A. y Joe, H. (2014). Assessing approximate fit in categorical data analysis. *Multivariate Behavioral Reseach.*, 49(11):305–328.

Maydeu-Olivares, A. y Montaño, R. (2013). How should we assess the fit of rasch-type models?. approximating the power of goodness-of-fit statistics in categorical data analysis. *Psychometrika.*, 78(1):116–133.

McCullagh, P. (1985). On the asymptotic distribution of pearson's statistics in linear exponential family models. *International Statistical Review*, 53:61–67.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics. Academic Press*, pages 105–142.

McFadden, D. (1989). A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrika*, 57:995–1026.

McKinley, R. y Mills, C. (1995). A comparison of several goodness-of-fit statistics. *Applied Psychological Assessment, American Psychologist.*, 50:741–749.

McNabb, R. y Wass, V. (1997). Male-female salary differentials in british universities. *Oxford Economic Papers, New Series.*, 49:328–343.

Mehta, C. y Patel, N. (1995). Exact logistic regression: theory and examples. *Statistics in Medicine*, 14:2143–2160.

Meier, L., van de Geer, S., y Buhlmann, P. (1984). The group lasso for logistic regression. *J. R. Statist. Soc.*, 70(1):53–71.

Mellenbergh, G. (1992). Generalized linear item response theory. *Psychological bulletin*, 115:300–307.

Metropolis, N., Rosenbluth, A., Teller, A., y Teller, E. (1953). Equations of state space calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1091.

Montesinos López, A. (2011). Estudio del aic y bic en la selección de modelos de vida con datos censurados. Master's thesis, Centro de Investigación en Matemáticas, A.C.(CIMAT). Guanajuato.

Moustaki, I. (1996). A latent trait and a latent class model for mixed observed variables. *British Journal of Mathematical and Statistical Psychology.*, 49:313.334.

Moustaki, I. (1999). *LATENT: A computer program for fittin a one or two factor latent variable model to categorical, metric and mixed observed item with missing values (Technical Report).* London School of Economics and Political Science, Statistics Department.

Moustaki, I. (2000). A latent variable model for ordinal variables. *Applied Psychological Measurement*, 24:211–223.

Moustaki, I. y Knott, M. (2000). Generalized latent trait models. *Psychometrika.*, 65:391–411.

Moustaki, I. y Papageorgiou, I. (2004). Latent class models for mixed outcomes with applications in archaeometry. *Computational Statistics & Data analysis.*, 48:659–675.

Muraki, E. (1990). Fitting a polytomous item response model to likert-type data. *Applied Psychological Measurement*, 14:59–71.

Muraki, E. (1992). A generalized partial credit model: Application of an em algorith. *Applied Psychological Measurement*, 16(2):159–176.

Muraki, E. y Bock, R. (1991). *PARSCALE: Parameter scaling of rating data.* Chicago: Scientific Software.

Muraki, E. y Carlsson, E. (1995). Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement*, 19:73–90.

Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika.*, 47:337–347.

Muthén, B. (1984). A general structural model with dichotomous, ordered categorical and continuous latent variable indicators. *Psychometrika*, 49(1):115–132.

Muthén, B. (1987). *LISCOMP: Computer Program.* Chicago: Scientific Software.

Muthén, L. y Muthén, B. (1984). *Analysis of Multivariate Social Science Data, 2nd edn.* Muthén and Muthén, Los Angeles.

Nagelkerke, N. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, 78:691–692.

Naidoo, S., Harris, A., Swanevelder, S., y Lombard, C. (2006). Fetal alcohol syndrome: A cephalometric analysis of patients and controls. *European Journal of Orthodontics*, 28:254–261.

Nenadić, O. y Greenacre, M. J. (2007). Correspondence analysis in R, with two- and three-dimensional graphics: The ca package. *Journal of Statistical Software*, 20(3).

Nieto, A., Galindo, M., Leiva, V., y Vicente-Galindo, P. (2014). A methodology for biplots based on bootstrapping with r. *Revista Colombiana de Estadística*, 37(2):367–397.

Nieto, F. y Vicente-Villardón, J. (2012). The problem of separation in binary, multinomial and ordinal regression. Master's thesis, Trabajo de Fin de Grado. Universidad de Salamanca.

Okabe, A., Boots, B., Sugihara, K., y Chiu, S. (2000). *Spatial Tesselations(2nd edition)*. Wiley, Chichester.

Orlando, M. y Thissen, D. (2000). New item fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, 24:50–64.

Orlando, M. y Thissen, D. (2003). Further investigation of the performance of s-x^2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement*, 27:289–298.

Osius, G. y Rojek, D. (1992). Normal goodness-of-fit tests for multinomial models with large degrees of freedom. *Journal of American Statistical Association*, 87(420):1145–1152.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series 5*, 50(302):157–175.

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

Pigeon, J. y Heyse, J. (1991). An improved goodness of fit test for probability prediction models. *Biometrical Journal*, 41:71–82.

Pigeon, J. y Heyse, J. (1999). A cautionary note about assessing the fit of logistic regression models. *Journal of Applied Statistics*, 26(7):847–853.

Pulkstenis, E. y Robinson, T. (2002). Two goodness-of-fit tests logistic regression models with continuous covariates. *Statistics in Medicine*, 21:79–93.

Rabe-Hesketh, S., Skrondan, A., y Pickles, A. (2002). Reliable estimation of generalized linear mixed models using adaptative quadrature. *Stata J.*, 2:1–21.

Ramsay, J. (1975). Solving implicit equations in psychometric data analysis. *Psychometrika*, 40:337–360.

Rao, C. (1995). *Use of Hellinger distance in graphical displays.* In E.-M. Tiit, T. Kollo, and H. Niemi (Ed.): Multivariate statistics and matrices in statistics. Leiden (Netherland): Brill Academic Publisher, pp. 143-161.

Reckase, M. (2009). *Multidimensional Item Response Theory.* Springer-Verlag, New York.

Redner, R. y Walker, H. (1984). Mixture densities, maximum likelihood, and the em algorithm. *SIAM Review*, 26:195–239.

Reiser, M. (1996). Analysis of residuals for the multinomial item response model. *Psychometrika*, 61:509–528.

Reiser, M. y VandenBerg, M. (1994). Validity of the chi-square test in dichotomous variable factor analysis when expected frequencies are small. *British Journal of Mathematical and Statistical Psychology*, 47:85–107.

Rizopoulos, D. (2006). ltm: An R package for lateng variable modeling and item response theory analysis. *Journal of Statistical Software*, 17(5):1–25.

Royston, P. (1992). The use of cusums and other techniques in modeling continuous covariates in logistic regression. *Statistics in Medicine*, 11:1115–1129.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika, Monograph Supplement*, 17.

Sammel, M., Ryan, L., y Legler, J. (1997). Latent variable models for mixed discrete and continous outcomes. *Journal of the Royal Statistical Society, Series B.*, 59:667–678.

Schilling, S. y Bock, R. D. (2005). High-dimensional maximum marginal likelihood item factor analysis by adaptative quadrature. *Psychometrika*, 70:533–555.

Schoenberg, F., Ferguson, T., y Li, C. (2003). Inverting dirichlet tesselations. *Computer journal*, 46(1):76–83.

Schwabe, M. (2011). The careers paths of doctoral graduates in austria. *European Journal of Education.*, 46(1):153–168.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464.

Sclove, S. (1987). Application of model-selection criteria to some problems of multivariate analysis. *Psychometrika.*, 52:333–343.

Scrucca, L. (2013). Graphical tools for model-based mixture discriminant analysis. *Advances in Data Analysis and Classification*, pages 1–19.

Sheng, Y. (2010). Bayesian estimation of mirt models with general and specific latent traits in matlab. *Journal of Statistical Software*, 34(3):1–27.

Silvapulle, M. (1981). On the existence of maximum likelihood estimates for the binomial response models. *Journal of the Royal Statistical Society. Series B.*, 43(3):310–313.

Sinharay, S., Haberman, S., y Jia, H. (2011). Fit of item response theory models: A survey of data from several operational tests. *ETS Research Report. No.RR-11-29.*

Skinner, C., Holt, D., y Smith, T. (1989). *Analysis of Complex Surveys*. New York: John Wiley.

Smith, R. (2002). The family approach to assessing fit in rasch measurement. *Paper presented at the 11th International Objective Measurement Workshop, New Orleans*.

Spector, P. E. (1976). *Job Satisfaction: Application, Assessment, Causes, and Consequences*. London: Sage.

Steiger, J. y Lind, J. (1980). Statistically-based tests for the number of common factors. *Paper presented at the annual meeting of the Psychometric Society, Iowa City*.

Steiger, J. H. (1990). Structural model evaluation and modification: an interval estimation approach. *Multivariate Behavioural Research*, 25:173–180.

Stern, S. (1997). Simulation-based estimation. *Journal of Economic Literature.*, 35:2006–2039.

Su, J. y Wei, L. (1991). A lack-of-fit test for the mean function in a generalized lineal model. *Journal of American Statistical Association*, 86(414):420–426.

Sun, H. y Wang, S. (2012). Penalized logistic regression for high-dimensional dna methylation data with case-control studies. *Bioinformatics*, 28(10):1368–1375.

Suzuki, A. y Iri, M. (1986). Approximation of a tesselation of the plane by a voronoi diagram. *Journal of the Operations Research. Society of Japan.*, 29(1):69–97.

R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*. *R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0*.

Thissen, D. (1991). *MULTILOG: Multiple, categorical items analysis and test scoring using item response theory*. Chicago: Scientific Software.

Tjur, T. (2009). Coefficients of determination in logistic regression models. a new proposal: The coefficient of discrimination. *The American Statistician*, 63:366–372.

Tollenaar, N. y Mooijaart, A. (2003). Type i errors and power of the parametric bootstrap goodness-of-fit test: Full and limited information. *British Journal of Mathematical and Statistical Psychology.*, 56:271–288.

Toutkoushian, R. (1998). Racial and marital status differences in faculty pay. *Journal of Higher Education*, 69:513–541.

Toutkoushian, R. (1999). The status of academic women in the 1990s: no longer outsiders, but not yet equals. *The Quarterly Review of Economics and Finance*, 39:679–698.

Trinchet-Almaguer, D. y Pérez-Roses, H. (2007). An algorithm to solve the generalized inverse voronoi problem. *Revista Cubana de Ciencias Informáticas*, 1(4):58–71.

Tsiatis, A. (1980). A note on a goodness-of-fit test for the logistic regression model. *Biometrika*, 67:250–251.

van der Linden, W. y Hambleton, R. (1997). *Handbook of Modern Item Response Theory*. Springer, New York.

van Eeuwijk, F. (1995a). Linear and bilinear models for the analysis of multienvironment trials: I. an inventory of models. *Euphytica*, 84:1–7.

van Eeuwijk, F. (1995b). Multiplicative interaction in generalized linear models. *Biometrics*, 51:1017–1032.

Vicente-Galindo, P., de Noronha Vaz, T., y Nijkamp, P. (2011). Institutional capacity to dynamically innovate: An application to the portuguese case. *Technological Forecasting and Social Change*, 78(1):3 – 12.

Vicente-Villardón, J. L. (1992). Una alternativa a las técnicas factoriales basada en una generalización de los métodos biplot. tesis doctoral. universidad de salamanca. españa.

Vicente-Villardón, J. L. (2010). *MULTBIPLOT: A package for Multivariate Analysis using Biplots*. University of Salamanca.Department of Statistics.

Vicente-Villardón, J. L. (2014). *PCABiplot: Classical PCA Biplot with added features. R package version 0.2.2*. University of Salamanca.Department of Statistics.

Vicente-Villardón, J. L., Galindo, M. P., y Blázquez-Zaballos, A. (2006). Logistic biplots. *Multiple Correspondence Analysis and related methods*, pages 491–509.

Vuk, M. y Curk, T. (2006). Roc curve, lift chart and calibration plot. *Metodoloski zvezki*, 3(1):89–108.

Ward, M. (2001). The gender salary gap in british academia. *Applied Economics*, 33:1669–1681.

Warman, C., Woolley, F., y Worswick, C. (2006). The evolution of male-female wages differentials in canadian universities: 1970-2001. *Queen's Economics Department Working Paper. No. 1099. Department of Economics, Queen's University.*

Wedderburn, R. (1976). On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. *Biometrika*, 63:27–32.

Wedel, M. y Wagner, A. (2001). Factor analysis with(mixed) observed and latent variables in the exponential family. *Psychometrika.*, 66(4):515–530.

Wu, C. (1983). On the convergence properties of the em algorithm. *The Annals of Statistics*, 11:95–103.

BIBLIOGRAFÍA

Xie, X., Pendergast, J., y Clarke, W. (2008). Increasing the power: A practical approach to goodness-of-fit test for logistic regression models with continuous predictors. *Computational Statistics & Data Analysis*, 1.

Yau, K. y McGilchrist, C. (1996). Simulation study of the glmm method applied to the analysis of clustered survival data. *J. Statis, Computn Simuln*, 55:189–200.

Yen, W. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, 5:245–262.

Zerrin, A. y Greenacre, M. (2011). Biplots of fuzzy coded data. *Fuzzy Sets and Systems*, 183(1):57–71.

Zhao, J., McMorris, R., Pruzek, R., y Chen, R. (2002). The robustness of the unidimensional 3-pl irt model when applied to two-dimensional data in coputerized adaptative tests. *Paper presented at the Annual Meeting of the American Educational Research Association in New Orleans, L.A.*

Zhu, J. y Hastie, T. (2004). Classification of gene microarrays by penalized logistic regression. *Biostatistics*, 54:167–179.

Zweig, M. y Campbell, G. (1993). Receiver operating characteristic (roc) plots: a fundamental evaluation tool in clinical medicine. *Clin. Chem.*, 39(4):561–577.

Índice de figuras

Índice de tablas

Glosario

A | **B** | **C** | **E** | **F** | **G** | **I** | **J** | **L** | **M** | **N** | **O** | **P** | **R** | **S**

A

AIC Criterio de información de Akaike. 173–175, 219

B

BIC Criterio de información bayesiano. 173, 175, 219

C

CA Análisis de Correspondencias. 2, 3, 31, 70, 71, 73, 76, 116

CATPCA Análisis de Componentes Principales Categórico. 6, 99, 230

CCA Análisis de Correlación Canónica. 31

CDH Desarrollo Profesional de las Personas con un Doctorado. 81

CLB Biplot Logístico de Variables categóricas.. 225, 226, 230

CLP Puntos Categoría (Category Level Points). 54, 55, 67, 68, 75, 90, 91, 185, 192, 194, 230

CVA Análisis Canónico. 14

E

EUROSTAT Oficina de Estadística de la Unión Europea. 80, 81

F

FA Análisis Factorial. 1, 3

G

GLIM Modelo Lineal General. 212

I

INE Instituto Nacional de Estadística. 81, 82, 241

IPSFL Instituciones Privadas Sin Fines de Lucro. 118

IRT Teoría de Respuesta al Ítem. 5, 6, 32, 99–101, 115, 137, 163, 164, 211–214, 238, 239

J

JCA Análisis de Correspondencias Conjunto. 2

L

LTA Análisis de Rasgos Latentes. 5, 239

M

MCA Análisis de Correspondencias Múltiple. 2, 5, 7, 10, 29, 46, 55, 69, 70, 73, 84, 87, 90–92, 192, 194, 225, 226, 239, 245

ML Máxima Verosimilitud. 142, 143

MML Máxima Verosimilitud Marginal. 138, 140, 144

MS Escalamiento Multidimensional. 1, 240

N

NLB Biplot Logístico Nominal. 5, 10, 84, 86–88, 90–92, 194, 239

O

OECD Organización para la cooperación y el desarrollo. 80, 81

OLB Biplot Logístico Ordinal. 6, 11, 100, 198

P

PCA Análisis de Componentes Principales. 3, 4, 14, 26, 29, 31, 35, 73–75, 230, 240

PCC Porcentajes de Clasificaciones Correctas. 124

PCoA Análisis de Coordenadas Principales. 1, 19

R

RMSE Raíz del Error Cuadrático Medio. 161, 162

RMSEA Raíz del Residuo Cuadrático Medio de Aproximación. 162

ROC Curva Característica de Operación. 165, 173

S

SVD Descomposición en Valores Singulares. 3, 15–18, 71, 75, 218, 243

Apéndices

Apéndice A

Artículo enviado a la revista "Advances in Data Analysis and Classification"

Advances in Data Analysis and Classification. manuscript No.
(will be inserted by the editor)

Logistic Biplot for Nominal Data

Julio César Hernández Sánchez
José Luis Vicente-Villardón.

the date of receipt and acceptance should be inserted later

Abstract Classical biplot methods allow for the simultaneous representation of individuals (rows) and variables (columns) of a data matrix. For binary data, logistic biplots have been recently developed. When data are nominal, linear or even binary logistic biplots are not adequate and techniques such as multiple correspondence analysis (MCA), latent trait analysis (LTA) or item response theory (IRT) for nominal items should be used instead.

In this paper we extend the binary logistic biplot to nominal data. The resulting method is termed "nominal logistic biplot"(NLB), although the variables are represented as convex prediction regions rather than vectors. Using the methods from computational geometry, the set of prediction regions is converted to a set of points in such a way that the prediction for each individual is established by its closest "category point".

Then interpretation is based on distances rather than on projections. We study the geometry of such a representation and construct computational algorithms for the estimation of parameters and the calculation of prediction regions. Nominal logistic biplots extend both MCA and LTA in the sense that give a graphical representation for LTA similar to the one obtained in MCA.

Keywords Biplot · Categorical variables · Logistic responses · Latent traits · Computational geometry · Inverse voronoi problem

Mathematics Subject Classification 62H25 Factor analysis and principal components, correspondence analysis · 62H30 Classification and

Julio César Hernández Sánchez (✉)
Spanish Statistics Institute
C/Torres Villarroel, 80, 6th floor C, Salamanca, 37005, Spain
E-mail: juliocesar.hernandez.sanchez@ine.es

José Luis Vicente-Villardón
University of Salamanca
Statistics Department
C/Alfonso X el Sabio S/N, 37007, Salamanca
E-mail: villardon@usal.es

Apéndice A. Artículo enviado a "Advances in Data Analysis and Classification"

discrimination, cluster analysis · 62H99 None of the above, but in this section(Multivariate Analysis) · 68U05 Computer graphics, computational geometry

1 Introduction

The biplot method [Gabriel, 1971] is a simultaneous graphical representation of the rows and columns of a data matrix. In practice, biplot fitting occurs either by computing the singular value decomposition (SVD) of the data matrix or by performing an alternating regressions procedure [Gabriel and Zamir, 1979]. Jongman et al. [1987] or Gower and Hand [1996] fit the biplot by alternating a regression and a interpolation step, essentially equivalent to the alternating regressions. Classical biplot methods are closely related to principal components or factor analysis, two very popular techniques that are still under development [Browne and McNicholas, 2013] even though these have been used for more than one hundred years.

Graphical techniques, as biplots, for visualization of data matrices or models associated to such data, are still popular in the literature, see for example Scrucca [2013].

For data with distributions from the exponential family, Gabriel [1998], describes "bilinear regression" as a method to estimate biplot parameters, but the procedure have never been implemented and the geometrical properties of the resulting representations have never been studied. De Leeuw [2006] proposes principal components analysis (PCA) for binary data based on an alternate procedure in which each iteration is performed using iterative majorization and Lee et al. [2010] extends the procedure for sparse data matrices, none of those describe a biplot representation for binary data. Vicente-Villardón et al. [2006] propose a biplot representation based on logistic responses called "logistic biplot" that is linear, the paper studies the geometry of this kind of biplots and uses an estimation procedure that is slightly different from Gabriel's method. A heuristic version of the procedure for large data matrices in which scores for individuals are calculated with an external procedure such as PCA is described in Demey et al. [2008] and this method is called "external logistic biplot". Binary logistic biplots have been successfully applied to different data sets, see for example, Gallego-Álvarez and Vicente-Villardón [2012], Vicente-Galindo et al. [2011] or Demey et al. [2008].

When data are nominal, there are many techniques to deal with them, some see the problem from a factor analytic point of view to obtain latent factors that explain the correlation among variables, others as some kind of non-parametric approximations to explore the similarities among individuals (principal coordinates analysis (PCoA) or multidimensional scaling (MS) but there is a lack of general exploratory techniques for the simultaneous representation of individuals and variables except MCA, based on the chi-squared distance, that is not always adequate to describe similarities among individuals and correlations among variables. As we will see, it is possible to combine

the factor analytic approach with the exploratory point of view to obtain a simultaneous representation of individuals and variables (biplot) that helps to explore the information provided by the data. In this paper we propose "nominal logistic biplots" that share characteristics from the previously mentioned techniques; on the one hand it is a procedure for dimension reduction, explaining the correlation among nominal variables with a reduced number of latent factors and on the other hand it can serve as an exploratory biplot technique. Nominal logistic biplots represent the rows of a data matrix as points on a reduced dimension representation (usually 2 or 3) and variables as prediction regions (convex polygons), in the same way as is done in Gower and Hand [1996] for multiple correspondence analysis. For MCA the category points are calculated first and then the prediction regions are obtained as regions of a voronoi diagram; in this case the prediction regions are obtained first by nominal logistic regression that defines tessellations of the space; the problem is then finding the voronoi diagram that is the closest to the "logistic tessellation" and a set of generators for such a diagram are the category points, the main advantage of doing so is that the interpretation of the biplot is done in terms of distances, for each individual the predicted category is the closest to it on the biplot.

There are several candidate methods for parameter estimation:

- **Alternated generalized regressions and interpolations:** (joint maximum likelihood, [Gabriel, 1998; Vicente-Villardón et al., 2006]).
- **Marginal maximum likelihood:** (as in IRT, [Baker, 1992; Bock and Aitkin, 1981; Chalmers, 2012]).
- **External logistic biplots:** heuristic approach for big data matrices. (logistic fits on the principal coordinates, [Demey et al., 2008]).

In the context of binary logistic biplots the first two procedures are particularly useful when the number of individuals (companies) is higher than the number of variables (indicators). When there are a high number of individuals the second procedure is more stable than the others. The third method is more useful when the number of variables is higher than the number of individuals, although it can be applied in any case. In this paper we have chosen a version of the second method. Final estimation of the variable parameters where calculated using an algorithm developed for this paper, standard logistic regressions on the scores provided by multidimensional item response theory [Chalmers, 2012] or by PCoA.

In Section 2 we describe linear and logistic biplots as a basis for the development of the nominal case, that is described in section 3. This section studies the model and its main geometrical characteristics and presents an algorithm to obtain the category points for each variable. Section 4 applies the nominal logistic biplot to a classical set of data and section 5 concludes the paper with a discussion and some suggestions for further research.

Apéndice A. Artículo enviado a "Advances in Data Analysis and Classification"

2 Linear and binary logistic biplots

2.1 Classical linear biplots and the singular value decomposition

Let $\mathbf{X}_{I\times J}$ be a data matrix containing the measures of J variables (continuous) on I individuals. A S-dimensional biplot is a graphical representation of a data matrix $\mathbf{X}$ by means of markers (points or vectors) $\mathbf{a}_1,\ldots,\mathbf{a}_I$ for its rows and markers $\mathbf{b}_1,\ldots,\mathbf{b}_J$ for its columns, in such a way that the product $\mathbf{a'}_i\mathbf{b}_j$ approximates the element x_{ij} as close as possible. Arranging the markers as row vectors in two matrices $\mathbf{A}$ and $\mathbf{B}$, the approximation of $\mathbf{X}$ can be written as $\mathbf{X}\approx\mathbf{AB'}$. Although the classical biplot is well known, we include here a short description, in terms of alternating regressions, related to our proposal.

The most typical way to obtain the biplot is from the singular value decomposition. Let R be the rank of $\mathbf{X}$, there exists a factorization of the form:

$$\mathbf{X}=\mathbf{U}\Lambda\mathbf{V'}=\sum_{r=1}^{R}\lambda_r\mathbf{u}_r\mathbf{v'}_r, \tag{1}$$

where $\mathbf{U}$ is an $I\times R$ unitary matrix, $\mathbf{\Lambda}$ is an $R\times R$ diagonal matrix with non-negative real numbers on the diagonal, and $\mathbf{V}$ an $J\times R$ unitary matrix. Such a factorization is called the singular value decomposition of $\mathbf{X}$. The diagonal entries λ_r of $\mathbf{\Lambda}$ are known as the singular values of $\mathbf{X}$, and are placed in decreasing order, and the columns $\mathbf{u}_r$ and $\mathbf{v}_r$ of $\mathbf{U}$ and $\mathbf{V}$ are known as left and right singular vectors. The SVD is closely related to the *eigen decomposition*, the columns of $\mathbf{U}$ are the eigenvectors of $\mathbf{XX'}$, the columns of $\mathbf{V}$ the eigenvectors of $\mathbf{X'X}$ and the diagonal elements of $\mathbf{\Lambda}$ are the squared roots of the non-null eigenvalues of both matrices.

It is known that the best *S-rank* approximation of $\mathbf{X}$ is given by its first S singular values and vectors:

$$\mathbf{X}\cong\sum_{s=1}^{S}\lambda_s\mathbf{u}_s\mathbf{v'}_s=\mathbf{U}_{(S)}\Lambda_{(S)}\mathbf{V'}_{(S)}. \tag{2}$$

From he SVD, it is easy to obtain a factorization in the biplot form with the desired restriction taking:

$$\mathbf{A}=\mathbf{U}_{(S)}\Lambda^{\gamma}_{(S)},\quad \mathbf{B}=\mathbf{V}^{1-\gamma}_{(S)}, \tag{3}$$

with $0\leq\gamma\leq 1$, as row and column coordinates respectively. This will be referred to as the PCA-biplot or classical biplot. For example, with $\gamma=1$, $\mathbf{A}$ are the coordinates of individuals on the principal components and $\mathbf{B}$ are the eigenvectors of the covariance matrix.

There is another way of obtaining biplots from alternated regressions. If we consider the row markers $\mathbf{A}$, as fixed, the column markers can be computed by regression:

$$\mathbf{B'}=(\mathbf{A'A})^{-1}\mathbf{A'X}. \tag{4}$$

In the same way, fixing $\mathbf{B}$, $\mathbf{A}$ can be obtained as:

$$\mathbf{A}' = (\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'\mathbf{X}'. \tag{5}$$

Alternating the steps (4) and (5) the product converges to the same subspace generated by SVD. The algorithm can then be completed with an orthogonalization step to ensure the uniqueness of its solution. The regressions in (1) and (2) can be separated for each row and column of the data matrix. This symmetrical process is commonly used to adjust bilinear (or bi-additive) models with symmetrical roles for rows and columns. For a data matrix of individuals by variables, the roles of rows and columns are non-symmetrical, nevertheless the algorithm is still valid and is interpreted as a two-step process, alternating a regression step and an interpolation step. The regression step adjusts a separate linear regression for each column (variable) and the interpolation step interpolates an individual using the column markers as the reference. The geometry of the interpolation step is described in Gower and Hand [1996, page 13].

2.2 Logistic biplots for binary data

Let $\mathbf{X}_{I\times J}$ be a data matrix in which the rows correspond to I individuals and the columns to J binary characters. Let $\pi_{ij} = E(x_{ij})$ be the expected probability that the character j be present at individual i, and x_{ij} the observed probability, either 0 or 1, resulting in a binary data matrix. The S-dimensional logistic biplot in the *logit* scale is formulated as:

$$logit(\pi_{ij}) = \log(\frac{\pi_{ij}}{1-\pi_{ij}}) = b_{j0} + \sum_{s=1}^{S} b_{js}a_{is} = b_{j0} + \mathbf{a'}_i\mathbf{b}_j, \tag{6}$$

where a_{is} and b_{js}, $(i = 1,\ldots,I; j = 1,\ldots,J; s = 1,...,S)$, are the model parameters used as row and column markers respectively. The model is a generalized (bi)linear model having the *logit* as a link function. In terms of probabilities rather than *logits*:

$$\pi_{ij} = \frac{e^{b_{j0}+\sum_k b_{jk}a_{ik}}}{1+e^{b_{j0}+\sum_k b_{jk}a_{ik}}}. \tag{7}$$

In matrix form:

$$logit(\mathbf{\Pi}) = \mathbf{1}_I\mathbf{b'}_0 + \mathbf{AB}', \tag{8}$$

where $\mathbf{\Pi}$ is the matrix of expected probabilities, $\mathbf{1}_I$ is a vector of ones and $\mathbf{b}_0 = (b_{j0})$ is the vector containing intercepts that have been added because it is not possible to center the data matrix in the same way as in linear biplots. The intercepts are the displacements of centroids in the same way as it is the first ordination axis in correspondence analysis (CA). The model is a latent trait model for binary data, the row coordinates being the scores of individuals

on the latent trait. Although the biplot in the logit scale may be useful, it would be more interpretable in a probability scale.

The points predicting different probabilities are on parallel straight lines on the biplot; this means that predictions on the logistic biplot are made in the same way as on the linear biplots, i. e., projecting a row marker $\mathbf{a}_i = (a_{i1}, a_{i2})$ onto a column marker $\mathbf{b}_j = (b_{j1}, b_{j2})$. (See Vicente-Villardón et al. [2006], Demey et al. [2008]).

The model in (6) is also a latent trait or item response theory model, in that ordination axes are considered as latent variables that explain the association between the observed variables. In this framework we suppose that individuals respond independently to variables, and that the variables are independent for given values of the latent traits. With these assumptions the likelihood function is:

$$\text{Prob}(x_{ij}|\,(\mathbf{b}_0, \mathbf{A}, \mathbf{B})) = \prod_{i=1}^{I}\prod_{j=1}^{J} \pi_{ij}^{x_{ij}}(1-\pi_{ij})^{1-x_{ij}}. \tag{9}$$

Taking the logarithm of the likelihood function yields:

$$L = \log \text{Prob}(x_{ij}|\,(\mathbf{b}_0, \mathbf{A}, \mathbf{B})) = \sum_{i=1}^{I}\sum_{j=1}^{J} [x_{ij}\log(\pi_{ij}) + (1-x_{ij})\log(1-\pi_{ij})]. \tag{10}$$

For $\mathbf{A}$ fixed, (10) can be separated into J parts, one for each variable:

$$L = \sum_{J=1}^{J} L_j = \sum_{J=1}^{J}\left(\sum_{i=1}^{I} [x_{ij}log(\pi_{ij}) + (1-x_{ij})log(1-\pi_{ij})]\right). \tag{11}$$

Maximizing each L_j is equivalent to performing a standard logistic regression using the j-th column of $\mathbf{X}$ as a response and the columns of $\mathbf{A}$ as regressors. In the same way the probability function can be separated into several parts, one for each row of the data matrix, $L = \sum_{i=1}^{I} L_i$.

Binary logistic biplots can be calculated using the package MULTBIPLOT [Vicente-Villardón, 2010] as a stand-alone application implemented in Matlab.

3 Logistic biplot for nominal data

3.1 Formulation

Let $\mathbf{X}_{I\times J}$ be a data matrix containing the values of J nominal variables, each with K_j $(j = 1, \ldots, J)$ categories, for I individuals, and let $\mathbf{G}_{I\times L}$ be the corresponding indicator matrix with $L = \sum_j K_j$ columns. The last (or the first) category of each variable will be used as a baseline. Let $\pi_{ij(k)}$ denote the expected probability that the category k of variable j be present at individual

i. A multinomial logistic latent trait model with S latent traits, states that the probabilities are obtained as:

$$\pi_{ij(k)} = \frac{e^{b_{j(k)0}+\sum_{s=1}^{S} b_{j(k)s}a_{is}}}{\sum_{l=1}^{K_j} e^{b_{j(l)0}+\sum_{s=1}^{S} b_{j(l)s}a_{is}}}, (k = 1, \ldots, K_j). \quad (12)$$

Using the last category as a baseline in order to make the model identifiable, the parameter for that category are restricted to be 0, i.e., $b_{j(K_j)0} = b_{j(K_j)s} = 0$, $(j = 1, \ldots, J; \quad s = 1, \ldots, S)$.The model can be rewritten as:

$$\pi_{ij(k)} = \frac{e^{b_{j(k)0}+\sum_{s=1}^{S} b_{j(k)s}a_{is}}}{1+\sum_{l=1}^{K_j-1} e^{b_{j(l)0}+\sum_{s=1}^{S} b_{j(l)s}a_{is}}}, (k = 1, \ldots, K_j - 1). \quad (13)$$

With this restriction we assume that the log-odds of each response (relative to the last category) follows a linear model:

$$\log\left(\frac{\pi_{ij(k)}}{\pi_{ij(K_j)}}\right) = b_{j(k)0} + \sum_{s=1}^{S} b_{j(k)s}a_{is} = b_{j(k)0} + \mathbf{a'}_i\mathbf{b}_{j(k)},$$

where a_{is} and $b_{j(k)s}$ $(i = 1, \ldots, I; \quad j = 1, \ldots, J; \quad k = 1, \ldots, K_j - 1; \quad s = 1, \ldots, S)$ are the model parameters. In matrix form:

$$\mathbf{O} = \mathbf{1}_I\mathbf{b'}_0 + \mathbf{AB'}, \quad (14)$$

where $\mathbf{O}_{I\times(L-J)}$ is the matrix containing the expected log-odds, defines a biplot for the odds. Although the biplot for the odds may be useful, it would be more interpretable in terms of predicted probabilities and categories. This biplot will be called "nominal logistic biplot", and it is related to the latent nominal models in the same way as classical linear biplots are related to factor or principal components analysis or binary logistic biplots are related to the IRT or LTA for binary data.

The points predicting different probabilities are no longer on parallel straight lines (see Fig 1 with the response surfaces); this means that predictions on the logistic biplot are not made in the same way as in the linear biplots, the surfaces now define prediction regions for each category as shown in the graph.

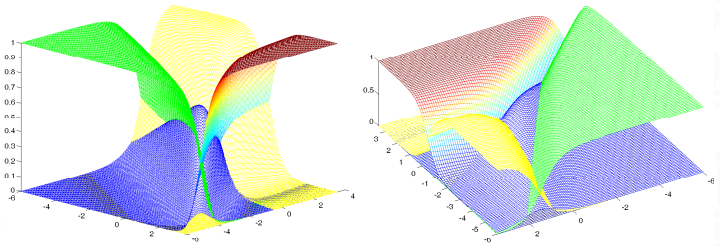
3.2 Geometry

Suppose we have a two-dimensional representation in which the row coordinates are defined by the first two columns of $\mathbf{A}$ in (14), let's call $\mathcal{L}$ the space generated by those columns. Equations (12), (13) and (14) define a set of probability response surfaces (one for each category and each variable) (Fig 1) that

are no longer sigmoid as in the binary case (Vicente-Villardón et al. [2006]). This means that the level curves are no longer straight lines and then, prediction of probabilities is not made by projection as in the usual linear biplots. Figure 2b shows the level curves for probability 0.5 and a hypothetical variable with four categories. We will show that in this case the predicted probabilities, for each variable, define a set of convex polygons that can be interpreted as "prediction" regions in the same way as in Gower and Hand [1996]. For each variable there are as many regions as categories and each one is formed by the set points in with the expected probability for a category is higher than the probability for the rest of categories. Let $\mathcal{R}_k$ denote the region for category j, then it can be defined as:

$$\mathcal{R}_k = \left\{ \mathbf{a}_h = (a_{h1}, a_{h2}) \in \mathcal{L} / \pi_{hj(k)} \geq \pi_{hj(m)}, \forall m \neq k; k, m = (1, \ldots, K_j) \right\}.$$

The prediction regions for a hypothetical variable with four categories are shown in Fig 2c. It is immediate to see that the prediction regions are closely related to the level curves.



(a) (b)

Fig. 1: Response surfaces of the nominal logistic model (a,b) , with 4 categories and 2 explanatory variables.

It has to be noted that there are some cases in which some of the categories are never predicted, those will be termed **hidden categories** and should be taken into account to construct the final representation.

3.3 Obtaining "prediction regions"

In the following paragraphs we will describe a procedure to obtain the prediction regions using methods taken from the computational geometry. The set of convex polygons predicting each category form a tessellation of the plane. Each cell of the tessellation is delimited by a set of straight lines that correspond

to points that have equal probabilities for two of the categories of the variable (the edges). We consider each variable j ($j = 1, \ldots, J$) separately. Each pair of response surfaces defined by (12) intersect in a straight line that, projected onto the space of predictors, is the set of points in which the probability of both categories is the same. Those lines are the candidates to be the edges of the convex polygons defining the prediction regions. That is, we search for the set of points $\mathcal{E}_{kl}$ in $\mathcal{L}$ such that the pair of categories k and l ($k, l = 1, \ldots, K_j$), that have the same expected probability $\pi_{ij(k)} = \pi_{ij(l)}$ i. e., $\mathcal{E}_{kl}$ is the set of points verifying:

$$\frac{e^{b_{j(k)0}+\sum_{s=1}^{2} b_{j(k)s}a_s}}{\sum_{m=1}^{K_j} e^{b_{j(m)0}+\sum_{s=1}^{2} b_{j(m)s}a_s}} = \frac{e^{b_{j(l)0}+\sum_{s=1}^{2} b_{j(l)s}a_s}}{\sum_{m=1}^{K_j} e^{b_{j(m)0}+\sum_{s=1}^{2} b_{j(m)s}a_s}}. \tag{15}$$

Then

$$b_{j(k)0} + \sum_{s=1}^{2} b_{j(k)s}a_s = b_{j(l)0} + \sum_{s=1}^{2} b_{j(l)s}a_s$$

or

$$(b_{j(k)1} - b_{j(l)1})a_1 + (b_{j(k)2} - b_{j(l)2})a_2 = (b_{j(l)0} - b_{j(k)0}).$$

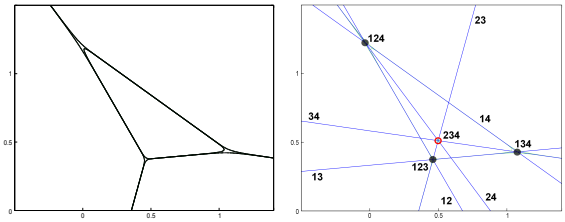
The above equation can be written as:

$$a_2 = \frac{(b_{j(l)0} - b_{j(k)0})}{(b_{j(k)2} - b_{j(l)2})} - \frac{(b_{j(k)1} - b_{j(l)1})}{(b_{j(k)2} - b_{j(l)2})}a_1,$$

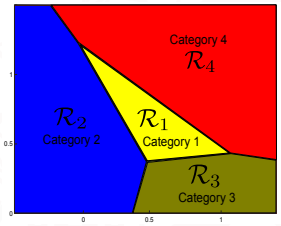
where a_1 and a_2 are generic coordinates on the dimensions of $\mathcal{L}$. Each variable j has $\binom{K_j}{2}$ of such lines as shown in Fig 2b for a hypothetical example with four categories.

Except for degenerate cases, any two lines with one index in common, $\mathcal{E}_{kl}$ and $\mathcal{E}_{km}$, intersect in a point $\mathcal{P}_{klm}$. The $\binom{K_j}{3}$ of such points are the candidates to be the vertices of the tessellation. Point $\mathcal{P}_{klm}$ is a vertex of the tessellation if there is not a $t \notin \{k, l, m\}$ such that $\pi_{(\mathcal{P}_{klm})t} > \pi_{(\mathcal{P}_{klm})r}$ for $r \in \{k, l, m\}$, where $\pi_{(\mathcal{P}_{klm})t}$ is the expected probability of category t at point $\mathcal{P}_{klm}$, i.e., the expected probability for one of the categories involved is the highest. If a point is a vertex of the tessellation it is termed a **real point**, otherwise it is a **virtual point**. Degenerate cases may have parallel lines but this is extremely unlikely to occur. The prediction regions $\mathcal{R}_k$ are delimited by all the lines $\mathcal{E}_{kl}$ with index k and its vertices are all the points $\mathcal{P}_{klm}$ with the index k. A category is hidden when its index is not present in any of the real points. The region of the hidden category is omitted in the representation. We now define the meaning of **join** two points $\mathcal{P}_{klm}$ and $\mathcal{P}_{kln}$ as follows (see Fig 3):

1. Two real points should be joined, if they have two indices in common, following the line $\mathcal{E}_{kl}$.
2. Two virtual points are never joined.
3. A virtual point and a real point are joined along the line $\mathcal{E}_{kl}$, starting from the real point and away from the virtual point.



(a) (b)



(c)

Fig. 2: Geometry for a two-dimensional solution and a hypothetical variable with four categories: Level curves of the response surfaces for $p = 0.5$ (a), lines of equal probability for each pair of categories and their intersection points (candidates for edges and vertices of the tessellation) (b), and tessellation of the plane defined by the prediction regions (c).

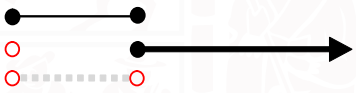


Fig. 3: Definition of join for constructing the tessellation. Black •: real point; Red ⊙: virtual point

Now it is easy to adapt the algorithm described in Gower and Hand [1996] to construct the tessellation generated by the probability responses:

1. Compute the coordinates of all $\binom{K_j}{3}$ points $\mathcal{P}_{klm}$.
2. Decide if the point is real or virtual.
3. Join all pairs of points sharing two suffices, interpreting “join” as described before.

The procedure is different from that in Gower and Hand [1996] in two aspects: they start from a set of points $\mathcal{C}_k, k = (1, \ldots, K_j)$ that they call “category points” arising from a MCA with some modifications, and then construct

the tessellation from those points using distances; we don't have the category points and use probabilities rather than distances. The tessellation based on distances is called a voronoi diagram and is quite a popular tool in a discipline called "computational geometry"; in this diagram the space is divided into a set of polygons or regions $\mathcal{R}_k$ in such a way that points in the region are closest to $\mathcal{C}_k$ than to any other point. The main advantage of doing so is that it provides a simple interpretation of the representation of row and column markers of the data matrix, the predicted category for each point is the category corresponding to its closest "category point". Representing points rather than "regions" produces a much cleaner and easier way to interpret the graph. We have the regions but not the points and, although from a formal point of view our problem is solved, and we have a simultaneous representation of individuals and variables, it would be more convenient to have also a set of "category points" to interpret the biplot in terms of distances. Let's call this set of points $\mathcal{C}_{j(k)}, j = (1, \ldots, J), k = (1, \ldots, K_j)$. This would be a fundamental contribution of our research becase the interpretation of distances among row and column points is simple and it is not an intrinsic property of most multivariate techniques as MCA. Three problems arise:

1. Is our tessellation a voronoi diagram?.
2. If not, is there any way to approximate it by its closest voronoi tessellation?.
3. Given a voronoi tessellation, is it possible to obtain a set of generators for it?.

In the next section we describe a procedure to obtain the generators given a tessellation.

3.4 Obtaining generators of the tessellation

The problem of testing if any convex tessellation consists of voronoi polygons and if so, obtain a set of centers or generators of the voronoi diagram, has been studied for example by Hartvigsen [1992] and Evans and Jones [1987]. The first paper establishes a set of equations of slope and distance that a tesselation must hold to be voronoi in such a way that solving a linear system it is possible to obtain the set of centers (Fig 4). Let's see it in more detail.

First consider the following result (we will omit the index j of the variable for simplicity): a tessellation of K polygons or convex regions $\mathcal{R}_k, k = 1, \ldots, K$ is a voronoi diagram with centres $\mathcal{C}_k = (x_k, y_k), k = (1, \ldots, K)$ iff $\mathcal{R}_k = \{(x, y) : (x - x_k)^2 + (y - y_k)^2 \leq (x - x_l)^2 + (y - y_l)^2, \forall l \neq k\}$, i.e., each polygon of the tessellation is the set of points that are nearer to its center than to any center of other polygon.

If we consider two adjacent polygons, $\mathcal{R}_l$ y $\mathcal{R}_m$, whose common edge is E_{lm} with equation $y = s_i x + b_i$, and contain the vertices (u_p, v_p) and (u_q, v_q), let $\mathcal{C}_l = (x_l, y_l)$ and $\mathcal{C}_m = (x_m, y_m)$ be the voronoi centers of the regions (our "category points"). The equations of slope and distance are:

$$\frac{(y_l - y_m)}{(x_l - x_m)} = \frac{-1}{s_i} \tag{16}$$

$$-s_i x_l + y_l - b_i = -s_i x_m + y_m - b_i, \tag{17}$$

where $s_i = \frac{(v_p - v_q)}{(u_p - u_q)}$ and $b_i = s_i u_p - v_p$.

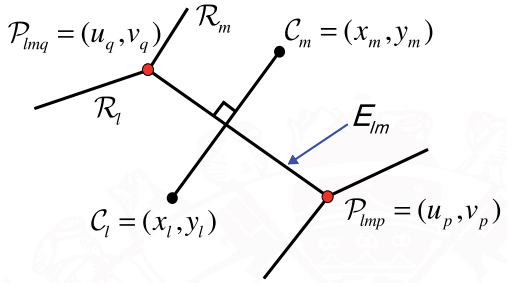


Fig. 4: Centers $\mathcal{C}_l = (x_l, y_l)$ and $\mathcal{C}_m = (x_m, y_m)$ are equidistant from the edge they share E_{lm} (16) and both lie on the line perpendicular to E_{lm} (17).

Those equations with, for example k edges and n polygons form a linear system with $2k$ equations and $2n$ unknowns, that can be solved by least squares. In matrix form the system is:

$$\mathbf{Bx} = \mathbf{0}$$
$$\mathbf{Ax} = \mathbf{b},$$

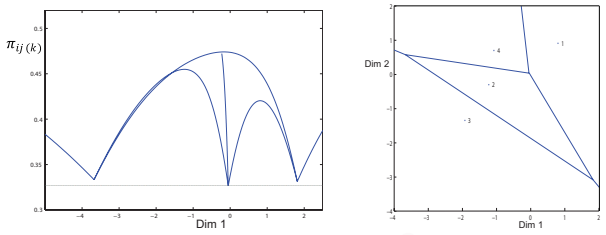
with $\mathbf{x} = [x_1, y_1, \ldots, x_n, y_n]'$, $\mathbf{b} = -2[b_1, \ldots, b_k]'$. Matrices $\mathbf{A}$ and $\mathbf{B}$ are sparse but that is not a problem because the number of categories is usually small. Calculations to obtain a solution are based on three algorithms that can produce different centers in the case that the polygons of the tessellation are not voronoi. The three methods are:

Algorithm 1: minimize the conditions of distance and slope, that is, search for $Min \left\| \begin{bmatrix} \mathbf{A} \\ \mathbf{B} \end{bmatrix} \mathbf{x} - \begin{bmatrix} \mathbf{b} \\ \mathbf{0} \end{bmatrix} \right\|^2$, with $\|.\|^2$ the euclidean norm.

Algorithm 2: minimize $\|\mathbf{Bx}\|^2$, subject to $\mathbf{Ax} = \mathbf{b}$.

Algorithm 3: minimize $\|\mathbf{Ax} - \mathbf{b}\|^2$, subject to $\mathbf{Bx} = \mathbf{0}$.

In practice, the main problem with the linear systems is the instability of the algorithms due to the ill conditioning of the matrices. Schoenberg et al. [2003] treat the problem and propose some alternatives to improve the stability of the final solution. Evans and Jones [1987] also proposes a measure of the goodness of fit, i.e., a measure of how near is the tessellation from a true voronoi diagram. For the hypothetical example in Fig 1 we show the result of inverting a tessellation obtained from the logistic response in Fig 5, for this case the tessellation is very close to a voronoi diagram.



(a) (b)

Fig. 5: Frontal view of the intersections among the 4 response curves obtained from the nominal logistic regression (a), and the generated tessellation with the result of the algorithm for inversion (b) with the category points.

3.5 Paramater estimation

Although the nominal case doesn't share the geometrical properties with the binary case, the alternated algorithm described in Vicente-Villardón et al. [2006], can be easily extended replacing the binary logistic regressions by multinomial logistic regressions. The problem with this approach is that the parameters for the individuals can not be estimated when the individual has 0 or 1 in all the variables for the binary case, or all the responses are at the baseline category for the nominal case. In this paper we use a procedure that is similar to the alternated regressions method, except that the interpolation step is "eliminated" by considering the row parameters as incidental. The technique assumes that the scores for individuals are random effects sampled from some larger distribution. The estimation procedure is an EM-algorithm that uses the Gauss-Hermite quadrature to approximate the integrals, considering the individual scores as missing data. More details of similar procedures can be found in Bock and Aitkin [1981] or Chalmers [2012].

The likelihood function is:

$$M\left(\mathbf{G}|\,\mathbf{b}_0,\mathbf{A},\mathbf{B}\right)=\prod_{i=1}^{I}\prod_{j=1}^{J}\prod_{k=1}^{K_j}\pi_{ij(k)}^{g_{ij(k)}},$$

where $g_{ij(k)}=1$ if individual i chooses category k of item j and $g_{ij(k)}=0$ otherwise. The log-likelihood is:

$$L\left(\mathbf{G}|\,\mathbf{b}_0,\mathbf{A},\mathbf{B}\right)=\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K_j}g_{ij(k)}\log\left(\pi_{ij(k)}\right). \tag{18}$$

Apéndice A. Artículo enviado a "Advances in Data Analysis and Classification"

If the parameters $\mathbf{A}$ for individuals where known, the log-likelihood could be separated into J parts, one for each variable:

$$L\left(\mathbf{G}|\,\mathbf{b}_0,\mathbf{B}\right)=\sum_{j=1}^{J}L_j(\mathbf{G}|\,\mathbf{b}_{j0},\mathbf{B}_j)=\sum_{j=1}^{J}\left(\sum_{i=1}^{I}\sum_{k=1}^{K_j}g_{ij(k)}\log\left(\pi_{ij(k)}\right)\right),\quad(19)$$

where $\mathbf{b}_{j0}$ and $\mathbf{B}_j$ are the submatrices of parameters for the jth variable. Maximizing the log-likelihood is equivalent to maximizing each part, i.e., obtaining the parameters for each variable separately. Maximizing each L_j is equivalent to performing a multinomial logistic regression using the jth column of $\mathbf{X}$ as response and the columns of $\mathbf{A}$ as predictors. We do not describe logistic regression here because it is a very well known procedure. It is also well-known that when the individuals for different categories are separated (or quasi-separated) on the space spanned by the explanatory variables, the maximum likelihood estimators don't exist (or are unstable). Because we view the biplot as a procedure to classify the set of individuals, it is probable that, for some variables, the individuals having different categories are completely separated. When this occurs the estimators obtained from the multinomial logistic regression tend to infinity. This is known as "*separation problem in logistic regression*". The problem of the existence of the estimators in logistic regression can be seen in Albert and Anderson [1984] and a solution for the binary case, based on Firth's method [Firth, 1993] is proposed by Heinze and Schemper [2002]. The extension to nominal logistic model was made by Bull et al. [2002]. All the procedures were initially developed to remove the bias but work well to avoid the problem of separation. Here we have chosen a simpler solution based on ridge estimators for logistic regression [Le Cessie and Van Houwelingen, 1992].

Rather than maximizing $L_j(\mathbf{G}|\,\mathbf{b}_{j0},\mathbf{B}_j)$ we maximize:

$$L_j(\mathbf{G}|\,\mathbf{b}_{j0},\mathbf{B}_j)-\lambda\left(\|\mathbf{b}_{j0}\|+\|\mathbf{B}_j\|\right).\quad(20)$$

where $\|.\|$ indicates the squared norm.

We don't describe here the procedure in great detail because it is also a standard procedure. Changing the values of λ we obtain slightly different solutions not affected by the separation problem.

In the same way, if parameters for variables were known, the log-likelihood could be separated into I parts, one for each individual:

$$L\left(\mathbf{G}|\,\mathbf{A}\right)=\sum_{i=1}^{I}L_i(\mathbf{G}|\,\mathbf{a}_i)=\sum_{i=1}^{I}\left(\sum_{j=1}^{J}\sum_{k=1}^{K_j}g_{ij(k)}\log\left(\pi_{ij(k)}\right)\right).$$

To maximize each part we could use Newton-Raphson with a penalization as before. Rather than that we will use expected a posteriori estimators for the individual markers. For each individual (or response pattern) $\mathbf{g}_i$, the likelihood is:

$$M_\ell\left(\mathbf{g}_i|\,\mathbf{b}_0,\mathbf{a}_i,\mathbf{B}\right)=\prod_{j=1}^{J}\prod_{k=1}^{K_j}\pi_{ij(k)}^{g_{ij(k)}}.$$

Assuming a distributional form $g(\mathbf{a})$ (multivariate normal, for example) the marginal distribution becomes:

$$P_l\left(\mathbf{b}_0, \mathbf{B} | \, \mathbf{g}\right) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} M_\ell\left(\mathbf{g}_i | \, \mathbf{b}_0, \mathbf{a}_i, \mathbf{B}\right) g(\mathbf{a}) d\mathbf{a},$$

and the observed likelihood:

$$M\left(\mathbf{b}_0, \mathbf{B} | \, \mathbf{G}\right) = \prod_{i=1}^{I} \left[\int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} M_\ell\left(\mathbf{g}_i | \, \mathbf{b}_0, \mathbf{a}_i, \mathbf{B}\right) g(\mathbf{a}) d\mathbf{a} \right].$$

We approximate the integral by S-dimensional Gauss-Hermite quadrature:

$$\tilde{P}_l = \sum_{qS=1}^{Q} \ldots \sum_{q1=1}^{Q} M_\ell\left(\mathbf{g}_\ell | \, \mathbf{b}_0, \mathbf{Y}, \mathbf{B}\right) g(y_{q1}) \ldots g(y_{qS}).$$

The multivariate S-dimensional quadrature, $\mathbf{Y}$, has been obtained as the product of S unidimensional quadratures $(y_1, \ldots, y_Q)$ with Q nodes each. Then the marginal expected a posteriori score for individual i at dimension s, a_{is}, is:

$$E(a_s / \mathbf{g}_\ell) = \frac{\sum_{q=1}^{Q} y_q M_\ell\left(\mathbf{g}_\ell | \, \mathbf{b}_0, \mathbf{Y}, \mathbf{B}\right) g(y_q)}{\tilde{P}_l}.$$

4 Case Study

Table 1, taken from Gower and Hand [1996], shows the observations of four variables observed on twenty farms from the dutch island of Terschelling. This table is reported in Jongman et al. [1987] and it is part of a much larger survey. It is concerned with environmental factors and different forms of farm management. We have chosen this data because it has been previously analysed in literature and can serve as a comparison with the methods proposed here.

The variables are:

- Moisture class, with 5 levels, although level 3 does not occur in the data. Levels are labelled M1, M2, M4 and M5.
- Grassland management type, with 4 levels (standard farming (SF), biological farming (BF), hobby farming (HF) and nature conservation management(NM))
- Grassland use, with three levels: (production(U1), intermediate(U2) and grazing(U3))
- Manure class, with 5 levels labelled C0, C1, C2, C3 and C4. The variable is probably ordinal because the levels assume an increasing level of manure but it will be treated as nominal here.

Apéndice A. Artículo enviado a "Advances in Data Analysis and Classification"

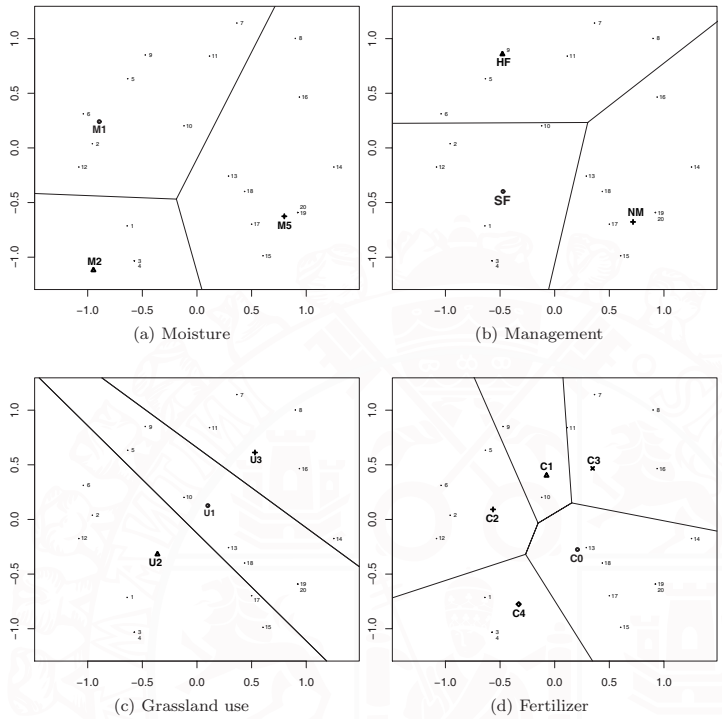
Table 1: Data on four variables observed at 20 farms on the island of Terschelling.

Farm number	Moisture class	Grassland management type	Grassland use	Manure class
1	1	SF	2	4
2	1	BF	2	2
3	2	SF	2	4
4	2	SF	2	4
5	1	HF	1	2
6	1	HF	2	2
7	1	HF	3	3
8	5	HF	3	3
9	4	HF	1	1
10	2	BF	1	1
11	1	BF	3	1
12	4	SF	2	2
13	5	SF	2	3
14	5	NM	3	0
15	5	NM	2	0
16	5	SF	3	3
17	2	NM	1	0
18	1	NM	1	0
19	5	NM	1	0
20	5	NM	1	0

The prediction regions obtained from the proposed algorithm together with the category points associated to them, are shown in Fig 6.

The four graphs could be superimposed although the resulting image would be almost unreadable (Fig 7) even with only four variables; with more variables the interpretation would be very complicated. The proposed procedure for obtaining a set of category points for each variable allows for a much simpler and easy to interpret representation. The final result is shown in Fig 8. We can see that farms having "nature management" (NM) are at the areas with higher moisture (M5), zero fertilizer (CO) and production (U1). Farms with "scientific management" (SF) are at the region with moisture M1 and M2, high values of fertilizer (C4) and intermediate grassland use (U2). Hobby farms (HF) are associated to dry places (M1), low use of fertilizer (C1) and a tendency toward U3. Farms of type BF are hidden on the prediction model because the probability of that category is never higher than the rest.

In order to compare the proposed method with MCA as in Gower and Hand [1996], and some alternatives for estimation described here, we have estimated the model parameters using our modification of the EM algorithm and the mirt R package [Chalmers, 2012] with an additional multinomial logistic regression. The prediction regions obtained for our method produce 14 incorrect classifications against the 21 obtained by MCA and the 31 by mirt (see Table 2). The table shows also true and predicted categories for all the data. There are no hidden categories for variable "Manuring" but for "Moisture" and "Management", categories M4 and BF, respectively, are hidden. The last



(a) Moisture (b) Management

(c) Grassland use (d) Fertilizer

Fig. 6: The prediction regions for each of the four variables, as given by NLB.

value is present in farmers 2, 10 and 11 and none of the methods is able to predict it correctly.

If we analyse the combined prediction regions for all the variables with EM parameter estimation, we can observe in Fig 7 that there are 28 separate convex regions. Except the region containing farms 13, 18, 19 and 20, most of the regions are small and have less points inside, emphasizing the richness of the technique for interpreting data. In the study described by Gower and Hand [1996], there were 16 different regions for MCA but only three were clearly populated, so we obtain a finer classification of the farms.

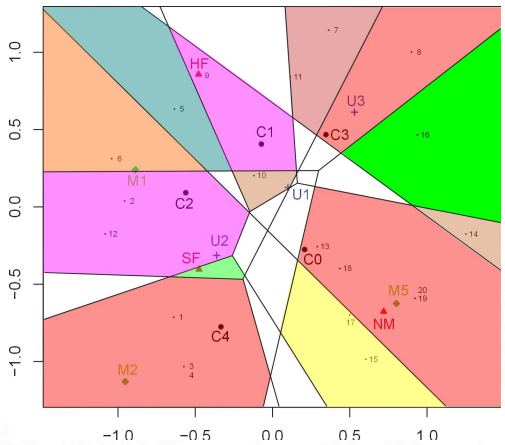


Fig. 7: Four tesselations superimposed.

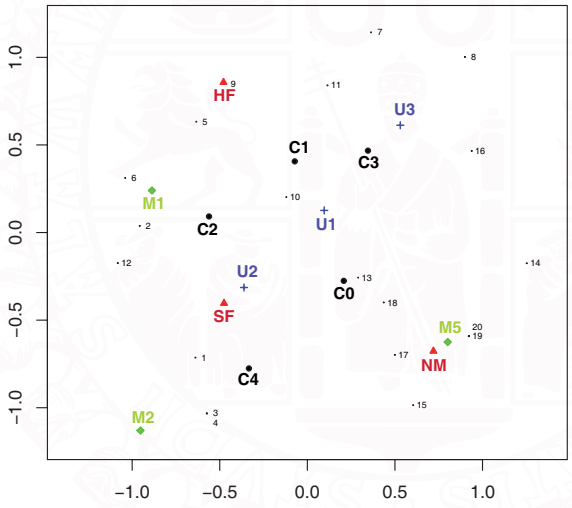


Fig. 8: Two-dimensional NLB of the categorical variables shown in Table 1.

Table 2: Predictions for the categorical variables for Table 1 given by a two-dimensional approximation. T denotes the true value given in Table 1, and MCA, Mirt, OP are the predictions using the row coordinates estimated by MCA, Mirt and Alternated Method(AM).

Farm	Moisture				Management				Grassland				Manuring			
	T	MCA	Mirt	AM	T	MCA	Mirt	AM	T	MCA	Mirt	AM	T	MCA	Mirt	AM
1	1	2*	2*	2*	1	1	1	1	2	2	2	2	4	4	4	4
2	1	1	5*	1	2	3*	3*	1*	2	2	1*	2	2	2	0*	2
3	2	2	2	2	1	1	1	1	2	2	2	2	4	4	4	4
4	2	2	2	2	1	1	1	1	2	2	2	2	4	4	4	4
5	1	1	5*	1	3	3	3	3	1	3*	1	1	2	1*	0*	2
6	1	1	5*	1	3	3	3	3	2	2	1*	2	2	2	0*	2
7	1	1	1	1	3	3	3	3	3	1*	3	3	3	1*	3	3
8	5	1*	1*	5	3	3	3	3	3	1*	3	3	3	3	3	3
9	4	1*	1*	1*	3	3	1*	3	1	3*	2*	1	1	1	3*	1
10	2	1*	5*	1*	2	3*	4*	1*	1	1	1	1	1	1	0*	1
11	1	1	5*	1	2	3*	3*	3*	3	3	3	3	1	1	2*	3*
12	3	1*	5*	1*	1	1	3*	1	2	2	3*	2	2	2	3*	2
13	5	2*	1*	5	1	1	1	4*	2	2	2	1*	3	4*	3	0*
14	5	5	5	5	4	4	4	4	3	1*	1*	3	0	0	0	0
15	5	5	5	5	4	4	4	4	2	1*	1*	2	0	0	0	0
16	5	5	1*	5	1	1	1	4*	3	2*	3	3	3	3	3	3
17	2	5*	5*	5*	4	4	4	4	1	1	1	1	0	0	0	0
18	1	5*	5*	5*	4	4	4	4	1	1	1	1	0	0	0	0
19	5	5	5	5	4	4	4	4	1	1	1	1	0	0	0	0
20	5	5	5	5	4	4	4	4	1	1	1	1	0	0	0	0
Errors	0	8	13	6	0	3	5	5	0	7	6	1	0	3	7	2

5 Conclusions and discussion

In the preceding sections we have proposed a biplot method for nominal data in which the individuals are represented as points in a low-dimensional subspace and the variables are represented as "prediction regions" or "category points" for the categories of each variable. Prediction regions are convex polygons that divide the representation space into as many regions as categories of the variable, except if there is some hidden category, and then define a tessellation of the space that, conveniently approximated by a voronoi diagram, provides a set of generators that can be considered as category points. The proposed representation is interpreted in terms of distances in the sense that the category predicted for each individual is defined by the closest category point. Although it is not described here in detail, linear biplots for the log-odds of each category with the baseline could also be constructed.

A simple adaptation of an EM-algorithm is proposed for estimation of model parameters. The usual alternated EM algorithm is modified to include penalized ridge estimation of the logistic model parameters in order to avoid the problems produced by the separation that makes the estimators undefined. Other penalized methods are the lasso for logistic regression [Meier et al., 1984] and the Firth method [Firth, 1993] applied to multinomial models by

Apéndice A. Artículo enviado a "Advances in Data Analysis and Classification"

Bull et al. [2002]. The estimators obtained from the package mirt [Chalmers, 2012] can also be used as a start point to construct the biplot, using the factor scores but with an additional step to refit the nominal logistic model for the variable parameters. This is so because mirt is designed for IRT, the scores are always calculated with an additional rotation but the parameters seems not to be rotated consequently. In some examples we have tried the numerical values are strange probably due to the fact that mirt does not take into account the separation problem. Both, our alternated method and mirt perform better when the number of individuals are much higher than the number of variables but there are many practical problems in which this is not so, for example, trying to classify a set of individuals with the genotypes resulting from thousands of single nucleotide polymorphisms [Demey et al., 2008]. For those cases it is probably more efficient to estimate the individual markers by principal coordinates of the matrix **G** of indicators defined previously and then fitting the nominal models on the coordinates. This is not a maximum likelihood solution but it is a good approximation when the other methods are unstable. The main advantage of using maximum likelihood is that it is possible to perform hypothesis testing to compare different models, for example to select the number of dimensions to retain. The proposed method share the characteristics of "formal" models as IRT or latent traits and "descriptive" models as MCA, could even be considered also as a graphical representation of the formal model. It has to be noted that the performance of the algorithm for approximation and inversion of the tessellation crucially depends on the goodness of fit of the nominal regression. Only variables with a reasonable fit should be represented on the graph.

6 Software Note

An R package containing the procedures described by this paper has been developed by the authors [Hernández and Vicente-Villardón, 2013].

Acknowledgements The authors would like to thank the anonymous referees and the editor very much for their careful reading of our manuscript and their valuable comments and suggestions that have improved significantly the paper.

References

Albert A, Anderson J.A (1984) On the existence of maximum likelihood estimates in logistic regression models. Biometrika 71(1):1–10

Baker F(1992) Item Response Theory. Parameter Estimation Techniques. Marcel Dekker.

Bock R, Aitkin M (1981) Marginal maximum likelihood estimation of item parameters: Application of an em algorithm. Psychometrika 46(4):443–459

Apéndice A. ARTÍCULO ENVIADO A "ADVANCES IN DATA ANALYSIS AND CLASSIFICATION"

Browne R.P, McNicholas P.D (2013) Estimating common principal components in high dimensions. Advances in Data Analysis and Classification pp 1–10

Bull S.B, Mak C, Greenwood C.M (2002) A modified score function for multinomial logistic regression. Computational Statistics and data Analysis 39:57–74

Chalmers R.P (2012) Mirt: A multidimensional item response theory package for the r environment. Journal of Statistical Software 48(6):1–29

De Leeuw J (2006) Principal component analysis of binary data by iterated singular value decomposition. Computational Statistics and Data Analysis 50(1):21–39

Demey J, Vicente-Villardón J.L, Galindo M.P, Zambrano A (2008) Identifying molecular markers associated with classification of genotypes using external logistic biplots. Bioinformatics 24(24):2832–2838

Evans D.G, Jones S.M (1987) Detecting voronoi (area of influence) polygons. Mathematical Geology 19(6):523–537

Firth D. (1993) Bias reduction of maximum likelihood estimates. Biometrika 80(1):27–38

Gabriel K.R (1971) The biplot graphic display of matrices with application to principal component analysis. Biometrika 58(3):453–467

Gabriel K.R (1998) Generalised bilinear regresin. Biometrika 85(3):689–700

Gabriel K.R, Zamir S (1979) Lower rank approximation of matrices by least squares with any choice of weights. Technometrics 21(4):489–498

Gallego-Álvarez I, Vicente-Villardón J.L (2012) Analysis of environmental indicators in international companies by applying the logistic biplot. Ecological Indicators 23(0):250–261

Gower J, Hand D (1996) Biplots. Monographs on statistics and applied probability. 54. London: Chapman and Hall., 277 pp.

Hartvigsen D (1992) Recognizing voronoi diagrams with linear programming. ORSA Journal on Computing 4:369–374

Heinze G, Schemper M (2002) A solution to the problem of separation in logistic regresion. Statistics in Medicine 21:2409–2419

Hernández J.C, Vicente-Villardón J.L (2013) Nominal Logistic Biplot: Biplot representations of categorical data. University of Salamanca.Department of Statistics, http://CRAN.R-project.org/package=NominalLogisticBiplot, R package, version 0.1

Jongman R.H.G, Braak C.J.F.Ter, Tongeren O.F.R.V (1987) Data Analysis in Community and Landscape Ecology. Cambridge University Press

Le Cessie S, Van Houwelingen J (1992) Ridge estimators in logistic regression. Applied Statistics 41(1):191–201

Lee S, Huand J, Hu J (2010) Sparse logistic principal component analysis for binary data. Annals of Applied Statistics 4(3):21–39

Meier L, van de Geer S, Buhlmann P (1984) The group lasso for logistic regression. J R Statist Soc 70(1):53–71

Apéndice A. Artículo enviado a "Advances in Data Analysis and Classification"

Schoenberg F, Ferguson T, Li C (2003) Inverting dirichlet tesselations. Computer journal 46(1):76–83

Scrucca L (2013) Graphical tools for model-based mixture discriminant analysis. Advances in Data Analysis and Classification pp 1–19

Vicente-Galindo P, de Noronha Vaz T, Nijkamp P (2011) Institutional capacity to dynamically innovate: An application to the portuguese case. Technological Forecasting and Social Change 78(1):3–12

Vicente-Villardón J.L (2010) MULTBIPLOT: A package for Multivariate Analysis using Biplots. University of Salamanca.Department of Statistics,http://biplot.usal.es/ClassicalBiplot/index.html

Vicente-Villardón J.L, Galindo M.P, Blázquez-Zaballos A (2006) Logistic biplots. Multiple Correspondence Analysis and related methods pp 491–509

Apéndice B

Artículo enviado a la revista "Statistics and Computing"

Statistics and Computing manuscript No.
(will be inserted by the editor)

Logistic Biplots for Ordinal Variables with an Application to Survey Data

José Luis Vicente-Villardón · Julio César Hernández Sánchez

Received: date / Accepted: date

Abstract Biplot methods allow for the simultaneous representation of rows and columns of a data matrix. For binary or nominal data, logistic biplots have been recently developed to extend the classical linear representations for continuous data. Linear, binary or nominal logistic biplots are not adequate for ordinal data and techniques as categorical principal component analysis (CATPCA) or item response theory (IRT) for ordinal items should be used instead.

In this paper we extend the biplot methodology to ordinal data. The resulting method is termed "ordinal logistic biplot" (OLB). Row scores are computed to have ordinal logistic responses along the dimensions and column parameters produce logistic response surfaces that, projected onto the space spanned by the row scores, define a linear biplot. A proportional odds model is used, obtaining a multidimensional model similar to a graded response model in the IRT. We study the geometry of such a representation and construct computational algorithms for estimating model parameters and calculating prediction directions for visualization. Ordinal logistic biplots extend both CATPCA and IRT in the sense that gives a graphical representation for IRT similar to the biplot for CATPCA.

The main theoretical results are applied to the study of job satisfaction of doctorate (PhD) holders in Spain. Holders of doctorate degrees or other research qualifications are crucial to the creation, commercialization and dissemination of knowledge and to innovation. The proposed methods are used to extract useful information from the Spanish data from the international 'Survey on the careers of doctorate holders (CDH)', jointly carried out by Eurostat, the Organisation for Economic Co-operation and Development (OECD) and UNESCO's Institute for Statistics (UIS).

Keywords Biplot · Ordinal Variables · Logistic Responses · Latent Traits

José Luis Vicente-Villardón ✉
Departamento de Estadística, Universidad of Salamanca
C/Alfonso X el Sabio S/N 37007, Salamanca, Spain
Tel.: +0034 923294500 Ext:1852
E-mail: villardon@usal.es

Julio César Hernández Sánchez
Spanish Statistical Office
Zamora(Spain)
E-mail: juliocesar.hernandez.sanchez@ine.es

1 Introduction

The Biplot method (Gabriel (1971), Gower and Hand (1996)) is becoming one of the most popular techniques for analysing multivariate data. Biplot methods are techniques for simultaneous representation of the I rows and J columns of a data matrix $\mathbf{X}$, in reduced dimensions, where rows usually represent to individuals, objects or samples and columns to variables measured on them. Classical Biplot methods are a graphical representation of a Principal Components Analysis (PCA) or Factor Analysis (FA) that are used to obtain linear combinations that successively maximize the total variability. From another point of view, Classical Biplots can be obtained from alternated regressions and calibrations (Gabriel and Zamir, 1979). This approach is essentially an alternated least squares algorithm equivalent to an EM-algorithm when data are normal.

For data with distributions from the exponential family, Gabriel (1998), describes "bilinear regression" as a method to estimate biplot parameters, but the procedure have never been implemented and the geometrical properties of the resulting representations have never been studied in detail. de Leeuw (2006) proposes Principal Components Analysis for Binary data based on an alternate procedure in which each iteration is performed using iterative majorization and Lee et al. (2010) extends the procedure for sparse data matrices, none of those describe the associated biplot. Vicente-

Villardon et al. (2006) propose a biplot based on logistic responses called "Logistic Biplot" that is linear; this paper studies the geometry of this kind of biplots and uses a estimation procedure that is slightly different from Gabriel's method. A heuristic version of the procedure for large data matrices in which scores for individuals are calculated with an external procedure as Principal Coordinates Analysis is described in Demey et al. (2008). Method is called "External Logistic Biplot". Binary Logistic Biplots have been successfully applied to different data sets, see for example, Demey et al. (2008), Vicente-Galindo et al. (2011) or Gallego and Vicente-Villardon (2012). For nominal data, Hernandez Sanchez and Vicente-Villardon (2013) propose a biplot representation based on convex prediction regions for each category of a nominal variable. EM algorithm is used for the parameter estimation. In section 2, biplots for continuous, binary and, in lesser extent, nominal variables are described.

Linear, binary or nominal logistic biplots are not adequate when data are ordinal and techniques as CATPCA or IRT for ordinal items should be used instead. In Section 3 we extend the logistic biplot to ordinal data. The resulting method is termed "ordinal logistic biplot" (OLB). Row scores are computed to have ordinal logistic responses along the dimensions and column parameters produce logistic response surfaces that, projected onto the space spanned by the row scores, define a linear biplot. A proportional odds model is used, obtaining a multidimensional model similar to a graded response model in the IRT. We study the geometry of such a representation and construct computational algorithms for estimating the model parameters and calculating the prediction directions (or axes) used for visualization on the biplot. Ordinal logistic biplots extend both CATPCA and IRT in the sense that gives a graphical representation for IRT similar in some way to the biplot for CATPCA. In Section 4 the main results are applied to the study of job satisfaction of doctorate (PhD) holders in Spain. Holders of doctorate degrees or other research qualifications are crucial to the creation, commercialization and dissemination of knowledge and to innovation. The proposed methods are used to extract useful information from the Spanish data from the international 'Survey on the careers of doctorate holders (CDH)', jointly carried out by Eurostat, the Organisation for Economic Co-operation and Development (OECD) and UNESCO's Institute for Statistics (UIS). Finally, in Section 5, there is a brief discussion concerning both, statistical and applied results.

2 Logistic biplot for continuous, binary or nominal data

In this section we describe the biplot for continuous, binary and nominal data, being the first two treated in a greater extent because of the closer relation to the proposal of this paper.

2.1 Continuous data

Let $\mathbf{X}_{I\times J}$ be a data matrix of continuous measures, and consider the following reduced rank model (S-dimensional)

$$\mathbf{X} = \mathbf{1}_I\mathbf{b}_0' + \mathbf{A}\mathbf{B}' + \mathbf{E} \tag{1}$$

where $\mathbf{b}_0'$ is a vector of constants, usually the column means($\mathbf{b}_0 = \overline{\mathbf{x}}$), $\mathbf{A}$ and $\mathbf{B}$ are matrices of rank S with I and J columns respectively, and $\mathbf{E}$ is a $I \times J$ matrix of errors or residuals. The reduced rank approximation of the centred data matrix (expected values), written as

$$\tilde{\mathbf{X}} = E\left[\mathbf{X} - \mathbf{1}_I\mathbf{b}_0'\right] = \mathbf{A}\mathbf{B}' \tag{2}$$

or

$$E\left[\mathbf{X}\right] = \mathbf{1}_I\mathbf{b}_0' + \mathbf{A}\mathbf{B}', \tag{3}$$

usually obtained from its singular value decomposition (SVD), is closely related to its principal components and its called a biplot (Gabriel, 1971) because it can be used to simultaneously plot the individuals and variables using the rows of $\mathbf{A} = (\mathbf{a}_1,\ldots,\mathbf{a}_I)'$ and $\mathbf{B} = (\mathbf{b}_1,\ldots,\mathbf{b}_J)'$ as markers, in such a way that the inner product $\mathbf{a}_i'\mathbf{b}_j$ approximates the element $\tilde{x}_{ij}$ as close as possible.

If we consider the row markers $\mathbf{A}$ as fixed and the data matrix previously centred, the column markers can be computed by regression trough the origin:

$$\mathbf{B}' = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'(\mathbf{X} - \mathbf{1}_I\overline{\mathbf{x}}). \tag{4}$$

In the same way, fixing $\mathbf{B}$, $\mathbf{A}$ can be obtained as:

$$\mathbf{A}' = (\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'(\mathbf{X} - \mathbf{1}_I\overline{\mathbf{x}})'. \tag{5}$$

Alternating the steps (4) and (5) the product converges to the same subspace generated by the SVD of the centred data matrix. Regression step in equation (4) adjusts a separate linear regression for each column (variable) and interpolation step in equation (5), interpolates an individual using the column markers as reference. The procedure is in some way a kind of EM-algorithm in which the regression step is the maximization part and the interpolation step is the expectation part. An extension for frequency matrices can be found in Gabriel et al. (1998). In summary, the expected values on the original data matrix are obtained on the biplot using a simple scalar product, that is, projecting the point $\mathbf{a}_i$ onto the direction defined by $\mathbf{b}_j$. This is why row markers are usually represented as points and column markers as vectors (also called biplot axis by Gower and Hand (1996)).

The biplot axis can be completed with scales to predict individual values of the data matrix. To find the point on the

biplot direction, that predicts a fixed value μ of the observed variable when an individual point is projected, we look for the point (x,y) lies on the biplot axis, i. e. that verifies

$$y = \frac{b_{j2}}{b_{j1}}x$$

and

$$\mu = b_{j0} + b_{j1}x + b_{j2}y$$

Solving for x and y, we obtain

$$x = (\mu - b_{j0})\frac{b_{j1}}{b_{j1}^2 + b_{j2}^2}$$

and

$$y = (\mu - b_{j0})\frac{b_{j2}}{b_{j1}^2 + b_{j2}^2}$$

or

$$(x,y) = (\mu - b_{j0})\frac{\mathbf{b}_j}{\mathbf{b}_j'\mathbf{b}_j}$$

Therefore, the unit marker for the j-th variable is computed by dividing the coordinates of its corresponding marker by its squared length, several points for specific values of μ can be labelled to obtain a reference scale. If data are centred then $b_{j0} = 0$ and labels can be calculated by adding the average to the value of μ ($\mu + \bar{x}_j$). The resulting representation will be like the one shown in Fig.1.

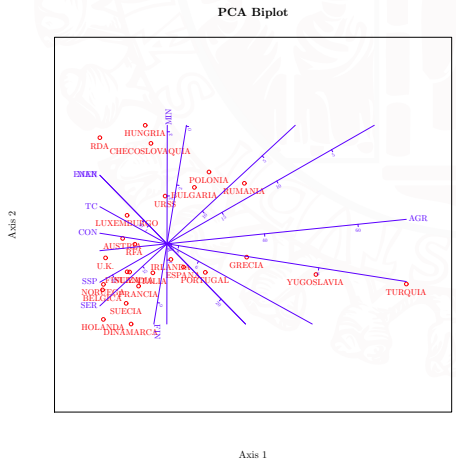


Fig. 1: PCA-Biplot with scales for the variables.

2.2 Binary data

Let $\mathbf{P}_{I\times J}$ be a binary data matrix. Let $\pi_{ij} = E(x_{ij})$ the expected probability that the character j be present at individual i, and p_{ij} the observed probability, either 0 or 1, resulting in a binary data matrix. The S-dimensional logistic biplot in the *logit* scale is formulated as:

$$logit(\pi_{ij}) = \log(\frac{\pi_{ij}}{1-\pi_{ij}}) = b_{j0} + \sum_{s=1}^{S} b_{js}a_{is} = b_{j0} + \mathbf{a}_i'\mathbf{b}_j, \quad (6)$$

where a_{is} and b_{js}, $(i = 1,\ldots,I; j = 1,\ldots,J; s = 1,\ldots,S)$, are the model parameters used as row and column markers respectively. The model is a generalized (bi)linear model having the *logit* as a link function. In terms of probabilities rather than *logits*:

$$\pi_{ij} = \frac{e^{b_{j0}+\sum_k b_{jk}a_{ik}}}{1+e^{b_{j0}+\sum_k b_{jk}a_{ik}}} = \frac{1}{1+e^{-(b_{j0}+\sum_k b_{jk}a_{ik})}}. \quad (7)$$

In matrix form:

$$logit(\boldsymbol{\Pi}) = \mathbf{1}_I\mathbf{b}'_0 + \mathbf{A}\mathbf{B}', \quad (8)$$

where $\boldsymbol{\Pi}$ is the matrix of expected probabilities, $\mathbf{1}_I$ is a vector of ones and $\mathbf{b}_0 = (b_{10},\ldots,b_{J0})$ is the vector containing intercepts that have been added because it is not possible to centre the data matrix in the same way as in linear biplots. The points predicting different probabilities are on parallel straight lines on the biplot; this means that predictions on the logistic biplot are made in the same way as on the linear biplots, i. e., projecting a row marker $\mathbf{a}_i = (a_{i1}, a_{i2})$ onto a column marker $\mathbf{b}_j = (b_{j1}, b_{j2})$. (See Vicente-Villardon et al. (2006) or Demey et al. (2008)). The calculations for obtaining the scale markers are simple. To find the marker for a fixed probability π, we look for the point (x,y) that predicts π and is on the $\mathbf{b}_j$ direction, i. e., on the line joining the points (0, 0) and (b_{j1}, b_{j2}), that is

$$y = \frac{b_{j2}}{b_{j1}}x$$

and

$$logit(\pi) = b_{j0} + b_{j1}x + b_{j2}y$$

We obtain

$$x = \frac{(logit(\pi) - b_{j0})b_{j1}}{b_{j1}^2 + b_{j2}^2}$$

and

$$y = \frac{(logit(\pi) - b_{j0})b_{j2}}{b_{j1}^2 + b_{j2}^2}$$

Several points for specific values of π can be labelled to obtain a reference scale. From a practical point of view the most interesting value is $\pi = 0.5$ because the line passing trough that point and perpendicular to the direction, divides the representation into two regions, one that predicts

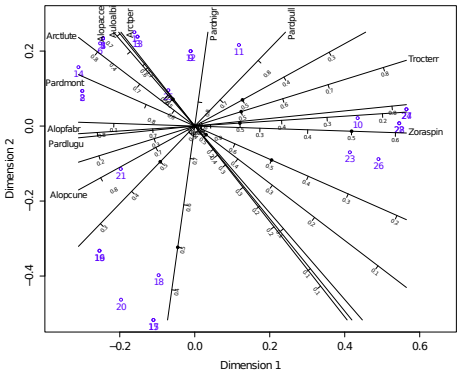


Fig. 2: Binary logistic biplot with probability scales for the variables.

presence and other *absence*. Plotting that point and an arrow pointing to the direction of increasing probabilities should be enough for most practical applications.

The model in (6) is also a latent trait or item response theory model, in that ordination axes are considered as latent variables that explain the association between the observed variables. In this framework we suppose that individuals respond independently to variables, and that the variables are independent for given values of the latent traits. With these assumptions the likelihood function is:

$$\text{Prob}(x_{ij}|\,(\mathbf{b}_0,\mathbf{A},\mathbf{B})) = \prod_{i=1}^{I}\prod_{j=1}^{J}\pi_{ij}^{x_{ij}}(1-\pi_{ij})^{1-x_{ij}}. \tag{9}$$

Taking the logarithm of the likelihood function yields:

$$L = \log \text{Prob}(x_{ij}|\,(\mathbf{b}_0,\mathbf{A},\mathbf{B})) = \\ = \sum_{i=1}^{I}\sum_{j=1}^{J}[x_{ij}\log(\pi_{ij}) + (1-x_{ij})\log(1-\pi_{ij})]. \tag{10}$$

For $\mathbf{A}$ fixed, (10) can be separated into J parts, one for each variable:

$$L = \sum_{J=1}^{J} L_j = \sum_{J=1}^{J}\left(\sum_{i=1}^{I}[x_{ij}log(\pi_{ij}) + (1-x_{ij})log(1-\pi_{ij})]\right). \tag{11}$$

Maximizing each L_j is equivalent to performing a standard logistic regression using the *j-th* column of $\mathbf{X}$ as a response and the columns of $\mathbf{A}$ as regressors. In the same way the probability function can be separated into several parts, one for each row of the data matrix, $L = \sum_{i=1}^{I} L_i$. The details of the procedure to calculate the row or individual scores can be found in Vicente-Villardon et al. (2006).

Binary logistic biplots can be calculated using the package MULTBIPLOT (Vicente-Villardón, 2010). A typical representation of a binary logistic biplot is given in Fig.2.

2.3 Nominal data

Let $\mathbf{X}_{I\times J}$ be a data matrix containing the values of J nominal variables, each with K_j $(j = 1,\ldots,J)$ categories, for I individuals, and let $\mathbf{P}_{I\times L}$ be the corresponding indicator matrix with $L = \sum_j K_j$ columns. The last (or the first) category of each variable will be used as a baseline. Let $\pi_{ij(k)}$ denote the expected probability that the category k of variable j be present at individual i. A multinomial logistic latent trait model with S latent traits, states that the probabilities are obtained as:

$$\pi_{ij(k)} = \frac{e^{b_{j(k)0}+\sum_{s=1}^{S} b_{j(k)s}a_{is}}}{\sum_{l=1}^{K_j} e^{b_{j(l)0}+\sum_{s=1}^{S} b_{j(l)s}a_{is}}}, (k = 1,\ldots,K_j). \tag{12}$$

Using the last category as a baseline in order to make the model identifiable, the parameter for that category are restricted to be 0, i.e., $b_{j(K_j)0} = b_{j(K_j)s} = 0$, $(j = 1,\ldots,J; \quad s = 1,\ldots,S)$.The model can be rewritten as:

$$\pi_{ij(k)} = \frac{e^{b_{j(k)0}+\sum_{s=1}^{S} b_{j(k)s}a_{is}}}{1+\sum_{l=1}^{K_j-1} e^{b_{j(l)0}+\sum_{s=1}^{S} b_{j(l)s}a_{is}}}, (k = 1,\ldots,K_j-1). \tag{13}$$

With this restriction we assume that the log-odds of each response (relative to the last category) follows a linear model:

$$\log\left(\frac{\pi_{ij(k)}}{\pi_{ij(K_j)}}\right) = b_{j(k)0} + \sum_{s=1}^{S} b_{j(k)s}a_{is} = b_{j(k)0} + \mathbf{a}_i'\mathbf{b}_{j(k)},$$

where a_{is} and $b_{j(k)s}$ $(i = 1,\ldots,I; \quad j = 1,\ldots,J; \quad k = 1,\ldots,K_j - 1; \quad s = 1,\ldots,S)$ are the model parameters. In matrix form:

$$\mathbf{O} = \mathbf{1}_I\mathbf{b}_0' + \mathbf{A}\mathbf{B}', \tag{14}$$

where $\mathbf{O}_{I\times(L-J)}$ is the matrix containing the expected log-odds, defines a biplot for the odds. Although the biplot for the odds may be useful, it would be more interpretable in terms of predicted probabilities and categories. This biplot will be called “nominal logistic biplot”, and it is related to the latent nominal models in the same way as classical linear biplots are related to factor or principal components analysis or binary logistic biplots are related to the item reponse theory or latent trait analysis for binary data.

The points predicting different probabilities are no longer on parallel straight lines; this means that predictions on the

logistic biplot are not made in the same way as in the linear biplots, the response surfaces define now prediction regions for each category as shown in Hernandez Sanchez and Vicente-Villardon (2013). The nominal logistic biplot is described here in less detail because its geometry is less related to our proposal than linear or binary logistic biplots.

3 Logistic biplot for ordinal data

3.1 Formulation and geometry

Let $\mathbf{X}_{I\times J}$ be a data matrix containing the measures of I individuals on J ordinal variables with $K_j, (j = 1,\ldots,J)$ ordered categories each, and let $\mathbf{P}_{I\times L}$ the indicator matrix with $L = \sum_j(K_j)$ columns. The indicator $I \times K_j$ matrix for each categorical variable $\mathbf{P}_j$ contains binary indicators for each category and $\mathbf{P} = (\mathbf{P}_1,\ldots,\mathbf{P}_J)$. Each row of $\mathbf{P}_j$ sums 1 and each row of $\mathbf{P}$ sums J. Then $\mathbf{P}$ is the matrix of observed probabilities for each category of each variable.

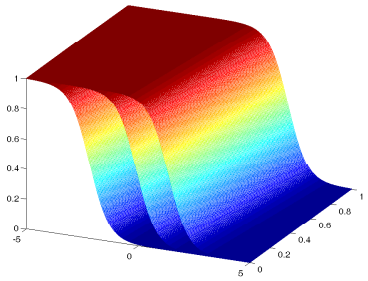


Fig. 3: Cumulative response curves for a two-dimensional latent trait and a variable with four categories.

Let $\pi^*_{ij(k)} = P(x_{ij} \leq k)$ be the (expected) cumulative probability that individual i has a value lower than k on the $j-th$ ordinal variable, and let $\pi_{ij(k)} = P(x_{ij} = k)$ the (expected) probability that individual i takes the $k-th$ value on the $j-th$ ordinal variable. Then $\pi^*_{ij(K_j)} = P(x_{ij} = K_j) = 1$ and $\pi_{ij(k)} = \pi^*_{ij(k)} - \pi^*_{ij(k-1)}$ (with $\pi^*_{ij(0)} = 0$). A multidimensional (S-dimensional) logistic latent trait model for the cumulative probabilities can be written for $(1 \leq k \leq K_j - 1)$ as

$$\pi^*_{ij(k)} = \frac{1}{1+e^{-(d_{jk}+\sum_{s=1}^{S} a_{is}b_{js})}} = \frac{1}{1+e^{-(d_{jk}+\mathbf{a}_i'\mathbf{b}_j)}} \qquad (15)$$

where $\mathbf{a}_i = (a_{i1},\ldots,a_{iS})'$ is the vector of latent trait scores for the $i-th$ individual and d_{jk} and $\mathbf{b}_j = (b_{j1},\ldots,b_{jS})'$ the parameters for each item or variable. Observe that we have defined a set of binary logistic models, one for each category, where there is a different intercept for each but a common set of slopes for all. In the context of IRT, this is known as the "*Graded Response Model*" or *Samejima's model* (Samejima, 1969). The main difference with IRT models is that we don't have the restriction that the probability of obtaining a higher category must increase along the dimensions. Our variables are not necessarily items of a test, but the models are formally the same for both cases. For the unidimensional case that corresponds to a model with an unique discrimination b_j for all categories and different threshold, boundaries, difficulties or location parameters $d_{j(k)}$. The two-dimensional cumulative model is shown in Fig.3.

The $\mathbf{a}_i$ scores can be represented in a scatter diagram and used to establish similarities and differences among individuals or searching for clusters with homogeneous characteristics, i. e., the representation is like the one obtained from any multidimensional scaling method. In the following we will see that the $\mathbf{b}_j$ parameters can also be represented on the graph as directions on the scores space that best predict probabilities and are used to help in searching for the variables or items responsible for the differences among individuals.

In logit scale, the model is

$$logit(\pi^*_{ij(k)}) = d_{j(k)} + \sum_{s=1}^{S} a_{is}b_{js} = d_{j(k)} + \mathbf{a}_i'\mathbf{b}_j, k = 1,\ldots,K_j - 1 \qquad (16)$$

That defines a binary logistic biplot for the cumulative categories.

In matrix form:

$$logit(\boldsymbol{\Pi}^*) = \mathbf{1}\mathbf{d}' + \mathbf{A}\mathbf{B}' \qquad (17)$$

where $\boldsymbol{\Pi}^* = (\Pi^*_1,\ldots,\Pi^*_J)$ is the $I \times (L-J)$ matrix of expected cumulative probabilities, $\mathbf{1}_I$ is a vector of ones and $\mathbf{d} = (\mathbf{d}'_1,\ldots,\mathbf{d}'_J)$, with $\mathbf{d}'_j = (d_{j(1)},\ldots,d_{j(K_j-1)})$, is the vector containing thresholds, $\mathbf{A} = (\mathbf{a}'_1,\ldots,\mathbf{a}'_I)'$ with $\mathbf{a}'_i = (a_{i1},\ldots,a_{iS})$ is the $I \times S$ matrix containing the individual scores matrix and $\mathbf{B} = (\mathbf{B}'_1,\ldots,\mathbf{B}'_J)'$ with $\mathbf{B}_j = \mathbf{1}_{K_j-1} \otimes \mathbf{b}'_j$ and $\mathbf{b}'_j = (b_{j1},\ldots,b_{jS})$, is the $(L-J) \times S$ matrix containing the slopes for all the variables. This expression defines a biplot for the odds that will be called "ordinal logistic biplot". Each equation of the cumulative biplot shares the geometry described for the binary case (Vicente-Villardon et al., 2006), moreover, all curves share the same direction when projected on the biplot. The set of parameters $\{d_{jk}\}$ provide a different threshold for each cumulative category, the second part of (16) does not depend on the particular category, meaning that all the $K_j - 1$ curves share the same slopes. In the following paragraphs we will obtain the geometry for the general case an algorithm to perform the calculations.

The expected probability of individual i responding in category k to item j, with $(k = 1,\ldots,K_j)$, that we denote by $\pi_{ij(k)} = P(x_{ij} = k)$ must be obtained by subtracting cumulative probabilities:

$$\pi_{ij(k)} = \pi^*_{ij(k)} - \pi^*_{ij(k-1)}$$

then using the equations in (15):

$$\begin{aligned}
\pi_{ij(1)} &= P(x_{ij} = 1) = \frac{1}{1+e^{-(d_{j1}+\mathbf{a}_i'\mathbf{b}_j)}} \\
\pi_{ij(k)} &= P(x_{ij} = k) = P(x_{ij} \leq k) - P(x_{ij} \leq (k-1)) \\
&= \frac{1}{1+e^{-(d_{jk}+\mathbf{a}_i'\mathbf{b}_j)}} - \frac{1}{1+e^{-(d_{j(k-1)}+\mathbf{a}_i'\mathbf{b}_j)}} \\
&= \frac{e^{-(\mathbf{a}_i'\mathbf{b}_j)}(e^{-d_{j(k-1)}} - e^{-d_{jk}})}{(1+e^{-(d_{jk}+\mathbf{a}_i'\mathbf{b}_j)})(1+e^{-(d_{j(k-1)}+\mathbf{a}_i'\mathbf{b}_j)})}, \; 1 < k < K_j \\
\pi_{ij(K_j)} &= P(x_{ij} = K_j) = 1 - \frac{1}{1+e^{-(d_{j(K_j-1)}+\mathbf{a}_i'\mathbf{b}_j)}}
\end{aligned} \tag{18}$$

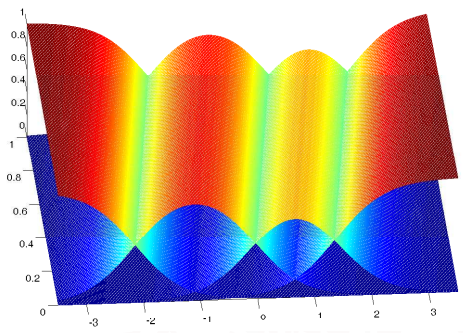


Fig. 4: Response curves for an ordinal variable with four ordered categories.

If the row scores where known, obtaining the parameters of the model in 18 is equivalent to fitting a proportional odds model using each item as a response and the row scores as regressors. The response surfaces for such a model are shown in Fig.4. Although the response surfaces are no longer sigmoidal, the level curves are still straight lines, so the set of points on the representation (generated by the columns of **A**) predicting a particular value for the probability of a category lie on a straight line, and different probabilities for all the categories of a particular variable or item lie on parallel straight lines. A perpendicular to all those lines can be used as ”biplot axis” in the sense of Gower and Hand (1996) and is the direction that better predicts the probabilities of all the categories, that is, projecting any individual point onto that direction, we should obtain an optimal prediction of the category probabilities. As all the categories share the same biplot direction, it would be very difficult to place a different graded scales for each and we will represent just the line segments in which the probability of a category is higher then the probability of the rest. That will result, except for some pathological cases, in as many segments as categories (K_j), separated by $K_j - 1$ points in which the probabilities of two (contiguous) categories are equal. See Fig.5 in which we show the parallel lines representing the points that predict equal probabilities for two contiguous categories and a line, perpendicular to all, that is the ”biplot axis”. The three parallel lines divide the space spanned by the columns of **A** into four regions, each predicting a particular category of the variable. For a biplot representation we don’t need the whole set of lines but just the ”axis” and the points on it, intersecting the boundaries of the prediction regions.

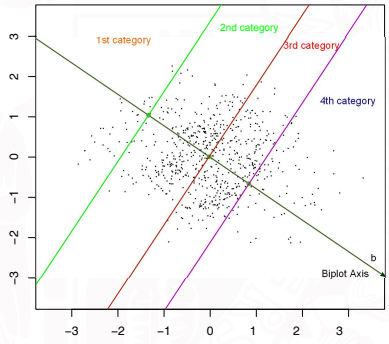


Fig. 5: Prediction regions determined by three parallel straight lines for an ordinal variable with four categories.

3.2 Obtaining the biplot representation

So, if we denote (x,y) one of those intersection points, it must be on the biplot direction, that is,

$$y = \frac{b_{j2}}{b_{j1}}x \tag{19}$$

and the probability of two, possibly contiguous, categories (for example l and m) at this point, must be equal,

$$\pi_{j(l)} = \pi_{j(m)} \quad (\pi^*_{j(l)} - \pi^*_{j(l-1)} = \pi^*_{j(m)} - \pi^*_{j(m-1)}). \tag{20}$$

We have omitted index i because probabilities are for a general point and not for a particular individual. Using the condition in 19 we can rewrite the cumulative probabilities (or its *logit*) as

$$logit(\pi^*_{j(k)}) = d_{j(k)} + xb_{j1} + yb_{j2} = d_{j(k)} + z \tag{21}$$

with

$$z = x\left(\frac{b^2_{j1} + b^2_{j2}}{b_j}\right) \tag{22}$$

Changing the values of z we can obtain the probabilities of each category along the biplot axis. So, finding the point (x,y) is equivalent to find the values of z in which 20 holds. From those values the original point is obtained solving for x in 22 and then calculating y from 19.

There are some pathological cases in which the probability of one or several categories are never higher than the probability of the rest, in such cases we say that the category is "hidden" or "never predicted" and the number of separating points will be lower than $K_j - 1$. Those pathological cases have to be taken into account when calculating the intersection points.

The existence of abnormal cases means that, not just contiguous, but any pair of categories may have to be compared. Then, many comparisons are possible because the equations are different for each case

1. 1-2
2. 1-l $(l < K_j)$
3. 1-K_j
4. l-K_j $(l > 1)$
5. l-$(l+1)$ with $l > 1$
6. l-j with $j > (l+1), l > 1$
7. $(K_j - 1)$-K_j

For example, in case (3) 1-K_j, is simple to deduce that

$$z = \frac{-(d_{j(K_j-1)} + d_{j(1)})}{2}.$$

Cases (1), (3), (5) and (7) are simple. In the other 3 combinations we have to solve a quadratic equation to obtain the intersection points. For example, in case (2), the first with the $l-th$ categories, we have to solve $\pi_{j(1)} = \pi_{j(l)}$, that is

$$\frac{1}{1+e^{-(d_{j(1)}+z)}} = \frac{e^{-z}(e^{-d_{j(l-1)}} - e^{-d_{j(l)}})}{(1+e^{-(d_{j(1)}+z)})(1+e^{-(d_{j(l-1)}+z)})}$$

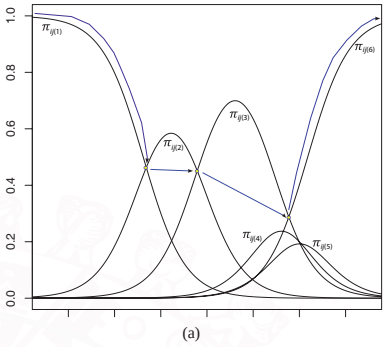
Taking

$$w = e^{-z}$$

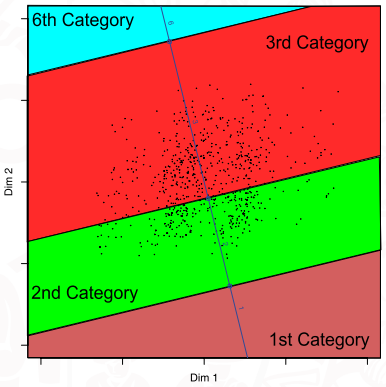
we have to solve the quadratic equation

$$\alpha w^2 - \beta w - 1 = 0$$

with $\alpha = (e^{-(d_{j(1)}+d_{j(l-1)})} - e^{-(d_{j(1)}+d_{j(l)})} - e^{-(d_{j(l-1)}+d_{j(l)})})$ and $\beta = 2e^{-d_i}$.



(a)



(b)

Fig. 6: Probability curves for a variable with 6 categories in which two (4 and 5) are hidden or never predicted. (a) Projection of the response curves onto a plane prependicular to the biplot axis. (b) Final representation without the hidden categories.

If the roots of the equation are both negative, curves don't intersect. If it has a positive root, we can calculate the intersection points solving for w and then reversing the transformations to obtain (x,y). In a similar way we can calculate the intersection points for cases (4) i-K_j $(i > 1)$ and (6) i-j with $j > (i+1)$.

A procedure to calculate the representation of an ordinal variable on the biplot would be as follows:

1. Calculate the biplot axis with equation $y = \frac{b_{j2}}{b_{j1}}x$.
2. Calculate the intersection points z and then (x,y) of the biplot axis with the parallel lines used as boundaries of the prediction regions for each pair of categories, in the

following order:

$$\pi_{j(1)} = \pi_{j(2)}$$
$$\pi_{j(l-1)} = \pi_{j(l)}\ , 1 < l < (K_j - 1)$$
$$\pi_{j(K_j-1)} = \pi_{j(K_j)}$$

3. If the values of z are ordered, there are not hidden categories and the calculations are finished.
4. If the values of z are not ordered we can do the following:
   (a) Calculate the z values for all the pairs of curves, and the probabilities for the two categories involved.
   (b) Compare each category with the following, the next to represent is the one with the highest probability at the intersection.
   (c) If the next category to represent is K_j the process is finished. It not go back to the previous step, starting with the new category.

A simpler algorithm based on a numeric procedure could also be developed to avoid the explicit solution of the equations.

1. Calculate the predicted category for a set of values for z. For example a sequence from -6 to 6 with steps of 0.001. (The precision of the procedure can be changed with the step)
2. Search for the z values in which the prediction changes from one category to another.
3. Calculate the mean of the two z values obtained in the previous step and then the (x,y) values. Those are the points we are searching for.

Hidden categories are the ones with zero frequencies in the predictions obtained by the algorithm.

3.3 Paramater estimation

The alternated algorithm described in Vicente-Villardon et al. (2006), can be easily extended replacing binary logistic regressions by ordinal logistic regressions. The problem with this approach is that the parameters for the individuals can not be estimated when the individual has 0 or 1 in all the variables for the binary case, or all the responses are at the baseline category for the ordinal case. In this paper we use a procedure that is similar to the alternated regressions method, except that the interpolation step is "changed" by *a posteriori* expected values. The estimation procedure is an EM-algorithm that uses the Gauss-Hermite quadrature to approximate the integrals, considering the individual scores as missing data. More details of similar procedures can be found in Bock and Aitkin (1981) or Chalmers (2012).

The likelihood function is:

$$M(\mathbf{P}|\mathbf{d},\mathbf{A},\mathbf{B}) = \prod_{i=1}^{I}\prod_{j=1}^{J}\prod_{k=1}^{K_j} \pi_{ij(k)}^{p_{ij(k)}},$$

where $p_{ij(k)} = 1$ if individual i chooses category k of item j and $p_{ij(k)} = 0$ otherwise. The log-likelihood is:

$$L(\mathbf{P}|\mathbf{d},\mathbf{A},\mathbf{B}) = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K_j} p_{ij(k)} \log\left(\pi_{ij(k)}\right). \tag{23}$$

If the parameters $\mathbf{A}$ for individuals where known, the log-likelihood could be separated into J parts, one for each variable:

$$L(\mathbf{P}|\mathbf{d},\mathbf{B}) = \sum_{j=1}^{J} L_j(\mathbf{P}|\mathbf{d}_j,\mathbf{b}_j) = \sum_{j=1}^{J}\left(\sum_{i=1}^{I}\sum_{k=1}^{K_j} p_{ij(k)} \log\left(\pi_{ij(k)}\right)\right), \tag{24}$$

where $\mathbf{d}_j$ and $\mathbf{b}_j$ are the submatrices of parameters for the jth variable. Maximizing the log-likelihood is equivalent to maximizing each part, i.e., obtaining the parameters for each variable separately. Maximizing each L_j is equivalent to performing an ordinal logistic regression using the jth column of $\mathbf{X}$ as response and the columns of $\mathbf{A}$ as predictors. We do not describe logistic regression here because it is as a very well known procedure. It is also well-known that when the individuals for different categories are separated (or quasi-separated) on the space spanned by the explanatory variables, the maximum likelihood estimators don't exist (or are unstable). Because we are seen the biplot as a procedure to classify the set of individuals and searching for the variables responsible for it, accounting for as much of the information as possible, it is probable that, for some variables, the individuals are separated and then the procedure does not work just because the solution is good. The problem of the existence of the estimators in logistic regression can be seen in Albert and Anderson (1984), a solution for the binary case, based on the Firth's method (Firth, 1993) is proposed by Heinze and Schemper (2002). All the procedures were initially developed to remove the bias but work well to avoid the problem of separation. Here we have chosen a simpler solution based on ridge estimators for logistic regression (Le Cessie and Van Houwelingen, 1992).

Rather than maximizing $L_j(\mathbf{P}|\mathbf{d}_j,\mathbf{b}_j)$ we maximize:

$$L_j(\mathbf{P}|\mathbf{d}_j,\mathbf{b}_j) - \lambda\left(\|\mathbf{d}_j\|^2 + \|\mathbf{b}_j\|^2\right). \tag{25}$$

We don't describe here the procedure in great detail because that is also a standard procedure. Changing the values of λ we obtain slightly different solutions not affected by the separation problem.

In the same way, if parameters for variables were known, the log-likelihood could be separated into I parts, one for each individual:

$$L(\mathbf{P}|\mathbf{A}) = \sum_{i=1}^{I} L_i(\mathbf{P}|\mathbf{a}_i) = \sum_{i=1}^{I}\left(\sum_{j=1}^{J}\sum_{k=1}^{K_j} p_{ij(k)} \log\left(\pi_{ij(k)}\right)\right).$$

To maximize each part we could use Newton-Raphson with a penalization as before. Rather than that we will use expected a posteriori estimators for the individual markers. For each individual (or response pattern) $\mathbf{p}_i$, the likelihood is:

$$M(\mathbf{P}=\mathbf{p}_i|\,\mathbf{d},\mathbf{A}=\mathbf{a}_i,\mathbf{B})=\prod_{j=1}^{J}\prod_{k=1}^{K_j}\pi_{ij(k)}^{p_{ij(k)}}.$$

Assuming a distributional form $g(\mathbf{a})$ (multivariate normal, for example) the marginal distribution becomes:

$$P(\mathbf{P}=\mathbf{p}_i|\mathbf{d},\mathbf{B})=\int M(\mathbf{P}=\mathbf{p}_i|\,\mathbf{d},\mathbf{A}=\mathbf{a},\mathbf{B})g(\mathbf{a})d\mathbf{a},$$

and the observed likelihood:

$$M(\mathbf{P}|\mathbf{d},\mathbf{B})=\prod_{i=1}^{I}\left[\int M(\mathbf{P}=\mathbf{p}_i|\mathbf{d},\mathbf{A}=\mathbf{a},\mathbf{B})g(\mathbf{a})d\mathbf{a}\right].$$

We approximate the integral by S-dimensional Gauss-Hermite quadrature:

$$\tilde{P}(\mathbf{P}=\mathbf{p}_i|\mathbf{d},\mathbf{B})= =\sum_{qS=1}^{Q}\cdots\sum_{q1=1}^{Q}M(\mathbf{P}=\mathbf{p}_i|\,\mathbf{d},\mathbf{Y}=y,\mathbf{B})g(y_{q1})\ldots g(y_{qS}). \tag{26}$$

The multivariate S-dimensional quadrature, $\mathbf{Y}$, has been obtained as the product of S unidimensional quadratures $(y_1,\ldots,y_Q)$ with Q nodes each, $\{g(y_q):q=1,\cdots,Q\}$ are associated weights in the quadrature and $y_{q1\cdots qS}\overset{not}{=}y$ represents each S-dimensional quadrature points . Then the marginal expected a posteriori score for each individual $E(\mathbf{a}|\mathbf{p}_i)$ can be approximated by:

$$\frac{\sum_{qS=1}^{Q}\cdots\sum_{q1=1}^{Q}y_{q1\cdots qS}P(\mathbf{P}=\mathbf{p}_i|\mathbf{d},\mathbf{Y}=y_{q1\cdots qS},\mathbf{B})g(y_{q1})\ldots g(y_{qS})}{\tilde{P}(\mathbf{P}=\mathbf{p}_i|\mathbf{d},\mathbf{B})}, \tag{27}$$

being $y_{q1\cdots qS}$ the S-dimensional points of the multivariate quadrature, as it was denoted before, and $\tilde{P}(\mathbf{P}=\mathbf{p}_i|\mathbf{d},\mathbf{B})$ given by (26).

The ability for individual i has S components (as much as dimensions of spanned space, i.e. $(\mathbf{a}_i=(a_{i1},\cdots,a_{iS}))$, and each one $\{a_{is},s=1,\cdots,S\}$ will be approximated by the expression (27), which depends on each S-dimensional coordinate of $y_{q1\cdots qS}$.

3.4 Goodness of fit

The log-likelihood can be used as a measure of the overall goodness of fit, specially for comparison of different models containing, for example, a different number of dimensions. Statistical tests similar to those proposed in the context of IRT models could be used here, in particular, tests based on ordinal logistic regressions. Mair et al. (2008) propose tests based on logistic regressions for binary data that could easily be extended to ordinal data. On one hand, the statistical tests for this situations have many problems and on the other hand we see the procedure more as a descriptive exploratory model, so we are less interested in global statistical tests and more in goodness of fit indices or tests for each separate variable. For IRT models, the items are closely related to one or several latent dimensions and all are useful to describe the latent factors, so there must be an adequate overall goodness of fit. In a more general exploratory situation some of the variables may not be useful for describing the problem, inducing a lower overall fit that can lead to miss-interpretations. Demey et al. (2008), conduct a simulation study in which some irrelevant variables and noise are added to a known structure, showing that the known structure is recovered even when the overall goodness of fit is not high. Using fit indices for each variable, they are able to identify the relevant and eliminate the irrelevant variables. A similar result can be found in Gabriel (2002).

Several tests and goodness of fit indices can be defined for each variable considering that has been fitted using an ordinal logistic regression with proportional odds. Here we use the likelihood ratio test to compare the model with constant probability (no latent dimensions) with the complete model in exactly the same way as in the standard logistic regression. The test should be interpreted here with care because the latent variables are also estimated inside the procedure and its variation is not considered explicitly in the test; nevertheless is an indication of the significance of the variable to describe the data. For classical linear biplots no such tests are usually performed but goodness of fit indices are calculated. Gardner-Lubbe et al. (2008) define what they call *predictivities* as the percentage of the variance of each variable explained by the dimensions, i. e., is a measure of the prediction accuracy on the biplot. The predictivity is used as a measure of goodness of fit measure by the package BiplotGUI (la Grange et al. (2009)) In the context of Correspondence Analysis those quantities were called relative contributions of the axis to the elements (variables or individuals) (see Benzecri (1976), Greenacre (1984) or Benzecri (1976)) extended to biplots by Galindo-Villardon (1986) or Greenacre (1984). The package MULTBIPLOT (Vicente-Villardón (2010)) uses contributions calculated in that way.

From another point of view, the *predictivity*, is the coefficient of determination R_j^2 for the regressions in 4. For ordi-

nal responses, a pseudo R_j^2 as the Cox-Senell or Nagelkerke can be used.

4 An empirical study

In 2008, for a set of 26 countries worldwide, among which was Spain, setting 2006 as reference year and following the guidelines set by the Organization for Cooperation and Development(OECD), the department of statistics of UNESCO and Eurostat (Statistical Office of the European Union), began to do surveys to people that had obtained a PhD degree, and that therefore are doctorates, with the objective of having a clearer information about their characteristics. Most of the pioneer countries belonged the European Union, although members of the OECD, as USA or Australia, also participated. In the Spanish case, it was the National Institute of Statistics (Instituto Nacional de Estadistica - INE) which focused all efforts to carry out this new operation with the objective that the availability of information in this field had continuity in time. Thus, the so-called “Survey on Human Resources in Science and Technology” was established as part of the general plan of science and technology statistics carried out by the European Union Statistics Office (Eurostat). This need for information is evident in the European Regulation 753/2004 on Science and Technology, which specifies the production of statistics on human resources in science and technology.

Surveys on doctorates (CDH: Careers of doctorate holders) try to measure specific demographic aspects related to employment, so that the level of investigation of this group, the professional activity carried out, the satisfaction with their main job, the international mobility and the income of this group can be quantified in Spain.

The study focused on all the doctorate holders resident in Spain, younger than 70, that obtained their degree in a Spanish university, both public or private, between 1990 and 2006. The frame of this statistical operation was a directory of doctorate holders provided to the National Statistic Institute by the University Council, which includes all the persons who have defended a doctoral thesis in any Spanish university, according with their electronic databases, which was comprised of approximately 80000 people. Doctors belong to the level 6 of the international classification of education ISCED-97. This level is reserved for tertiary programmes which lead to the award of an advanced research qualification and devoted to advanced study and original research and not based on course-work only.

As for the sampling design, a representative sample was designed for each region at NUTS-2 level[1], using a sampling with equal probabilities. The doctors were grouped according to their place of residence, and the selection has been done independently in each region by equal probability with random start systematic sampling. A sample of 17000 doctors was selected. The sample has been distributed between the regions assigning the 50% in an uniform way and the rest proportional to the size of them, measured in number of doctorate holders that have their residence in those regions.

The INE used a questionnaire harmonized at European level, structured in several modules, that can be found at the website of the Institute (http: www.ine.es/metodologia/t14 /t1430225-09-cues.pdf). As a result of the collection process it was obtained a response rate at the national level of 72%.

We have 12193 doctorate’s answers to the questionnaire in Spain to develop this study. We will focus our attention on the module C (Employment situation) and specifically, in the subsection C.6.4, that tries to find out the level of satisfaction of the doctorate holders in aspects related to their principal job. This question has several points of interest coded on a likert scale from 1 to 4 (see Fig.7).

Please rate your satisfaction with your PRINCIPAL JOB's...

Mark (X) ONLY one for each item	Very satisfied	Somewhat satisfied	Somewhat dissatisfied	Very dissatisfied
A. Salary	1	2	3	4
B. Benefits	1	2	3	4
C. Job security	1	2	3	4
D. Job location	1	2	3	4
E. Working conditions	1	2	3	4
F. Opportunities for advancement*	1	2	3	4
G. Intellectual challenge	1	2	3	4
H. Level of responsibility*	1	2	3	4
I. Degree of independence*	1	2	3	4
J. Contribution to society	1	2	3	4
K. Social status	1	2	3	4

Fig. 7: Question C.6.4 of the questionnaire.

Each item will be considered as ordinal, then we have 11 items or variables in total.

Using the alternated algorithm to estimate the parameters of the two-dimensional model, we have obtained the indicators in table 1 and the factor loadings and communalities in table 2. The percentages of correct classifications are very high for the variables salary and intellectual challenge, presenting Nagelkerke’s *pseudo* $-R^2$ values close to one. This makes us thinking that it could be a quasi-separation between categories of the variables, a problem in logistic regression that has been conveniently considered and solved by the estimation method.

Analysing the interpretation of factors using their loadings, the first has higher weights for the variables, opportunities for advancement, degree of independence, intellectual

[1] The NUTS classification (Nomenclature of Territorial Units for Statistics) is a hierarchical system for dividing up the economic territory of the EU for different purposes. (see *http* : *//epp.eurostat.ec.europa.eu/portal/page/nuts_nomenclature /introduction*)

challenge, level of responsibility and contribution to society, which are features associated to research activity. The second factor presents higher values for salary, benefits, job security and working conditions, all related with economical and working conditions. Then we have two main almost independent factors, the first related to the research activity and the second to the conditions of the job. The job location and social status variables have similar loadings in both factors.

Table 1: Fitting indicators for the eleven variables.

Variable	logLik	df	p-value	PCC	Nagelk.
Salary	-2829.4	2	0	0.94	0.96
Benefits	-8158.9	2	0	0.80	0.83
Job Security	-12689.8	2	0	0.58	0.15
Job Location	-11037.4	2	0	0.60	0.10
Working Conditions	-10072.8	2	0	0.63	0.46
Opp.for Advancement	-11963.3	2	0	0.56	0.54
Intelectual Challenge	-1845.1	2	0	0.96	0.97
Level of responsability	-9857.2	2	0	0.59	0.24
Degree of independence	-10049.2	2	0	0.61	0.39
Contribution to society	-9407.9	2	0	0.62	0.25
Social Status	-8991.1	2	0	0.71	0.47

Table 2: Factor loadings and communalities.

Variable	F1	F2	Communalities
Salary	0.105	0.991	0.994
Benefits	0.109	0.986	0.984
Job Security	0.287	0.858	0.819
Job Location	0.684	0.442	0.664
Working Conditions	0.613	0.749	0.938
Opportunities for Adv.	0.876	0.403	0.930
Intelectual Challenge	0.988	-0.137	0.995
Level of responsability	0.902	0.173	0.843
Degree of independence	0.911	0.275	0.906
Contribution to society	0.922	0.018	0.851
Social Status	0.732	0.626	0.929

Item response functions of three of the items can be observed in Fig.8.

It should be pointed out that in the variable relative to job security, the second category is hidden, which is partially satisfied, concentrating all of the information in the other three categories in such a way, that in this aspect, it appears that either the satisfaction is maximum or is low, which it is in line with the organization of the spanish public administration, that concentrates the job of the majority of doctorates.

The biplot of ordinal data can be seen in Fig.9, in which cut-off points, for each variable, from each of the curves and their projections in the reduced space corresponds to the points scored in each of the lines. In this representation, as mentioned before, the angle between the principal axe with some variables, such as Salary, Benefits, and Job security, is

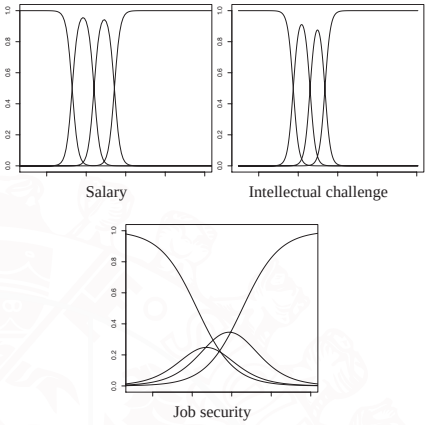


Fig. 8: Item information curves for the items of each of the variables.

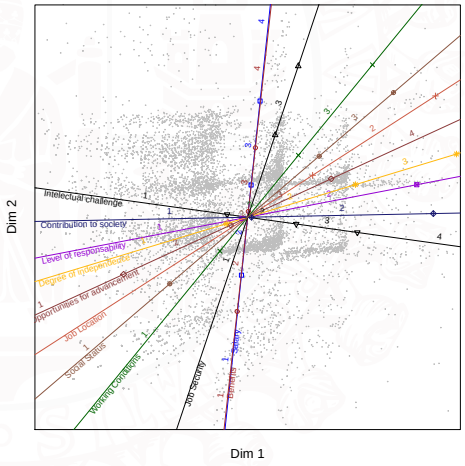


Fig. 9: Ordinal logistic biplot. Satisfaction of the doctorate holders with their principal job in Spain.

near to 90°, presenting a region in the first quadrant, away from the origin, in which doctorates are very dissatisfied in these aspects and others related with the research activity. The representation shows that satisfaction with income not appear to correlate with the doctorates levied on intellectual aspects such as intellectual challenge ot the degree of re-

sponsability, something that also appears in other european countries, as is evident in the Austrian case Schwabe (2011).

Some of the variables have a similar behavior, such as level of responsibility or contribution to society, in which the points that define the position of each category are distributed in a similar way and their biplot axes present a slope very similar. Although there are groups of variables whose directions are very similar, the position of the categories are quite different from each other within those groups. This happens with opportunities for advancement and degree of independence.

If we color the individuals according to the answer to the variables represented in the previous curves of information (see Fig.10), the situation of quasi-separation can be appreciated in the intellectual challenge variable and not in the job security with an individual behavior more disperse. The salary also presents this problem of separation, with a graph with horizontal stripes corresponding to each category. Those variables (salary and intellectual challenge) seem to be important in the interpretation of the information and the understanding of the aspect of the individuals cloud.

Fig.11 shows the convex hulls and centers of sets of points whose category of response is one of the possible for each variable. Considering an ordinary least squares regression using the average of the 11 responses to the question of satisfaction as dependent variable and a wide battery of variables of the questionnaire as independent ones, Canal Domnguez (2013) has showed that there were constraints that seem to be closely linked to job satisfaction, as the sector in which the doctorate works, appearing that private sector aglutinates doctorates more satisfied that public one. Another is the age that seems to have a positive but not significant effect and shows that women are more satisfied overall, or the greater the relationship between training and employment is, the greater worker satisfaction. However, jobs occupated by more qualified workers that required have a negative effect on staff satisfaction, which is consistent with the traditional literature. In addition there seems to be a fundamental aspect whose effect is decisive given the coefficients presented, which is income doctorate. If these increase, higher satisfaction levels are obtained, affecting positively to this variable. The wages range between 20 and 30 thousand euros brings to 24.6% of the sample and more than half of PhDs earn less than 35000 euros. The range of earnings of over 50000 euros gathers percentages similar to those that are in the stretch from 10 to 20 thousand euros (12% doctorates).

There have been studied in the graphical configuration some variables already known, for example, the sector of employment or the source of funding and sex of doctorate, in order to locate the positions of the predicted categories and

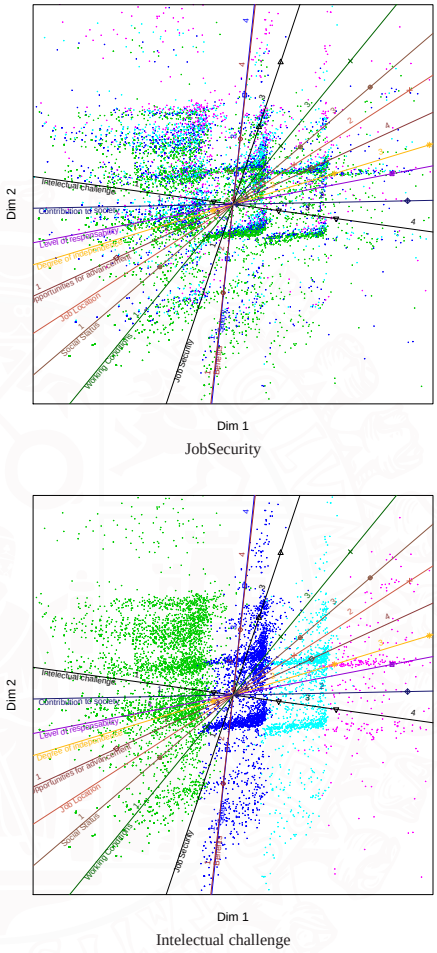
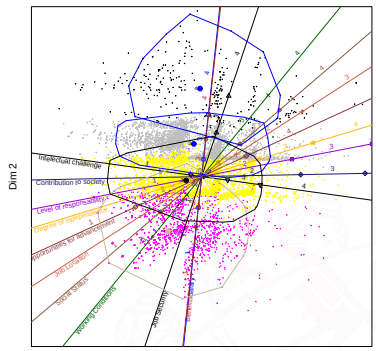


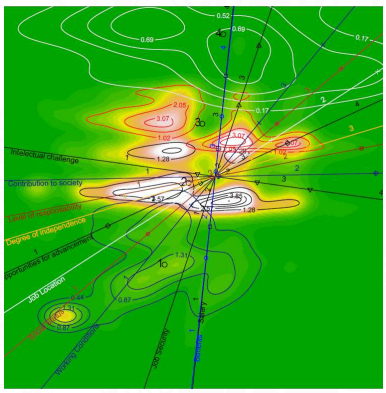
Fig. 10: Coloration according to the category answered by the doctorates in the ordinal logistic biplot.

to investigate the posible relationship with the satisfaction in all of their components.

According to the employment sector, a clear distinction between the higher education sector and other sectors appears. For the first one, it exists a marked intellectual challenge, with a degree of independence and obvious opportunities for job improvement. In terms of salary and benefits, the enterprise sector(BES) shows a more attractive compo-



(a)



(b)

Fig. 11: Convex hulls and density graph with contour lines for the variable Salary

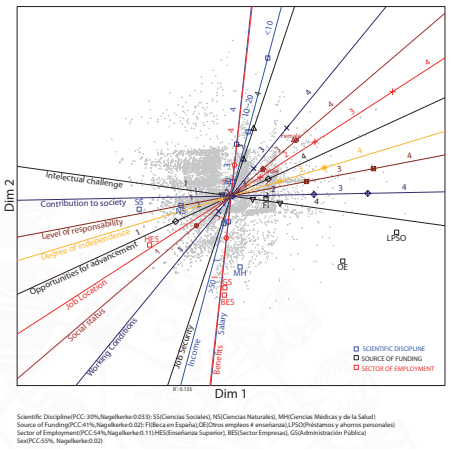


Fig. 12: OLB with ordinal variable income(gross annual salary) fitted on the coordinates of the biplot, and nominal variables sex, scientific discipline, source of funding overlapping.

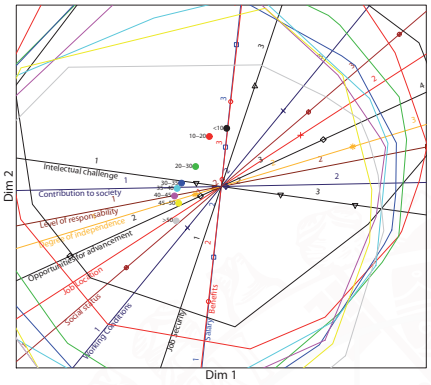
nent for doctorates(Fig 12). It further appears that the fact that these wage expectations penalize the university career makes the professional fate of doctorates is increasingly geared towards private companies and to other areas of public administration, as shown by authors such as Canal and Muiz (2012).

The position of men and women, calculating the centroid of the coordinates of the biplot, is virtually indistinguishable(Fig 13(b)), given the large sample on which we work and there is almost parity, but if we treat the variable as nominal and we adjust it with the dimensions of the biplot we obtain different positions on both sexes, as shown on the same graph, which can be related to variables such as income and wages, confirming the known wage gap between them and consistent with different studies on this aspect.

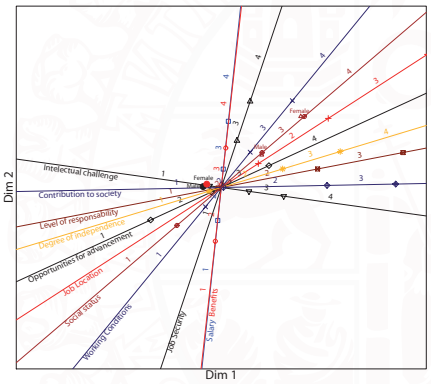
If it is calculated the centroids of different wage brackets for doctorates according to their coordinates in the two-dimensional biplot, figure 13(a) is obtained. However, figure 12 shows the variable relative to the income, treated as ordinal and adjusted on the biplot axes, and with the same categories as presented in the questionnaire [2]. It can be appreciated its relation to the axis related with the workplace itself, and showing that not all categories are predicted, but only some of them. Instead of being considered the variable income as ordinal we could have treated it as nominal with one of the public functions available in the R package "NominalLogisticBiplot" Hernandez Sanchez and Vicente-Villardon (2013), resulting in a really similar to the known configuration, and with the same predicted categories(see fig 14).

Now we could ask how different wages influences in the development of the elements satisfaction, for which let's consider Fig 15; it shows ordinal logistic biplot calculated for the doctorates with the lowest income range (below 10,000 euros). We positioned some nominal variables, such as age, employment sector and scientific discipline associated with this group. It is significant, as expected, that in this context the youngest doctoral are placed, since only the two lower

[2] Interval 1= ¡10000 euros; interval 2: 10000-20000 euros; interval 3: 20000-30000 euros; interval 4: 30000-35000 euros; interval 5: 35000-40000 euros; interval 6: 40000-45000 euros; interval 7: 45000-50000 euros; interval 8: ¿ 50000 euros



(a) Zoom de los centros de la imagen anterior



(b) Centros y ajuste de la variable nominal Sexo sobre el biplot logstico ordinal

Fig. 13: Biplots Logsticos Ordinales con variables con informacin externa situadas en los grficos.

sections are predicted (section 1: ¡34 years, section 2; 35 to 45; section 3; between 45 and 55; section 4; 55 to 65; section 5; between 65 and 70 years), which are mostly in the higher education sector (HES) or in the business sector(BES), and whose scientific guidance is given for the natural sciences (NS), humanities (H) and Social Sciences (SS). There are determinants of satisfaction that seem to appear in this tight wage band and human section, as shown in the first axe, with all matters relating to wages, job security, working conditions and opportunities to promote, bringing together the business sector improved expectations and satisfaction. Moreover it appears issues such as intellectual challenge, level of responsibility and contribution to the society in which, unlike in the previous case, the higher education sector is able to provide higher degrees of satisfaction among doctorates.

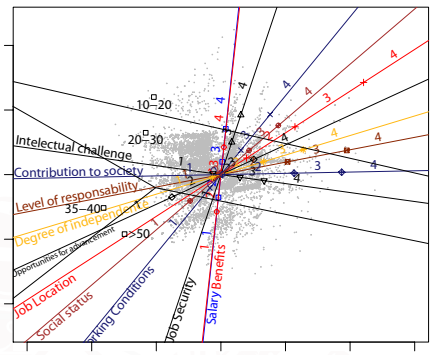


Fig. 14: OLB con la variable Ingresos(Salario Bruto Anual) considerada como variable nominal y ajustada sobre los resultados del biplot.

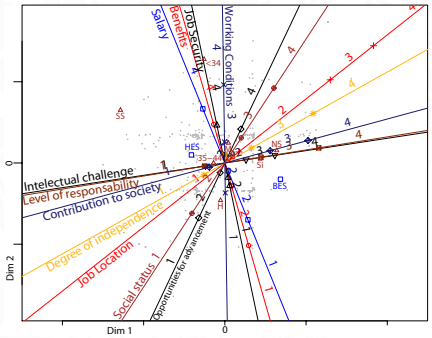


Fig. 15: OLB de los doctorados con menores ingresos, con las variables nominales sector, edad y disciplina cientficas ajustadas sobre el biplot.

You could go exploring how vary the perceptions of the group with increasing income levels, since these can change, as shown in Fig 16, in which the logistic biplot is built with the doctorates with higher income(earnings greater than 50000 euros). It would be interesting to analyze a generation of PhDs to see how they vary their insights and perspectives, but with this data it is not possible because the information

for this group is cross, so the sample composition is very diverse in the sections wage and we have to perform comparisons with some caution. In this chart we have been adjusted on the biplot nominal variables such as sex, financing or scientific discipline, being that you can only predict characteristics for men and not women, as well as the main mechanisms for financing the PhD studies are either a grant (of the institution where they completed their doctorate, government, business or nonprofit institution; FI) or an occupation full or part time (OE). Moreover, for this group they seem to be predicted scientific disciplines related to the social sciences (SS) and Medical Sciences (MH), although a more detailed analysis is necessary, taking into consideration some additional dimension which would be capable of capturing the collective complex variability.

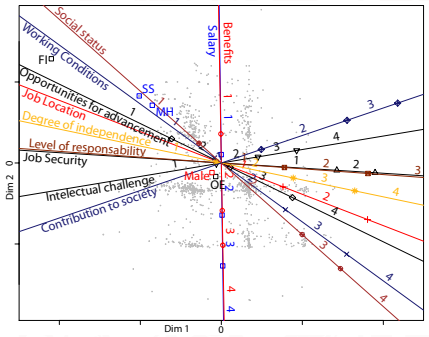


Fig. 16: OLB de los doctorados con mayores ingresos, con las variables nominales sexo, financiacin y sector ajustadas sobre el biplot.

5 Software Note

An R package containing the procedures described by this paper has been developed by the authors (Hernández and Vicente-Villardón, 2013).

References

Albert, A., Anderson, J. (1984) On the existence of maximum likelihood estimates in logistic regression models. Biometrika 71(1):1–10

Benzecri, J. P. (1976) L'analyse des donnees. Dunod

Bock, R., Aitkin, M. (1981) Marginal maximum likelihood estimation of item parameters: Application of an em algorithm. Psychometrika 46(4):443–459

Canal, J. F., Muiz, M. A. (2012) Professional doctorates and the careers: Present and future. the spanish case. European journal of education 47(1):153–171

Canal Domnguez, J. (2013) Earnings and job satisfaction of employed spanish doctoral graduates. Revista Espaola de Investigaciones Sociolgicas 144:49–72

Chalmers, R. P. (2012) Mirt: A multidimensional item response theory package for the r environment. Journal of Statistical Software 48(6):1–29

Demey, J., Vicente-Villardón, J. L., Galindo, M. P., Zambrano, A. (2008) Identifying molecular markers associated with classification of genotypes using external logistic biplots. Bioinformatics 24(24):2832–2838

Firth, D. (1993) Bias reduction of maximum likelihood estimates. Biometrika 80(1):27–38

Gabriel, K. (1971) The biplot graphic display of matrices with application to principal component analysis. Biometrika 58(3):453–467

Gabriel, K. R. (1998) Generalised bilinear regresin. Biometrika 85(3):689–700

Gabriel, K. R. (2002) Goodness of fit of biplots and correspondence analysis. Biometrika 89(2):423–436

Gabriel, K. R., Zamir, S. (1979) Lower rank approximation of matrices by least squares with any choice of weights. Technometrics 21(4):489–498

Gabriel, K. R., Galindo, M. P., Vicente-Villardon, J. L. (1998) Use of Biplots to diagnose independence models in contingency tables., Academic Press, pp. 391–404

Galindo-Villardon, M. P. (1986) Una alternativa de representacion simultanea: Hj-biplot. Questiio 10(1):13–23

Gallego, I., Vicente-Villardon, J. L. (2012) Analysis of environmental indicators in international companies by applying the logistic biplot. Ecological Indicators 23(0):250 – 261, DOI http://dx.doi.org/10.1016/j.ecolind.2012.03.024

Gardner-Lubbe, S., Le Roux, N. J., Gower, J. C. (2008) Measures of fit in principal component and canonical variate analyses. Journal of Applied Statistics 35(9):947–965

Gower, J., Hand, D. (1996) Biplots. Monographs on statistics and applied probability. 54. London: Chapman and Hall., 277 pp.

la Grange, A., le Roux, N., Gardner-Lubbe, S. (2009) Biplotgui: Interactive biplots in r. Journal of Statistical Software 30(12)

Greenacre, M. J. (1984) Theory and Applications of Correspondence Analysis. Academic Press

Heinze, G., Schemper, M. (2002) A solution to the problem of separation in logistic regresion. Statistics in Medicine 21(1):2409?2419

Hernández, J. C., Vicente-Villardón, J. L. (2013) Ordinal Logistic Biplot: Biplot representations of ordinal variables. Universidad de Salamanca.Department of Statistics, r package version 0.3

Hernandez Sanchez, J. C., Vicente-Villardon, J. L. (2013) Logistic biplot for nominal data. ArXiv e-prints 1309.5486

Le Cessie, S., Van Houwelingen, J. (1992) Ridge estimators in logistic regression. Applied Statistics 41(1):191–201

Lee, S., Huand, J., Hu, J. (2010) Sparse logistic principal component analysis for binary data. Annals of Applied Statistics 4(3):21–39

de Leeuw, J. (2006) Principal component analysis of binary data by iterated singular value decomposition. Computational Statistics and Data Analysis 50(1):21–39

Mair, P., Reise, S. P., Bentler, P. (2008) Irt goodness-of-fit using approaches from logistic regression. UCLA Statistics Preprint Series 540

Samejima, F. (1969) Estimation of latent ability using a response pattern of graded scores. Psychometric Monograph Supplement 4(34)

Schwabe, M. (2011) The careers paths of doctoral graduates in austria. European Journal of Education 46(1):153–168

Vicente-Galindo, P., de Noronha Vaz, T., Nijkamp, P. (2011) Institutional capacity to dynamically innovate: An application to the portuguese case. Technological Forecasting and Social Change 78(1):3 – 12, DOI http://dx.doi.org/10.1016/j.techfore.2010.08.004

Vicente-Villardon, J., Galindo, M., Blazquez-Zaballos, A. (2006) Logistic Biplots., Chapman and Hall., pp. 503–521

Vicente-Villardón, J. L. (2010) MULTBIPLOT: A package for Multivariate Analysis using Biplots. University of Salamanca.Department of Statistics, URL http://biplot.usal.es/ClassicalBiplot/index.html

Apéndice C

Poster presentado en el 27º congreso: International Biometric Conference(IBS), en Florencia, 2014.

Logistic Biplots for Nominal and Ordinal Data

José L. Vicente-Villardón[1] & Julio César Henández-Sánchez [2]
[1]Department of Statistics, Universidad Salamanca, Spain. [2] Instituto Nacional de Estadística, (INE)., Spain
http://biplot.usal.es/logisticbiplot villardon@usal.es

The Biplot method is becoming one of the most popular techniques for analysing multivariate data. Classical Biplot Methods allow for the simultaneous representation of individuals (rows) and variables (columns) of a data matrix. For Binary data, Logistic Biplots have been recently developed. When data are nominal or ordinal, linear or even binary logistic biplots are not adequate. In this paper we extend the binary logistic biplot to nominal and ordinal data. The resulting methods are termed Nominal and Ordinal Logistic Biplot respectively (NLB and OLB). In a NLB the variables are represented as convex prediction regions rather than vectors while in an OLB as directions divided into prediction segments are used. Using the methods from Computational Geometry, the set of prediction regions, in the NLB, is converted to a set of points in such a way that the prediction for each individual is established by its closest "category point". Then interpretation is based on distances rather than on projections. For the OLB, row scores are computed to have ordinal logistic responses along the dimensions and column parameters produce logistic response surfaces that, projected onto the space spanned by the row scores, define a linear biplot. A proportional odds model is used, obtaining a multidimensional model similar to the graded response model in the IRT literature.

LOGISTIC BIPLOT FOR BINARY DATA

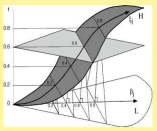
Let $\mathbf{X}_{IJ}$ be a binary data matrix. Let $\pi_{ij} = E(x_{ij})$ the expected probability that the variable j be present at individual i, and x_{ij} the observed probability, either 0 or 1. The *S-dimensional* logistic biplot in the logit scale is formulated as

$$logit(\pi_{ij}) = \log(\frac{\pi_{ij}}{1-\pi_{ij}}) = b_{j0} + \sum_{s=1}^{S} b_{js}a_{is} = b_{j0} + \mathbf{a}_i'\mathbf{b}_j, \qquad \left[\pi_{ij} = \frac{e^{b_{j0}+\sum_s b_{js}a_{is}}}{1+e^{b_{j0}+\sum_s b_{js}a_{is}}}\right]$$

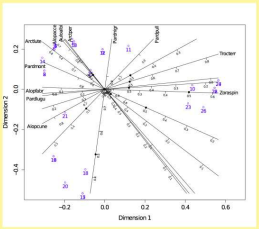
where a_{is} and b_{js} ($i=1, \ldots, I$; $j=1, \ldots, J$; $s=1, \ldots, S$) are the model parameters used as row and column markers respectively. The model is a generalized (bi)linear model having the logit as a link function. In matrix form

$$\log it(\Pi) = \mathbf{1}_I\mathbf{b}_0' + \mathbf{AB}'$$

Projecting a row marker $\mathbf{a}_i=(a_{i1}, a_{i2})$ onto a column marker $\mathbf{b}_j=(b_{j1}, b_{j2})$ it is possible to estimate the odds or the probability.



Logistic response surface and probability prediction points on the biplot axis.



Binary Logistic Biplot with prediction scales for the probabilities

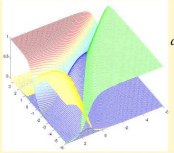
See Vicente-Villardón et al. (2006) and Demey et al. (2008).

LOGISTIC BIPLOT FOR NOMINAL DATA

Let $\mathbf{X}_{IJ}$ be a data matrix containing the values of J categorical variables -each with K_j ($j=1, \ldots, J$) categories- for I individuals, and let $\mathbf{G}_{IL}$ ($L = \sum_j K_j$) be the corresponding indicator matrix. The last category of each variable will be used as a baseline. Let $\pi_{ij(k)}$ the expected probability that the category k of variable j be present at individual i. In the multinomial logistic latent trait model with S latent traits

$$\pi_{ij(k)} = \frac{e^{b_{j(k)0}+\sum_{s=1}^{S} b_{j(k)s}a_{is}}}{1+\sum_{l=1}^{K_j-1} e^{b_{j(l)0}+\sum_{s=1}^{S} b_{j(l)s}a_{is}}}, (k = 1, \ldots, K_j - 1).$$

where a_{is} and $b_{j(k)s}$ ($i=1, \ldots, I$; $j=1, \ldots, J$; $k=1, \ldots, K_j-1$ $s=1, \ldots, S$) are the model parameters.

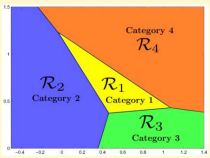


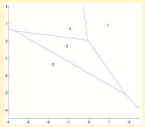
Response surfaces for a variable with four categories (The level curves predicting different probabilities for each category are no longer on straight lines, but the intersections are straight lines).

The intersections, projected on the subspace defined by the row markers, define a tessellation (A set of convex regions, as many as categories).

Each convex region predicts a category. The probability for that category, in the is higher than the probability of the rest.

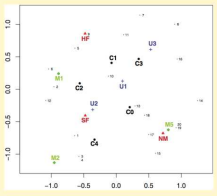
We search for a set of points (generators) whose Voronoi diagram is as close as possible to our "probability tessellation". Those points are called "category points".





The representation is interpreted in terms of distances in the sense that the category predicted for each individual is defined by the closest category point.

See Hernández-Sánchez & Vicente-Villardón (2013).



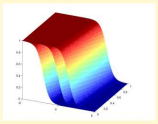
The final representation for several variables would be something like this.

LOGISTIC BIPLOT FOR ORDINAL DATA

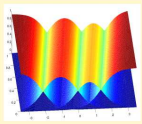
Let $\mathbf{X}_{IJ}$ be a data matrix containing the values of J ordinal variables -each with K_j ($j=1, \ldots, J$) categories- for I individuals, and let $\mathbf{G}_{IL}$ ($L = \sum_j K_j$) be the cumulative indicator matrix. The last category of each variable will be used as a baseline. Let $\pi^*_{ij(k)} = P(x_{ij} \le k)$ and $\pi_{ij(k)} = P(x_{ij}=k) = \pi^*_{ij(k)} - \pi^*_{ij(k-1)}$ (with $\pi^*_{ij(0)} = 0$). The ordinal logistic latent trait model for the cumulative probabilities is

$$\pi^*_{ij(k)} = \frac{1}{1+e^{-\left(d_{j(k)}+\sum_{s=1}^{S} a_{is}b_{js}\right)}} \qquad logit(\pi^*_{ij(k)}) = d_{j(k)} + \sum_{s=1}^{S} a_{is}b_{js} = d_{j(k)} + \mathbf{a}_i'\mathbf{b}_j, k = 1, \ldots, K_j - 1$$

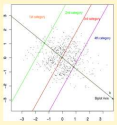
The equations define a biplot in the logit scale that shares the geometry of the binary case for each category. Each category have a different constant but the same slope, that means that the prediction direction is common to all categories and just the prediction markers are different. The parameters b define the direction of the projection; the representation subspace can be divided into prediction regions for each category, delimited by parallel straight lines.



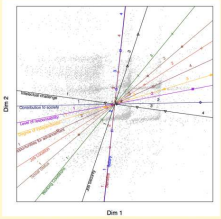
Cumulative probabilities.



Expected probabilities.



Prediction regions.



Ordinal Logistic Biplot: Predictions are obtained by projecting the individual points onto the directions for the variables.

See Vicente-Villardón & Hernández-Sánchez (2014).

PARAMETER ESTIMATION

- Alternated generalized regressions and interpolations. (Maximum Likelihood).
- Marginal Maximum Likelihood (As in Item Response Theory).
- Separation problem (Maximum likelihood does not converge). Penalized Maximum Likelihood.
- Heuristic approach for big data matrices: External Logistic Biplots (Logistic fits on the Principal Coordinates).

APPLICATIONS (mainly binary)

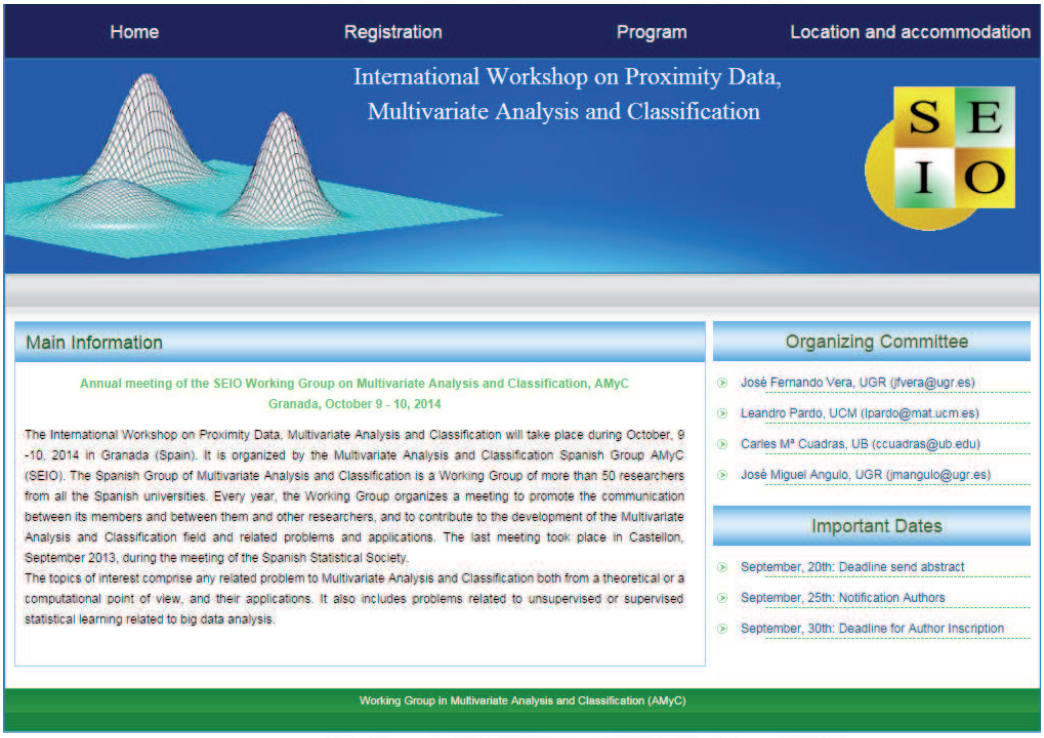
- HAPMAP Data (molecular markers). (Demey, et al., 2008)
- Sugar cane germoplasm. (Demey, 2008)
- Genetic diversity of Caricaceae family (García, et al. (2013)
- Greenhouse gas emissions (Gallego & Vicente-Villardón, 2012)
- Innovation profiles in Portugal. (Vicente et al., 2011)
- Job Satisfaction of Doctorate Degree Holders in Spain. Vicente-Villardón & Hernández-Sánchez, 2014).
- Many potential applications for categorical data.

REFERENCES

1.- VICENTE-VILLARDON, J. L., GALINDO M. P. & BLAZQUEZ, A. (2006). Logistic Biplots. In *"Múltiple Correspondence Análisis And Related Methods"*. Grenacre, M & Blasius, J. Eds. Chapman and Hall. Boca Ratón.
2.- DEMEY, J., VICENTE-VILLARDON, J. L., GALINDO, M.P. & ZAMBRANO, A. (2008) Identifying Molecular Markers Associated with Classification of Genotypes Using External Logistic Biplots. *Bioinformatics*, 24(24):2832-2838.
3.- HERNÁNDEZ SÁNCHEZ, J. C., & VICENTE-VILLARDÓN, J. L. (2013). Logistic biplot for nominal data. *arXiv preprint arXiv:1309.5486*.
4.- VICENTE-VILLARDÓN, J. L., & HENÁNDEZ-SÁNCHEZ, J. C. (2014). Logistic Biplots for Ordinal Data with an Application to Job Satisfaction of Doctorate Degree Holders in Spain. *arXiv preprint arXiv:1405.0294*.
5.- GALLEGO-ÁLVAREZ, L. & VICENTE-VILLARDÓN, J. L. (2012). Analysis of environmental indicators in international companies by applying the logistic biplot. *Ecological Indicators*, *23*, 250-261.
6.- VICENTE, P. , VAZ, T. D. N., & NIJKAMP, P. (2011). Institutional capacity to dynamically innovate: an application to the Portuguese case. *Technological Forecasting and Social Change*, *78*(1), 3-12.
7.- VICENTE-VILLARDÓN, J. L. (2011). Logistic Biplots for Binary, Nominal and Ordinal Data. *Agrocampus Ouest, Rennes FRANCE*, 63.
8.- R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.
9.- HERNÁNDEZ SÁNCHEZ, J. C., & VICENTE-VILLARDÓN, J. L. (2013). NominalLogisticBiplot: Biplots for Nominal Data in R. R package. Version 0.2.
10.- HERNÁNDEZ SÁNCHEZ, J. C., & VICENTE-VILLARDÓN, J. L. (2013). OrdinalLogisticBiplot: Biplots for Ordinal Data in R. R package. Version 0.3.
11.- VICENTE-VILLARDÓN, J. L. (2014). PrinCoord: Principal coordinates Analysis for binary, categorical, continuous or mixed data with External Logistic Biplots. R package version 0.1.0. http://biplot.usal.es/classicalbiplot/multbiplot-in-r/
12.- VICENTE-VILLARDÓN, J. L. (2014). PCABiplot: Classical PCA Biplot with added features. R package version 0.2.2. http://biplot.usal.es/classicalbiplot/multbiplot-in-r/
13.- GARCÍA, A. V., MILIANI, A., RODRÍGUEZ, D., ZAMBRANO, A. Y., VICENTE-VILLARDÓN, J. L., & DEMEY, J. R. (2013). Diversidad genética de la colección venezolana de la familia Caricáceas. *Interciencia*, *38*(3), 171-178.
14.- DEMEY, J. R. Diversidad genética en bancos de germoplasma: un enfoque biplot. Tesis Doctoral. Universidad de Salamanca.

Apéndice D

Ponencia en "Annual meeting of the SEIO Working Group on Multivariate Analysis and Classification, AMyC Granada, October 9-10, 2014": "Logistic Biplots for Nominal and Ordinal Data"


Home
Registration
Program
Location and accommodation
International Workshop on Proximity Data, Multivariate Analysis and Classification
S E I O
Main Information
Annual meeting of the SEIO Working Group on Multivariate Analysis and Classification, AMyC
Granada, October 9 - 10, 2014
The International Workshop on Proximity Data, Multivariate Analysis and Classification will take place during October, 9 -10, 2014 in Granada (Spain). It is organized by the Multivariate Analysis and Classification Spanish Group AMyC (SEIO). The Spanish Group of Multivariate Analysis and Classification is a Working Group of more than 50 researchers from all the Spanish universities. Every year, the Working Group organizes a meeting to promote the communication between its members and between them and other researchers, and to contribute to the development of the Multivariate Analysis and Classification field and related problems and applications. The last meeting took place in Castellon, September 2013, during the meeting of the Spanish Statistical Society.
The topics of interest comprise any related problem to Multivariate Analysis and Classification both from a theoretical or a computational point of view, and their applications. It also includes problems related to unsupervised or supervised statistical learning related to big data analysis.
Organizing Committee
José Fernando Vera, UGR (jfvera@ugr.es)
Leandro Pardo, UCM (lpardo@mat.ucm.es)
Carles Mª Cuadras, UB (ccuadras@ub.edu)
José Miguel Angulo, UGR (jmangulo@ugr.es)
Important Dates
September, 20th: Deadline send abstract
September, 25th: Notification Authors
September, 30th: Deadline for Author Inscription
Working Group in Multivariate Analysis and Classification (AMyC)

11.30 Invited Talk: Robust Statistical Inference based on the Density Power Divergence Approach

Leandro Pardo, *Universidad Complutense de Madrid*

In any parametric inference problem, the robustness of the procedure is a real concern. A procedure which retains a high degree of efficiency under the model and simultaneously provides stable inference under data contamination is preferable in any practical situation over another procedure which achieves its efficiency at the cost of robustness or vice versa. The density power divergence family provides a flexible class of divergences where the adjustment between efficiency and robustness is controlled by a single parameter. Robust estimation based on the minimization of density power divergences has proved to be a useful alternative to the classical maximum likelihood based technique. The most popular hypothesis testing procedure, the likelihood ratio test, is known to be highly non-robust in many real situations. An alternative robust procedure of hypothesis testing based on the density power divergence is presented.

12.30 Session 3

Chair: Eva Boj, Universitat de Barcelona

Merging classes

Josep A. Martín-Fernández, *Universitat de Girona*

In model based clustering, any element from a given sample is assigned to a particular class according to its posterior probability to belong to that class. Similarly, in fuzzy clustering such posterior probability is substituted by the weight of belonging to that class. In this presentation, we are going to introduce a general method to explore how these probabilities or weights of belonging may allow us to combine classes and build a hierarchy from a set of classes. Our approach is based on the log-ratio methodology for compositional data, as the posteriors probabilities or weights vectors can be viewed as compositions. Previous known methods to build hierarchies over classes will be discussed as special cases of our approach, and improved by incorporating new strategies.

Logistic Biplots for Nominal and Ordinal Data

José Luis Vicente, *Universidad de Salamanca*

The Biplot method is a popular technique for analysing multivariate data. Recently, Logistic Biplots for binary data have been developed. In this paper we extend the Logistic Biplot to nominal and ordinal data. For nominal data the variables are represented as convex prediction regions rather than vectors while for ordinal data as directions divided into prediction segments. We study the geometry of such representations and construct computational algorithms for estimation of the parameters and representation of the prediction regions or directions. Two R packages developed for the new methods and an application to a survey on job satisfaction of doctorate holders in Spain are also presented.

Claim reserving with DB-GLM: extending the Chain-Ladder method

Eva Boj, *Universitat de Barcelona*

As is demonstrated in the bibliography, generalized linear models (GLM) can be considered as a stochastic version of the classical Chain-Ladder method of claim reserving in non-life insurance. We refer, e.g., to England (1999) and England and Verrall (2002) for a detailed description. In particular, the deterministic Chain-Ladder model is reproduced when a GLM is fitted to a run-off-triangle by assuming overdispersed Poisson error distribution and logarithmic link. In this presentation, we propose the use of distance-based generalized linear models (DB-GLM) in the claim reserving problem. We refer to Boj et al. (2012) where the main characteristics of the DB-GLM are studied. DB-GLM can be considered a generalization of the classical GLM to the distance-based analysis. The only information required to fit these models is a predictor distance matrix. DB-GLM can be fitted using the dbstats package for R (Boj et al., 2013). It is important to point out that DB-GLM contains as a particular instance ordinary GLM. Then it can be considered too as a stochastic Chain-Ladder claim reserving method. To complement the methodology and estimate reserve distributions and standard errors we develop a bootstrap technique adequate to the DB-GLM. We make an application with the well known run-of-triangle of Taylor and Ashe (1983). This research is part of the project: Semiparametric and distance-based methodologies with applications in bioinformatics, finance and risk management (grant MTM2010-17323).

Apéndice E

Ponencia presentada en el congreso titulado "Conference of the International Federation of Classification Societies, ifcs Bolonia, Italy, July 6-8, 2015": "Prediction accuracy in Logistic Biplots for categorical data"

Prediction Accuracy in Logistic Biplots for categorical data.

Jose Luis Vicente-Villardon, Julio Cesar Hernandez-Sanchez

Last modified: 2015-05-19

Abstract

Classical biplot methods allow for the simultaneous representation of individuals (rows) and variables (columns) of a numerical data matrix. When data are binary, nominal or ordinal, classical linear biplots are not adequate; other techniques such as multiple correspondence analysis (MCA), latent trait analysis (LTA) or item response theory (IRT) for categorical items should be used instead.

We have recently extended the biplot to categorical data. The resulting method is termed `` logistic biplot''(LB) because the resulting procedure is related to logistic responses in the same way classical biplots are related to linear responses. For the nominal case, variables are represented as convex prediction regions rather than vectors; using the methods from computational geometry, the set of prediction regions is converted to a set of points in such a way that the prediction for each individual is established by its closest ``category point''.

Then interpretation is based on distances rather than on projections. For the binary and ordinal cases, the final representation is more like a traditional biplot with straight lines for predicting probabilities for each variable. The prediction regions are delimited by parallel straight lines and then a line with the adequate marks is enough to visualize the model.

We evaluate prediction accuracy of logistic biplots compared to MCA and IRT. The main differences between the LB and MCA are shown with data from demographic and labor market variables of doctorate (PdH) holders in the region of Castilla-Leon in Spain, using the 'Survey on the careers of doctorate holders (CDH)' carried out by Spanish Statistical Institute jointly with Eurostat, the Organization for Economic Co-operation and Development (OECD) and UNESCO's Institute for Statistics (UIS).

Keywords

logistic biplot; categorial data

References

Vicente-Villardón, J. L., Galindo Villardón, M. P., & Blázquez Zaballos, A. (2006): Logistic biplots. In: Greenacre, M & Blasius, J. (Eds.): *Multiple correspondence analysis and related methods.* Chapman & Hall, London, 503-521.

Demey, J. R., Vicente-Villardón, J. L., Galindo-Villardón, M. P., & Zambrano, A. Y. (2008): Identifying molecular markers associated with classification of genotypes by External Logistic Biplots. *Bioinformatics*, *24*, 24, 2832-2838.

Hernández Sánchez, J. C., & Vicente-Villardón, J. L. (2013): Logistic biplot for nominal data. *arXiv preprint arXiv:1309.5486*.

Vicente-Villardón, J. L., & Sánchez, J. C. H. (2014). Logistic Biplots for Ordinal Data with an Application to Job Satisfaction of Doctorate Degree Holders in Spain. *arXiv preprint arXiv:1405.0294*.

Gower, J. C., & Hand, D. J. (1995): *Biplots* (Vol. 54). CRC Press.

Gower, J. C., Lubbe, S. G., & Le Roux, N. J. (2011): *Understanding biplots*. John Wiley & Sons.

Sanchez, J. C. H., Vicente-Villardon, J. L., (2014): 'NominalLogisticBiplot': Logistic Biplot Representations for Nominal Data. R Package versión 0.4. http://cran.r-project.org/web/packages/NominalLogisticBiplot/index.html

Sanchez, J. C. H., Vicente-Villardon, J. L., (2015): 'OrdinalLogisticBiplot': Logistic Biplot Representations for Ordinal Data. R Package versión 0.4. http://cran.r-project.org/web/packages/OrdinalLogisticBiplot/index.html

Apéndice E. Ponencia "Prediction accuracy in Logistic Biplots for categorical data". ifcs, Bolonia.

Tuesday, 7th July 2015

9.00 – 10.00	**SP6: Density-based clustering** - G. Galimberti	**SP7: Accuracy and validation in clustering and scaling models I** - J. F. Vera	**SP8: Multi-Dimensional Scaling for Sparse Association Matrices** - T. Imaizumi	**SP9: Data science in Biomedical Research** - B. Lausen	**CONTR11: Change-point Detection**
	Geoff McLachlan, Sharon Lee	Eva Boj del Val, Teresa Costa Cor, Josep Fortiana Gregori	Tadashi Imaizumi	Hans Kestler	Jedelyn Cabrieto, Francis Tuerlinckx, Eva Ceulemans
	Giovanna Menardi, Domenico De Stefano	Jose Luis Vicente-Villardon, Julio Cesar Hernandez-Sanchez	Atsuho Nakayama	Matthias Schmid, Marvin Wright, Andreas Ziegler	Carlo Drago, Fabio Matano, Germana Scepi
	Surajit Ray	José Fernando Vera	Satoru Yokoyama	Berthold Lausen, Asma Gul, Zardad Khan, Osama Mahmoud	Kuniyoshi Hayashi, Koji Kurihara
10.15 – 11.00	**Plenary invited: Focused graphical model estimation (President's invited lecture) - Gerda Claeskens**				
11.00 – 11.30	Coffee Break				
11.30 – 13.10	**SP10: Cluster Analysis of Asymmetric Relationship** - A. Okada	**CONTR12: Dissimilarity and Distance Measures**	**CONTR13: Multiple Correspondence Analysis**	**CONTR14: Methods for DNA data**	**CONTR15: Applied classification and clustering**
	Innar Liiv	Ahmed Najeeb Albatineh	Barbara Batóg, Jacek Batóg, Wanda Skoczylas, Andrzej Niemiec, Piotr Waśniewski	Alia Dehman, Guillem Rigaill, Pierre Neuvial, Christophe Ambroise	Nicolas Greffard, Pascale Kuntz
	Facundo Memoli	Agnieszka Bernadetta Kozera, Aleksandra Łuczak, Feliks Wysocki	Belchin Kostov, Mónica Bécue-Bertaut, François Husson	Marcelo Ferreira, Ivan Costa	Guillaume Guex, Théophile Emmanouilidis, François Bavaud
	Miki Nakai	Gabriel Martos Venturini	Odysseas Moschidis, Theodore Chadjipadelis	Evgeny Mirkes, Thomas Walsh, Edward J Louis, Alexander N Gorban	Fuchen Liu, Yves Rozenholc, Charles-André Cuénod
	Akinori Okada, Satoru Yokoyama	Pascal Préa, François Brucker	Fionn Murtagh	Teppei Shimamura, Yusuke Matsui	Boris G. Mirkin, Mikhail A. Orlov
	Donatella Vicari	Zdenek Sulc, Hana Rezankova	Johané Nienkemper-Swanepoel, Sugnet Lubbe, Niël le Roux, Emilee Smith, Heather Zar, Mark Nicol	Makoto Tomita	Gabriella Schoier, Patrizia De Luca
13.10 – 14.10	Lunch Break				
14.10 – 15.40	**Presidential Address "Hierarchical Disjoint Non-negative Factor Analysis" – M. Vichi and Award Session**				
15.40 – 16.10	Coffee Break				
16.10 – 17.30	**SP11: Benchmarking in cluster analysis II** - I. Van Mechelen - 2	**SP12: New trends and applications of alternating least squares in data analysis** - Y. Mori, K. Adachi	**SP13: Analysis of multivariate longitudinal data** - P. Giordani	**CONTR16: Symbolic Data**	**CONTR17: Non gaussian mixture models and model selection**
	Friedrich Leisch and the IFCS Task Force on Benchmarking	Maria Brigida Ferraro, Paolo Giordani, Maurizio Vichi	Francesca Martella, Marco Alfò, Paolo Giordani	Carmela Cappelli, Pierpaolo D'Urso, Francesca Di Iorio	Ryan Browne
	Anne-Laure Boulesteix	Alfonso Iodice D'Enza, Michel van de Velden, Patrick J.F. Groenen	Mai Sherif Hafez, Irini Moustaki, Jouni Kuha	Yusuke Matsui, Teppei Shimamura	Brian C. Franczak, Ryan P. Browne, Paul D. McNicholas
	Nema Dean, Duncan Lee, and Craig Anderson	Michio Sakakihara, Msahiro Kuroda, Yuichi Mori, Msaya Iizuka	Joke Heylen, Iven Van Mechelen, Eiko Fried, Eva Ceulemans	Masahiro Mizuta, Hiroyuki Minami	Fumitake Sakaori
	Discussant: Rainer Dangl	Jun Tsuchida, Hiroshi Yadohisa	Marieke Timmerman, Eva Ceulemans, Henk Kiers	Marcin Pelka, Aneta Rybicka, Justyna Brzezinska-Grabowska	Cristina Tortora
17.30 – 19.00	Guided tour Bologna / Council Meeting				
20.15	**Conference Dinner** at Palazzo Albergati				

Apéndice F

Ponencia presentada en el congreso "Correspondence Analysis and Related Methods", Naples, Italy, September 20-23, 2015": "A comparison Between Nominal Logistic Biplots and Multiple Correspondence Analysis"

Apéndice F. Ponencia "A comparison Between Nominal Logistic Biplots and Multiple Correspondence Analysis". Nápoles.

Home
Organisers
Registration
Submission
Venue
Information participants
Program
SVD workshop

Important announcement: the new deadline for abstract submission is June 14, 2015.

The objective of this conference is to spotlight the very latest research in correspondence analysis and related methods (CARME) of multidimensional visualization, as well as to discuss future developments. We aim to bring together theoretical and applied researchers in all the areas where correspondence analysis and related methods are currently being used, notably sociology, psychology, education, ecology, archaeology, geology, linguistics, philosophy, genetics, biomedical research, health economics, marketing and management. Interdisciplinary contributions will be particularly welcome.

The conference will take place at the seaside in the **conference centre "Federico II"**, which is an historic building of the University of Naples, seen on the left in the above panoramic photo, and below (building, Aula Magna and entrance).



Invited speakers:

- Eric J. Beh
- John Gower
- Carlo Lauro
- Ludovic Lebart
- Brigitte Le Roux
- Fionn Murtagh
- Rosanna Verde

Themes of the conference include all forms of correspondence analysis and related fields:

- Simple correspondence analysis
- Multiple correspondence analysis
- Joint correspondence analysis
- Multiway correspondence analysis
- Canonical correspondence analysis
- Nonsymmetrical correspondence analysis
- Dual scaling
- Optimal scaling
- Homogeneity analysis
- Geometric data analysis
- Multidimensional scaling of categorical data
- Visualization of categorical data
- Visualization of compositional data
- Correspondence analysis in the social sciences
- Correspondence analysis in ecology and the environmental sciences
- Correspondence analysis in the health sciences
- Correspondence analysis in management and marketing research
- Correspondence analysis in food and sensory research

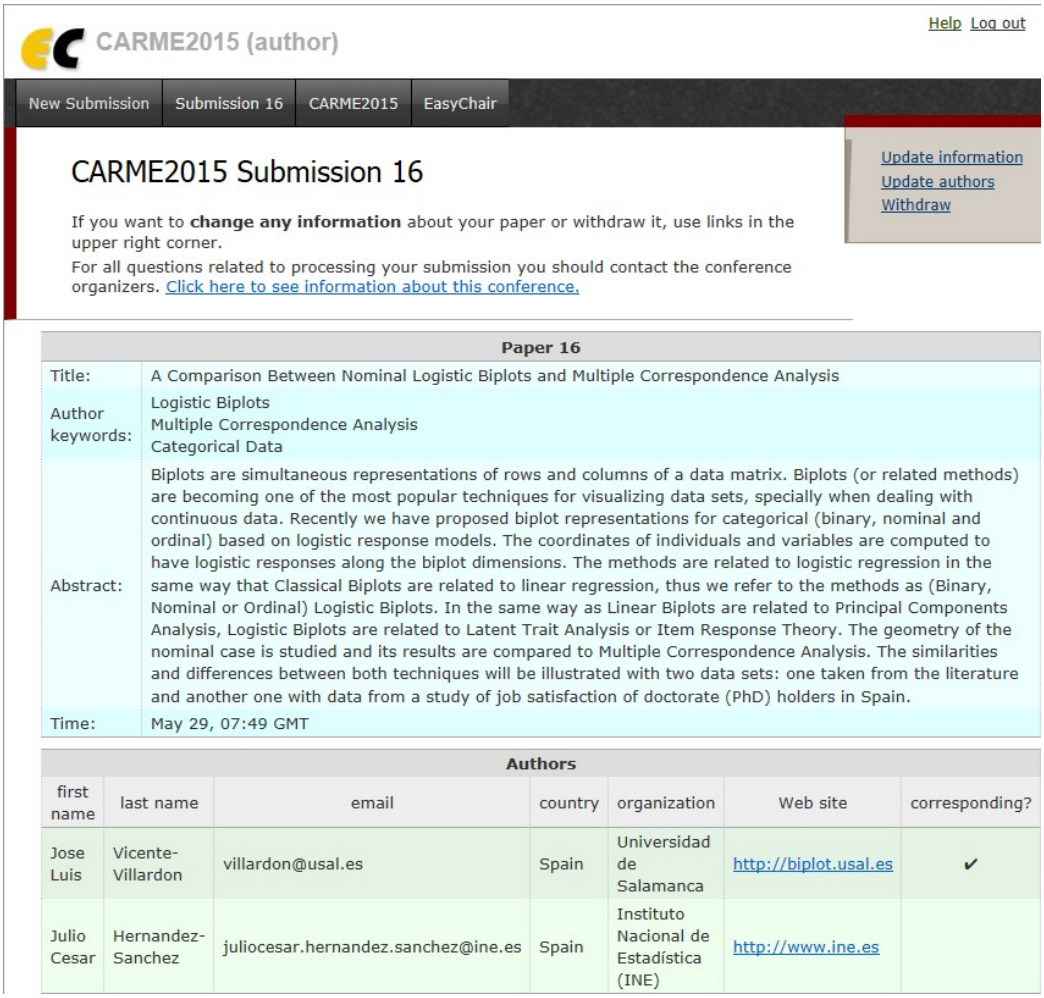
Workshop on the Singular Value Decomposition (SVD)

Directly following the main CARME conference we plan an additional event, a one-day workshop on the singular value decomposition (SVD), with guest speaker **Trevor Hastie**. You can see details of this workshop here or by clicking on the "SVD workshop" button on the menu on the left.

Important Dates

- May 31, 2015: Submission abstract
- June 14, 2015: Notification of acceptance of conference papers
- Before July 27, 2015: Early registration fee
- After July 27, 2015: Normal registration fee

Società Italiana di Statistica fondata nel 1939 SIS
ASA Associazione per la Statistica Applicata Applied Statistics Association
Università degli Studi di Napoli Federico II Dipartimento di Scienze Economiche e Statistiche



CARME2015 (author)

Help Log out

New Submission | Submission 16 | CARME2015 | EasyChair

Update information
Update authors
Withdraw

CARME2015 Submission 16

If you want to **change any information** about your paper or withdraw it, use links in the upper right corner.
For all questions related to processing your submission you should contact the conference organizers. Click here to see information about this conference.

Paper 16	
Title:	A Comparison Between Nominal Logistic Biplots and Multiple Correspondence Analysis
Author keywords:	Logistic Biplots Multiple Correspondence Analysis Categorical Data
Abstract:	Biplots are simultaneous representations of rows and columns of a data matrix. Biplots (or related methods) are becoming one of the most popular techniques for visualizing data sets, specially when dealing with continuous data. Recently we have proposed biplot representations for categorical (binary, nominal and ordinal) based on logistic response models. The coordinates of individuals and variables are computed to have logistic responses along the biplot dimensions. The methods are related to logistic regression in the same way that Classical Biplots are related to linear regression, thus we refer to the methods as (Binary, Nominal or Ordinal) Logistic Biplots. In the same way as Linear Biplots are related to Principal Components Analysis, Logistic Biplots are related to Latent Trait Analysis or Item Response Theory. The geometry of the nominal case is studied and its results are compared to Multiple Correspondence Analysis. The similarities and differences between both techniques will be illustrated with two data sets: one taken from the literature and another one with data from a study of job satisfaction of doctorate (PhD) holders in Spain.
Time:	May 29, 07:49 GMT

Authors

first name	last name	email	country	organization	Web site	corresponding?
Jose Luis	Vicente-Villardon	villardon@usal.es	Spain	Universidad de Salamanca	http://biplot.usal.es	✔
Julio Cesar	Hernandez-Sanchez	juliocesar.hernandez.sanchez@ine.es	Spain	Instituto Nacional de Estadística (INE)	http://www.ine.es	

Apéndice F. Ponencia "A comparison Between Nominal Logistic Biplots and Multiple Correspondence Analysis". Nápoles.

Tuesday 22 September 2015

9.30 **Contributed papers**

Related methods - methodology (Chair: *Michael Friendly*)	**Textual data analysis** (Chair: *Eric Beh*)
Jose Luis Vicente-Villardon and Julio Cesar Hernandez-Sanchez **A comparison between nominal logistic biplots and multiple correspondence analysis**	*Mireille Gettler Summa, Sadika Rjiba, Myriam Touati and Saloua Benamou* **A visualization approach for textual analysis using Gath-Geva algorithm, fuzzy framework and symbolic data analysis**
Sugnet Lubbe **Including covariates in linear discriminant analysis**	*Maria Gabriella Grassia, Marina Marino, Enrica Amaturo, Biagio Aragona, Gabriella Punziano and Rosanna Cataldo* **New media communication strategies for the election campaign. Campania regional election, 2015**
Antoine de Falguerolles **Similar or/and different? The case of classical multidimensional scaling of two matched distance matrices**	*Massimo Aria, Corrado Cuccurullo and Fabrizia Sarto* **Performance management in business and public administration domains: a co-word analysis on 20 Years**
Elias Moreno, F. Javier Giron, M. Lina Martinez and Carles Cuadras **Heterogeneity: A Bayesian model**	*Kazuo Fujimoto* **On publishing the Japanese translation of "Applied correspondence analysis" and its comment part**

11:00 ***Coffee***

11.30 **Invited papers** (Chair: *Patrick Groenen*)

Fionn Murtagh **The semantic context of aggregated data in high dimensional, massive data analytics**
Carlo Natale Lauro, Pasquale Dolce, Vincenzo Esposito Vinzi **Non-symmetrical approach to component based SEM**

12.45 ***Lunch***

14.00 **Contributed papers**

Clustering methods (Chair: *Fionn Murtagh*)	**Social space analysis** (Chair: *Brigitte Le Roux*)
Ndèye Niang, Stéphanie Bougeard, Gilbert Saporta and Hervé Abdi **Clusterwise multiblock PLS**	*Jan Thorhauge Frederiksen* **Bureauratic and political transformations of the Danish field of welfare work**
Michel van de Velden, Alfonso Iodice D'Enza and Francesco Palumbo **Cluster correspondence analysis**	*Philippe Bonnet and Frédéric Lebaron* **Discourses in turmoil? A lexicometric-GDA study of French political debate about the European budget**
Dario Bruzzese, Davide Passaretti and Domenico Vistocco **DESPOTA: an algorithm to automatically detect a reliable partition on a dendrogram**	*Alice Barth and Andreas Schmitz* **Assessing correspondences between psychological and sociological indicators with MCA**

15.15 ***Coffee***

16.00 **Contributed papers**

MCA & related methods - new developments (Chair: *Maurizio Vichi*)	**CA & MCA - applications** (Chair: *Sugnet Lubbe*)
Jean-Luc Durand **On the variance of eigenvalues in PCA and MCA**	*Luigi D'Ambra and Jules de Tibeiro* **Different approaches to analyse familial budgets with external information**
Angelos Markos and Alfonso Iodice D' Enza. **Joint dimension reduction and fuzzy clustering**	*Natascha Kienstra and Peter van der Heijden* **Using correspondence analysis in multiple case studies**
Patrick Groenen and Julie Josse **Multinomial correspondence analysis**	*Ulaş Akkucuk and Mehmet Artemel* **Visualisation of patent data: a study on seven Turkic countries**
Michael Friendly and David Meyer **General models and graphs for log odds ratios**	*Emmanouil Bagkeris, Heather Bailey, Ruslan Malyuta, Claire Thorne and Mario Cortina Borja* **Advantages of a composite marker of socioeconomic status over educational level alone: modelling cART initiation for PMTCT in the Ukraine European Collaborative Study using joint correspondence analysis**

Apéndice G

Manual del paquete de R NominalLogisticBiplot

Package 'NominalLogisticBiplot'

July 2, 2014

Type Package

Title Biplot representations of categorical data

Version 0.2

Date 2014-05-01

Author Julio Cesar Hernandez Sanchez, Jose Luis Vicente-Villardon

Maintainer Julio Cesar Hernandez Sanchez <juliocesar_avila@usal.es>

Description Analysis of a matrix of polytomous items using Nominal Logistic Biplots (NLB) according to Hernandez-Sanchez and Vicente-Villardon (2013).
The NLB procedure extends the binary logistic biplot to nominal (polytomous) data.
The individuals are represented as points on a plane and the variables are represented as convex prediction regions rather than vectors as in a classical or binary biplot. Using the methods from Computational Geometry, the set of prediction regions is converted to a set of points in such a way that the prediction for each individual is established by its closest ``category point". Then interpretation is based on distances rather than on projections.
In this package we implement the geometry of such a representation and construct computational algorithms
for the estimation of parameters and the calculation of prediction regions.

License GPL (>= 2)

Encoding latin1

Repository CRAN

Depends R (>= 2.15.1),mirt,gmodels,MASS

LazyData yes

Archs i386, x64

NeedsCompilation no

Date/Publication 2014-05-02 07:13:20

R topics documented:

`NominalLogisticBiplot-package`

Nominal Logistic Biplot representations for polytomous data

Description

Analysis of a matrix of polytomous items using Nominal Logistic Biplots (NLB) according to Hernandez-Sanchez & Vicente-Villardon (2013). The NLB procedure extends the binary logistic biplot to nominal (polytomous) data. The individuals are represented as points on a plane and the variables are represented as convex prediction regions rather than vectors as in a classical or binaly biplot. Using the methods from the Computational Geometry, the set of prediction regions is converted to a set of points in such a way that the prediction for each individual is established by its closest "category point". Then interpretation is based on distances rather than on projections. In this package we implement the geometry of such a representation and construct computational algorithms for the estimation of parameters and the calculation of prediction regions

Details

Package:	NominalLogisticBiplot
Type:	Package
Version:	1.0
Date:	2013-08-05
License:	GPL (>=2)

Author(s)

Julio Cesar Hernandez Sanchez, Jose Luis Vicente-Villardon Maintainer: Julio Cesar Hernandez Sanchez <juliocesar_avila@usal.es>

See Also

NominalLogisticBiplot,NominalLogBiplotEM,multiquad,summary.nominal.logistic.biplot,plot.nominal.logis

Examples

```
data(HairColor)
nlbo = NominalLogisticBiplot(HairColor,sFormula=NULL,numFactors=2,
method="EM",penalization=0.2,show=FALSE)
summary(nlbo)
plot(nlbo,QuitNotPredicted=TRUE,ReestimateInFocusPlane=TRUE,
planex = 1,planey = 2,proofMode=TRUE,LabelInd=TRUE,AtLeastR2 = 0.01
,xlimi=-1.5,xlimu=1.5,ylimi=-1.5,ylimu=1.5,linesVoronoi = TRUE
,SmartLabels = FALSE, PlotInd=TRUE,CexInd = c(0.6,0.7,0.5,0.4,0.5,0.6,0.7)
,PchInd = c(1,2,3,4,5,6,7),ColorInd="black",PlotVars=TRUE,LabelVar = TRUE
,PchVar = c(1,2,3,4,5),ColorVar = c("red","black","yellow","blue","green")
,ShowResults=TRUE)
```

`afc` *Simple Correspondence Analysis*

Description

This function calculates simple correspondence analysis for a data matrix.

Usage

```
afc(x, dim = 2, alpha = 1)
```

Arguments

`x`	A frequency matrix or a binary matrix obtained from the original data set of nominal variables.
`dim`	Number of dimensions for the solution
`alpha`	Biplot weight for rows and columns. 1 means rows in principal coordinates and columns in standard coordinates, 0 means rows in standard coordinates and columns in principal coordinates.

Value

An object of class `"afc.sol"`.This has some components:

`Title`	Title of the statistical technique
`Non_Scaled_Data`	Original data
`Minima`	vector with the minimum values for each column of the initial data matrix
`Maxima`	vector with the maximum values for each column of the initial data matrix
`Initial_Transformation`	Name of the transformation for the data
`Scaled_Data`	Scaled data according to the transformation
`nrows`	Number of rows of the data set
`ncols`	Number of columns of the data set
`dim`	Number of dimensions for the solution
`CumInertia`	Acumulated Inertia
`Scale_Factor`	Scale factor for the transformation
`RowCoordinates`	Coordinates for the individuals in the reduced dimension space
`ColCoordinates`	Coordinates for the variables in the reduced dimension space
`RowContributions`	Contributions of the dimensions to explain the inertia of each row
`ColContributions`	Contributions of the dimensions to explain the inertia of each column
`Inertia`	Inertia for each dimension
`Eigenvalues`	Eigenvalues

Author(s)

Jose Luis Vicente-Villardon,Julio Cesar Hernandez Sanchez

Maintainer: Jose Luis Vicente-Villardon <villardon@usal.es>

References

BENZECRI, J.P. (1973) *L'analyse des Donnees*. Vol. 2. L'analyse des correspondences. Dunod. Paris.

See Also

`NominalMatrix2Binary`

Examples

```
  data(HairColor)
  G = NominalMatrix2Binary(data.matrix(HairColor))
mca=afc(G,dim=2)
  mca
```

`Env` *Ecological Factors in Farm Management.*

Description

The farms Env data frame has 20 rows and 4 columns. The rows are farms on the Dutch island of Terschelling and the columns are factors describing the management of grassland.

Usage

```
data(Env)
```

Format

This data frame contains the following columns:

Mois five levels of soil moisture,although level 3 does not occur in the data. Levels are labelled M1, M2, M4 and M5.

Manag Grassland management type (SF = standard farming, BF = biological farming, HF = hobby farming, NM = nature conservation management).

Use Grassland use (U1 = it exists production, U2 = intermediate,U3 = grazing).

Manure Manure usage (C0,C1,C2,C3 and C4)

Source

J.C. Gower and D.J. Hand (1996) *Biplots*. Chapman & Hall, Table 4.6.

Quoted as from:
R.H.G. Jongman, C.J.F. ter Braak and O.F.R. van Tongeren (1987) *Data Analysis in Community and Landscape Ecology.* PUDOC, Wageningen.

References

Venables, W. N. and Ripley, B. D. (2002) *Modern Applied Statistics with S.* Fourth edition. Springer.

Examples

```
data(Env)
```

`Generators`	*Generators (points) of the tesselation generated by a nominal variable.*

Description

With the parameters resulting from fitting a nominal logistic model to the row scores for a given variable, the function calculates all the information necessary to plot the tessellation generated by the fit. The final user will not normally use this function.

Usage

```
Generators(beta)
```

Arguments

`beta`	Matrix with the estimated parameters for a given nominal variable. It has as many rows as the number of categories minus one and three columns (one for the constant and other two for the x-y coordinates on the plane).

Value

An object of class `"voronoiprob"`. This has the components:

`x`	x-coordinates for the real points (Vertices of the tessellation).
`y`	y-coordinates for the real points (Vertices of the tessellation).
`n1`	vector with the first neighbours of the real points
`n2`	vector with the second neighbours of the real points
`n3`	vector with the third neighbours of the real points
`dummy.x`	x-coordinates for the dummy points
`dummy.y`	y-coordinates for the dummy points
`ndummy`	Number of dummies
`IndReal`	Matrix with the indices of each real point in the tessellation
`Centers`	Matrix with the points resulting from inverting the tessellation
`hideCat`	Vector to indicate if there are some hidden categories
`equivRegiones`	Matrix with the new re-numbered categories (when some are hidden)

Author(s)

Julio Cesar Hernandez Sanchez, Jose Luis Vicente-Villardon

Maintainer: Julio Cesar Hernandez Sanchez <juliocesar_avila@usal.es>

References

Hern\'andez S\'anchez, J. C., & Vicente-Villard\'on, J. L. (2013). Logistic biplot for nominal data. arXiv preprint arXiv:1309.5486.

Gower, J. & Hand, D. (1996), *Biplots, Monographs on statistics and applied probability* 54. London: Chapman and Hall., 277 pp.

Evans, D. & Jones, S. (1987), *Detecting voronoi (area of influence) polygons* ,Mathematical Geology 19(6), 523–537.

Hartvigsen, D. (1992), *Recognizing voronoi diagrams with linear programming*, ORSA Journal on Computing 4, 369–374.

Schoenberg, F., Ferguson, T. & Li, C. (2003), *Inverting dirichlet tesselations*, Computer journal 46(1), 76–83.

Examples

```
data(HairColor)
data = data.matrix(HairColor)
xEM = NominalLogBiplotEM(data, dim = 2,showResults = FALSE)
nomreg = polylogist(data[,2],xEM$RowCoordinates[,1:2],penalization=0.1)
tesselation = Generators(nomreg$beta)
tesselation
```

`HairColor` *Demographic Data*

Description

The sample data corresponds to 7 people and shows some demographic characteristics.

Usage

```
data(HairColor)
```

Format

This data frame contains 7 observation for the following 5 columns:

Sex two levels (M=male,F=female)

HairColor four levels of hair color (Dark, Grey, Fair and Brown)

Region (E = England,S = Scotland, W = Wales)

Work (Manual,Clerical,Professional)

Education (School,Univ,Postgrad)

Source

Gower, J., Gardner-Lubbe,S., Le Roux,N. (2011). “Understanding Biplots.” *Wiley*.

Examples

```
data(HairColor)
```

`hermquad` *Gauss-Hermite quadrature*

Description

Computes the Hermite Quadrature weights for a set of grid points

Usage

```
hermquad(N)
```

Arguments

`N`	Number of nodes for the quadrature

Value

An object of class `"GaussQuadrature"`. This has the components:

`X`	Coordinates of the nodes
`W`	Weights asociated to each node

Author(s)

Jose Luis Vicente-Villardon,Julio Cesar Hernandez Sanchez

Maintainer: Jose Luis Vicente-Villardon <villardon@usal.es>

References

Stroud, A.H. and Secrest, D. (1966) *Gaussian Quadrature Formulas*, Englewood Cliffs, NJ: Prentice-Hall.

Hildebrand,F. B. (1987) *Intoduction to Numerical Analysis 2nd Ed*, Dover Publications, New York, page 385

Examples

```
hermquad(10)
```

`multiquad` *Multidimensional Gauss-Hermite quadrature*

Description

This function computes the gauss-hermite quadrature in more than one dimension.

Usage

```
multiquad(nnodos, dims)
```

Arguments

`nnodos`	Number of nodes.
`dims`	Number of dimensions of the quadrature

Value

An object of class `"MultiGaussQuadrature"`. This has the components:

`X`	Coordinates of the nodes
`A`	Weights asociated to each node

Author(s)

Jose Luis Vicente-Villardon,Julio Cesar Hernandez Sanchez

Maintainer: Jose Luis Vicente-Villardon <villardon@usal.es>

References

Jackel, P. (2005) *A note on multivariate Gauss-Hermite quadrature*

http://www.pjaeckel.webspace.virginmedia.com/ANoteOnMultivariateGaussHermiteQuadrature.pdf

See Also

`hermquad`

Examples

```
multiquad(10,2)
```

`Nominal2Binary` *Transformation of a nominal variable into a binary indicator matrix*

Description

This function transforms a nominal variable into a binary matrix with as many colums as categories. Each row of the matrix has a value of 1 for the corresponding level of the category and 0 elsewhere.

Usage

```
Nominal2Binary(y)
```

Arguments

`y`	A vector containing the values of nominal variable measured on a set of individuals- The values must be integers starting at 1.

Value

An object of type matrix:

`Z`	The binary indicator matrix asociated to the nominal variable

Author(s)

Jose Luis Vicente-Villardon,Julio Cesar Hernandez Sanchez

Maintainer: Jose Luis Vicente-Villardon <villardon@usal.es>

Examples

```
data(HairColor)
Nominal2Binary(as.numeric(HairColor[,1]))
```

`NominalDistances` *Distances between indidividuals calculated from nominal variables.*

Description

This function calculates the hamming distances (or similarities) among individuals from a nominal data matrix.

Usage

```
NominalDistances(x, similarities = FALSE)
```

Arguments

`x`	This parameter is a matrix with the nominal variables
`similarities`	Boolean parameter to specify if the user wants a distances matrix or a similarities matrix. By default this parameter is FALSE, so the function calculates the distances.

Value

The function returns a matrix with the distances or similarities

Author(s)

Jose Luis Vicente-Villardon,Julio Cesar Hernandez Sanchez

Maintainer: Jose Luis Vicente-Villardon <villardon@usal.es>

References

Boriah, S., Chandola, V. & Kumar,V.(2008) *Similarity measures for categorical data: A comparative evaluation*. In proceedings of the eight SIAM International Conference on Data Mining, pp 243–254

Examples

```
data(HairColor)
NominalDistances(data.matrix(HairColor))
```

`NominalLogBiplotEM` *Alternated EM algorithm for Nominal Logistic Biplots*

Description

This function computes, with an alternated algorithm, the row and column parameters of a Nominal Logistic Biplot for polytomous data. The row coordinates (E-step) are computed using multidimensional Gauss-Hermite quadratures and Expected *a posteriori* (EAP) scores and parameters for each variable or items (M-step)using Ridge Nominal Logistic Regression to solve the separation problem present when the points for different categories of a variable are completely separataed on the representation plane and the usual fitting methods do not converge. The separation problem is present in almost avery data set for which the goodness of fit is high.

Usage

```
NominalLogBiplotEM(x, dim = 2, nnodos = 10, tol = 1e-04,
maxiter = 100, penalization = 0.2,initial=1,alfa=1, Plot = FALSE,
showResults = FALSE)
```

Arguments

x	Matrix with the nominal data. The matrix must be in numerical form.
dim	Dimension of the solution
nnodos	Number of nodes for the multidimensional Gauss-Hermite quadrature
tol	Value to stop the process of iterations.
maxiter	Maximum number of iterations in the process of solving the regression coefficients.
penalization	Penalization used in the diagonal matrix to avoid singularities.
initial	Value to decide the method(1-Correspondence analysis, 2-Mirt) that calculates the initial abilities values for the individuals.
alfa	If initial parameter method is correspondence analysis, this parameter determines the weight for rows and columns.
Plot	Boolean parameter to plot the row coordinates
showResults	Boolean parameter to show all the information about the iterations.

Value

An object of class "nominal.logistic.biplot.EM".This has components:

RowCoordinates	Coordinates for the individuals in the reduced space
ColumnModels	List with information about the Nominal Logistic Models calculated for each variable including: estimated parameters with covariances and standard errors, log-likelihood, deviances, percents of correct classifications, pvalues and pseudo-Rsquared measures

Author(s)

Julio Cesar Hernandez Sanchez, Jose Luis Vicente-Villardon

Maintainer: Julio Cesar Hernandez Sanchez <juliocesar_avila@usal.es>

References

Bock,R. & Aitkin,M. (1981),*Marginal maximum likelihood estimation of item parameters: Aplication of an EM algorithm*, Phychometrika 46(4), 443-459.

Gabriel, K. R. (1998). Generalised bilinear regression. Biometrika, 85(3), 689-700.

Vicente-VillardÃ³n, J. L., Galindo VillardÃ³n, M. P., & BlÃ¡zquez Zaballos, A. (2006). Logistic biplots. Multiple correspondence analysis and related methods. London: Chapman & Hall, 503-521.

Gabriel, K. R., & Zamir, S. (1979). Lower rank approximation of matrices by least squares with any choice of weights. Technometrics, 21(4), 489-498.

See Also

polylogist, multiquad

Examples

```
data(HairColor)
data = data.matrix(HairColor)
xEM = NominalLogBiplotEM(data, dim = 2,showResults = FALSE)
xEM
```

`NominalLogisticBiplot` *Nominal Logistic Biplot for polytomous data*

Description

Function that calculates the parameters of the Nominal Logistic Biplot according to Hernandez-Sanchez & Vicente-Villardon (2013).

Usage

```
NominalLogisticBiplot(datanom, sFormula = NULL, numFactors = 2,
method = "EM", rotation = "varimax", metfsco = "EAP",
nnodos = 10, tol = 1e-04, maxiter = 100, penalization = 0.1,
cte = TRUE,initial=1,alfa=1, show = FALSE)
```

Arguments

`datanom`	The data set, it can be a *matrix* with integers or a *data frame* with factors. All variables have to be nominal.
`sFormula`	This parameter follows the unifying interface for selecting variables from a data frame for a plot, test or model. The most common formula es of type y ~ x1+x2+x3. It has a default value of NULL if not specified.
`numFactors`	Number of dimensions of the solution. It should be lower than the number of variables. It has a default value of 2.
`method`	This parameter can be: "EM", "ACM", "MIRT" or "PCOA". Method to compute the row coordinates.
`rotation`	Rotation method to used with "MIRT" option in "coordinates". No effect fror other options.
`metfsco`	Calculation method for the fscores with "MIRT" option in "coordinates". No effect fror other options.
`nnodos`	Number of nodes for gauss quadrature in the EM algorithm.
`tol`	Tolerance for the EM algorithm.
`maxiter`	Maximum number of iterations in the EM algorithm.
`penalization`	Penalization for the ridge regression for each variable.
`cte`	Include constant in the logistic regression model. Default is TRUE.
`initial`	Value to decide the method(1-Correspondence analysis, 2-Mirt) that calculates the initial abilities values for the individuals.
`alfa`	If initial parameter method is correspondence analysis, this parameter determines the weight for rows and columns.
`show`	Show intermediate copmputations. Default is TRUE.

Details

The general algorithm used is essentially an alternated procedure in which parameters for rows and columns are computed in alternated steps repeated until convergence. Parameters for the rows are calculated by expectation (E-step) or by a external procedure (Multiple Correspondence Analysis or Principal Coordinates Analysis) and parameters for the columns are computed by maximization (M-step), i. e., by Nominal Logistic Regression. When the procedure for Row scores is external, only one iteration is performed and the procedure is called "External Nominal Logistic Biplot". Because the aim of the biplot is the representation

There are several options for the computation:

1.- Using the package **mirt** to obtain the row scores, i. e. using a solution obtained from a latent trait model. The column (item) parameters should be directly used by our biplot procedure but, because of the characteristics of the package that performs a default rotation after parameter estimation, we have to reestimate the item parametes to be coherent to the scores.

2.- Using our implementation of the EM algorithm alternating expected a porteriori scores and Ridge Nominal Logistic Regression for each variable.

3.- Using external coordinates for the rows taken from Multiple Correspondence Analysis or Principal Coordinates Analysis and fitting the response surfaces in just one step.

Equations that define a set of probability response surfaces (one for each category and each variable) are no longer sigmoid as in the binary case (Vicente-Villardon et al. (2006)). This means that the level curves are no longer straight lines and then, prediction of probabilities is not made by projection as in the usual linear biplots. For each variable, define a set of convex polygons that can be interpreted as "prediction" regions in the same way as in Gower & Hand (1996). Each pair of response surfaces defined by intersect in a straight line that, projected onto the space of predictors, is the set of points in which the probability of both categories is the same. Those lines are the candidates to be the edges of the convex polygons defining the prediction regions.

Value

An object of class `"nominal.logistic.biplot"`. This has some components:

`dataSet`	Data set of study with all the information about the name of the levels and names of the variables and individuals
`RowCoords`	Coordinates for the individuals in the reduced space
`VariableModels`	Information of the regression resuls for each variable.
`NumFactors`	Number of dimensions selected for the study
`Method`	Method for calculating the row positions
`Rotation`	Type of rotation if we have chosen mirt coordinates
`Methodfscores`	Method of calculation of the fscores in mirt process
`NumNodos`	Number of nodes for the gauss quadrature in EM algorith
`tol`	Cut point to stop the EM-algorith
`maxiter`	Maximum number of iterations in the EM-algorith
`penalization`	Value for the correction of the ridge regression

`cte`	Boolean value to choose if the model for each variable will have independent term
`show`	Boolean value to indicate if we want to see the results of our analysis

Author(s)

Julio Cesar Hernandez Sanchez, Jose Luis Vicente-Villardon

Maintainer: Julio Cesar Hernandez Sanchez <juliocesar_avila@usal.es>

References

Hernandez, J. C. & Vicente-Villardon, J. L. (2013) Logistic Biplots for Nominal Data. Submitted. Preprint available at : https://www.researchgate.net/publication/256428288_Logistic_Biplot_for_Nominal_Data?ev=prf_pub

Vicente-Villardon, J., Galindo, M.P & Blazquez-Zaballos, A. (2006), *Logistic biplots*,Multiple Correspondence Analysis and related methods pp. 491–509.

Demey, J., Vicente-Villardon, J. L., Galindo, M.P. & Zambrano, A. (2008) *Identifying Molecular Markers Associated With Classification Of Genotypes Using External Logistic Biplots*. Bioinformatics, 24(24), 2832-2838.

Baker, F.B. (1992): Item Response Theory. Parameter Estimation Techniques. Marcel Dekker. New York.

Gabriel, K. (1971), *The biplot graphic display of matrices with application to principal component analysis.*, Biometrika 58(3), 453–467.

Gabriel, K. R. (1998), *Generalised bilinear regression*, Biometrika 85(3), 689–700.

Gabriel, K. R. & Zamir, S. (1979), *Lower rank approximation of matrices by least squares with any choice of weights*, Technometrics 21(4), 489–498.

Gower, J. & Hand, D. (1996), *Biplots, Monographs on statistics and applied probability*. 54. London: Chapman and Hall., 277 pp.

Chalmers,R,P (2012). *mirt: A Multidimensional Item Response Theory Package for the R Environment*. Journal of Statistical Software, 48(6), 1-29. URL http://www.jstatsoft.org/v48/i06/.

See Also

`NominalLogBiplotEM`, `afc`, `PCoA`

Examples

```
data(HairColor)
nlbo = NominalLogisticBiplot(HairColor,sFormula=NULL,
numFactors=2,method="EM",penalization=0.2,show=FALSE)
nlbo

#data(PhD_nomCyL)
#cyl = NominalLogisticBiplot(PhD_nomCyL,sFormula=NULL,
#numFactors=2,method="EM",initial = 1,penalization=0.3,show=FALSE)
#summary(nlboPhD)
#plot(nlboPhD,QuitNotPredicted=TRUE,ReestimateInFocusPlane=TRUE,
```

```
#      planex = 1,planey = 2,proofMode=TRUE,LabelInd=FALSE,AtLeastR2 = 0.01,
#     xlimi=-1.5,xlimu=1.5,ylimi=-1.5,ylimu=1.5,linesVoronoi = TRUE,SmartLabels = FALSE,
#      PlotInd=TRUE,
#      CexInd = c(0.4),
#      PchInd = c(1),
#      ColorInd="azure3",
#      PlotVars=TRUE,LabelVar = TRUE,
#      PchVar = c(1,2,3,4,5,6,7,8,9),ColorVar = c("red","black","maroon","blue","green",
#      "chocolate4","coral3","brown","brown2"),
#      ShowResults=TRUE)
```

`NominalMatrix2Binary` *Indicator matrix of a set of nominal variables.*

Description

Constructs the indicator matrix for a nominal variables matrix.

Usage

```
NominalMatrix2Binary(Y)
```

Arguments

`Y`	A matrix with nominal variables measured for a set of individuals. Input must be a matrix with integer values.

Value

`G`	The binary indicator matrix asociated to the nominal matrix

Author(s)

Jose Luis Vicente-Villardon,Julio Cesar Hernandez Sanchez

Maintainer: Jose Luis Vicente-Villardon <villardon@usal.es>

See Also

`Nominal2Binary`

Examples

```
data(HairColor)
NominalMatrix2Binary(data.matrix(HairColor))
```

`PCoA` *Principal Coordinates Analysis*

Description

This function calculates principal coordinates analysis using a distante matrix among a set of objets.

Usage

```
PCoA(dis, r = 2)
```

Arguments

`dis`	Distance matrix between a set ob objects.
`r`	Number of dimensions for the solution.

Value

An object with has some components:

`EigenValues`	Eigenvalues of the inner products matrix
`Inertia`	Variance (Inertia) accounted for each dimension
`RowCoordinates`	Coordinates for the rows in the reduced space
`RowQualities`	Qualities of representation of the objects. Squared cosines between the points (vectors) in the full space and the points in the reduced space. Values near 1 indicate good quality

Author(s)

Jose Luis Vicente-Villardon,Julio Cesar Hernandez Sanchez

Maintainer: Jose Luis Vicente-Villardon <villardon@usal.es>

References

Gower,J.C. (1966) *Some distance properties of latent root and vector methods used in multivariate analysis*. Biometrika, 53, 325–338.

See Also

`NominalDistances`

Examples

```
data(HairColor)
dis = NominalDistances(data.matrix(HairColor))
PCoA(dis,2)
```

PhD_nomCyL *Data set from Survey on Human Resources in Science and Technology carried out by Spanish Statistical Office.*

Description

The sample data corresponds 681 answers from PhD holders, corresponding to people that in 2006 had a doctoral degree and with their residence in Castilla-Le\'on region in Spain. The data come from Survey on Careers of doctorate holders(CDH) carried out by Spanish Statistical Office in 2008.

Usage

```
data(PhD_nomCyL)
```

Format

This data frame contains 681 observation for the following 7 columns:

MS Marital Status:(1:M(Married),2:MLR(Living in a marriage-like relationship), 3:SD (Separated or Divorced),4:SW(Widowed or Single)

SECT Sector of employment(1:BES(Business Enterprise Sector), 2:GS (Government Sector),3:HES(Higher Education Sector), 4:PNP(Private Non Profit))

MIN Minimum education level required for the principal job: (1:mPD(Postdoc),2:mARQ(Advanced Research Qualification), 3:mPG(Post-graduate),4:mGL(Graduate or lower)

DES Desirable education level required for the principal job: (1:dPD(Postdoc),2:dARQ(Advanced Research Qualification), 3:dPG(Post-graduate),4:dGL(Graduate or lower)

PJREL Is your principal job related to your advanced research qualification degree:(1:H(High),2:M(Medium),3:L(Low))

FOSAT Field of science and technology (1:NS(Natural Sciences),2:ET(Engineering and technology), 3:MH(Medical and health sciences),4:AS(Agricultural sciences), 5:SS(Social Sciences),6:H(Humanities))

SOF Principal source of financial support during your research studies (1:F(Fellowship),2:T(Teaching),3:OE (Other Employment),4:R(Reimbursement) ,5:LPSO (LoanPersonalSavingsOther)

Source

Spanish Statistical Office (Survey on Human Resources in Science and Technology, 2006): http://www.ine.es/prodyser/micro_

Examples

```
data(PhD_nomCyL)
```

`plot.nominal.logistic.biplot`
Graphical representation of a Nominal Logistic Biplot.

Description

Plotting a Nominal Logistic Biplot. There are parameters related to the way in which the biplot is plotted. All the posible parameters have default values

Usage

```
## S3 method for class 'nominal.logistic.biplot'
## S3 method for class 'nominal.logistic.biplot'
plot(x, planex = 1, planey = 2,
QuitNotPredicted = TRUE, ReestimateInFocusPlane = TRUE,
proofMode = FALSE, AtLeastR2 = 0.01, xlimi = -1.5, xlimu = 1.5,
ylimi = -1.5, ylimu = 1.5, linesVoronoi = FALSE, ShowAxis = TRUE,
PlotVars = TRUE, PlotInd = TRUE, LabelVar = TRUE, LabelInd = TRUE,
CexInd = NULL, CexVar = NULL, ColorInd = NULL, ColorVar = NULL,
SmartLabels = FALSE, PchInd = NULL, PchVar = NULL,
LabelValuesVar = NULL, ShowResults = FALSE,...)
```

Arguments

`x`	An object of the class nominal.logistic.biplot.
`planex`	Dimension for X axis.
`planey`	Dimension for Y axis.
`QuitNotPredicted`	Should the non-predicted categories be represented on the graph?
`ReestimateInFocusPlane`	Shuld the item parameters be reestimated using only the dimensiona of the plot.? If FALSE the values of the parameters for other dimensions are set to 0. Default is FALSE
`proofMode`	Should each variable be plotted on a separate plot? If FALSE, a single plot with a legend for identifying each variable is made.
`AtLeastR2`	It establishes the cutting value to plot a variable attending to its Nagelkerke R^2 value. A variable is plotted if its R^2 is higher than this value.
`xlimi`	Minimum value on the x-axis.
`xlimu`	Maximum value on the x-axis.
`ylimi`	Minimum value on the y-axis.
`ylimu`	Maximum value on the y-axis.
`linesVoronoi`	Should the tesselation be plotted.? Default is FALSE and only the category points are plotted for a better reading of the plot.

ShowAxis	Should the axis be shown?
PlotVars	Should the variables (items) be pplotted?
PlotInd	Should the individuals be plotted?
LabelVar	Should the variable labels be shown?
LabelInd	Should the individual labels be shown?
CexInd	Size of the individual points. It can be an array with the cex information for each row.
CexVar	Size of the category points. It can be an array with the cex information for each variable.
ColorInd	Color of the individual points. It can be an array with the color information for each row.
ColorVar	Color for the variables. It can be an array with the color information for each variable.
SmartLabels	Should the text labels be printed accordind to its position on the plot?.
PchInd	Symbol for the individuals. It can be an array with the pch information for each row.
PchVar	Symbol for the variables. It could be an array with the pch information for each variable.
LabelValuesVar	List with the text labels for all the variables. If NULL, initial labels are used.
ShowResults	Should the results of the proccess of calculating the prediction regions be shown?
...	Additional parameters to plot.

Details

The function without parameters plots the nominal.logistic.biplot object with labels in the original data and default values for colors, symbols and sizes for points and lines. Other values of colors, symbols and sizes can be supplied. A single value applies to all the points but an array with different values can be used to improve the undestanding of the plot.-

Author(s)

Julio Cesar Hernandez Sanchez, Jose Luis Vicente-Villardon

Maintainer: Julio Cesar Hernandez Sanchez <juliocesar_avila@usal.es>

See Also

NominalLogisticBiplot

Examples

```
data(HairColor)
nlbo = NominalLogisticBiplot(HairColor,sFormula=NULL,
numFactors=2,method="EM",penalization=0.2,show=FALSE)
plot(nlbo,QuitNotPredicted=TRUE,ReestimateInFocusPlane=TRUE,
planex = 1,planey = 2,proofMode=TRUE,LabelInd=TRUE,
```

```
AtLeastR2 = 0.01,xlimi=-1.5,xlimu=1.5,ylimi=-1.5,
ylimu=1.5,linesVoronoi = TRUE,SmartLabels = FALSE,
PlotInd=TRUE,CexInd = c(0.6,0.7,0.5,0.4,0.5,0.6,0.7)
,PchInd = c(1,2,3,4,5,6,7),ColorInd="black",PlotVars=TRUE,
LabelVar = TRUE,PchVar = c(1,2,3,4,5),
ColorVar = c("red","black","yellow","blue","green")
,ShowResults=TRUE)
```

`plotNominalFittedVariable`

Function for plotting in the reduced space an unordered and fitted categorical variable.

Description

Graphical representation of a polytomous unordered variable previously fitted in the reduced space, according to the Nominal Logistic Biplot theory. It can be choosen some parameters related to the way in which the variable is plotted.

Usage

```
plotNominalFittedVariable(nameVar, numcateg, beta, varstudyC, rowCoords,
levelsVar = NULL, numFactors = 2, planex = 1, planey = 2, xi = -3.5, xu = 3.5,
yi = -3.5, yu = 3.5, CexVar = 0.7,ColorVar = "blue", PchVar = 0.7,
addToPlot = FALSE, QuitNotPredicted = TRUE, ShowResults = TRUE,
linesVoronoi = TRUE, LabelVar = TRUE)
```

Arguments

`nameVar`	Name of the variable to be plotted.
`numcateg`	Number of categories of the variable.
`beta`	Estimated coefficients matrix.
`varstudyC`	Values of the categorical variable to be plotted. It should be a factor with information about a nominal variable, i.e., an unordered variable.
`rowCoords`	Estimation coordinates for the individuals in the spanned space.
`levelsVar`	Vector with the labels for each level of the variable.
`numFactors`	Dimension of the reduced space.
`planex`	Dimension for X axis.
`planey`	Dimension for Y axis.
`xi`	Minimum value on the x-axis.
`xu`	Maximum value on the x-axis.
`yi`	Minimum value on the y-axis.
`yu`	Maximum value on the y-axis.
`CexVar`	Size of the category points.

ColorVar	Color for the variable.
PchVar	Symbol for the variable.
addToPlot	Should the graph be added to an existing representation?
QuitNotPredicted	Should the non-predicted categories be represented on the graph?
ShowResults	Should the results of the proccess of calculating the prediction regions be shown?
linesVoronoi	Should the tesselation be plotted.? Default is FALSE and only the category points are plotted for a better reading of the plot
LabelVar	Should the variable labels be shown?

Author(s)

Julio Cesar Hernandez Sanchez, Jose Luis Vicente-Villardon

Maintainer: Julio Cesar Hernandez Sanchez <juliocesar_avila@usal.es>

See Also

polylogist

Examples

```
data(Env)
nlbo = NominalLogisticBiplot(Env,sFormula=NULL,
numFactors=2,method="EM",penalization=0.2,show=FALSE)
nameVar = nlbo$dataSet$ColumNames[1]
numcateg = 4
beta = nlbo$VariableModels[,1]$beta
Nagelkerke = nlbo$VariableModels[,1]$Nagelkerke
varstudyC = as.matrix(as.numeric(Env[,1]))
rowCoords = nlbo$RowsCoords
levelsVar = c("M1","M2","M4","M5")
plotNominalFittedVariable(nameVar,numcateg,beta,varstudyC,rowCoords,levelsVar=NULL,
          numFactors=2,planex = 1,planey = 2,xi=-3.5,xu=3.5,yi=-3.5,yu=3.5,
          CexVar=0.7,ColorVar="blue",PchVar=0.7,addToPlot=FALSE,
          QuitNotPredicted=TRUE,ShowResults=TRUE,linesVoronoi=TRUE,LabelVar=TRUE)
```

`plotNominalVariable` *Function for plotting in the reduced space an unordered categorical variable.*

Description

Graphical representation of a polytomous unordered variable in the reduced space, according to the Nominal Logistic Biplot theory. Inside the function, the estimations needed for the variable will be done. It can be choosen some parameters related to the way in which the variable is plotted.

Usage

```
plotNominalVariable(nameVar, nominalVar, estimRows, planex = 1, planey = 2,
xi = -3.5, xu = 3.5, yi = -3.5, yu = 3.5, CexVar = 0.7, ColorVar = "blue",
PchVar = 0.7, addToPlot = FALSE, QuitNotPredicted = TRUE, ShowResults = FALSE,
linesVoronoi = TRUE, LabelVar = TRUE, tol = 1e-04, maxiter = 100,
penalization = 0.1, showIter = FALSE)
```

Arguments

`nameVar`	Name of the variable to be plotted.
`nominalVar`	Values of the categorical variable to be plotted. It should be a factor with information about a nominal variable, i.e., a variable without ordered values.
`estimRows`	Estimation coordinates for the individuals in the spanned space.
`planex`	Dimension for X axis.
`planey`	Dimension for Y axis.
`xi`	Minimum value on the x-axis.
`xu`	Maximum value on the x-axis.
`yi`	Minimum value on the y-axis.
`yu`	Maximum value on the y-axis.
`CexVar`	Size of the category points.
`ColorVar`	Color for the variable.
`PchVar`	Symbol for the variable.
`addToPlot`	Should the graph be added to an existing representation?
`QuitNotPredicted`	Should the non-predicted categories be represented on the graph?
`ShowResults`	Should the results of the proccess of calculating the prediction regions be shown?
`linesVoronoi`	Should the tesselation be plotted.? Default is FALSE and only the category points are plotted for a better reading of the plot
`LabelVar`	Should the variable labels be shown?
`tol`	Value to stop the process of iterations.
`maxiter`	Maximum number of iterations in the process of solving the regression coefficients.
`penalization`	Penalization used in the diagonal matrix to avoid singularities.
`showIter`	Boolean parameter to show the information about the iterations.

Author(s)

Julio Cesar Hernandez Sanchez, Jose Luis Vicente-Villardon

Maintainer: Julio Cesar Hernandez Sanchez <juliocesar_avila@usal.es>

See Also

`polylogist`

Examples

```
data(HairColor)
nlbo = NominalLogisticBiplot(HairColor,sFormula=NULL,
numFactors=2,method="EM",penalization=0.2,show=FALSE)
nameVar = nlbo$dataSet$ColumNames[2]
nominalVar = HairColor[,2]
estimRows = nlbo$RowsCoords
plotNominalVariable(nameVar,nominalVar,estimRows,planex = 1,planey = 2,
    xi=-1.5,xu=1.5,yi=-1.5,yu=1.5,CexVar=0.7,ColorVar="blue",PchVar=0.7,
    addToPlot=FALSE,QuitNotPredicted=TRUE,ShowResults=TRUE,
    linesVoronoi=TRUE,LabelVar=TRUE,tol = 1e-04, maxiter = 100,
    penalization = 0.3,showIter = FALSE)
```

`polylogist` *Multinomial logistic regression with ridge penalization*

Description

This function does a logistic regression between a dependent variable y and some independent variables x, and solves the separation problem in this type of regression using ridge regression and penalization.

Usage

```
polylogist(y, x, penalization = 0.2, cte = TRUE, tol = 1e-04, maxiter = 200, show = FALSE)
```

Arguments

`y`	Dependent variable.
`x`	A matrix with the independent variables.
`penalization`	Penalization used in the diagonal matrix to avoid singularities.
`cte`	Should the model have a constant?
`tol`	Tolerance for the iterations.
`maxiter`	Maximum number of iterations.
`show`	Should the iteration history be printed?.

Details

The problem of the existence of the estimators in logistic regression can be seen in Albert (1984), a solution for the binary case, based on the Firth's method, Firth (1993) is proposed by Heinze(2002). The extension to nominal logistic model was made by Bull (2002). All the procedures were initially developed to remove the bias but work well to avoid the problem of separation. Here we have chosen a simpler solution based on ridge estimators for logistic regression Cessie(1992).

Rather than maximizing $L_j(\mathbf{G}|\,\mathbf{b}_{j0}, \mathbf{B}_j)$ we maximize

$$L_j(\mathbf{G}|\,\mathbf{b}_{j0},\mathbf{B}_j) - \lambda\,(\|\mathbf{b}_{j0}\| + \|\mathbf{B}_j\|)$$

Changing the values of λ we obtain slightly different solutions not affected by the separation problem.

Value

An object of class `"polylogist"`. This has components

`fitted`	Matrix with the fitted probabilities
`cov`	Covariance matrix among the estimates
`Y`	Indicator matrix for the dependent variable
`beta`	Estimated coefficients for the multinomial logistic regression
`stderr`	Standard error of the estimates
`logLik`	Logarithm of the likelihood
`Deviance`	Deviance of the model
`AIC`	Akaike information criterion indicator
`BIC`	Bayesian information criterion indicator

Author(s)

Julio Cesar Hernandez Sanchez, Jose Luis Vicente-Villardon

Maintainer: Julio Cesar Hernandez Sanchez <juliocesar_avila@usal.es>

References

Albert,A. & Anderson,J.A. (1984),*On the existence of maximum likelihood estimates in logistic regression models*, Biometrika 71(1), 1–10.

Bull, S.B., Mak, C. & Greenwood, C.M. (2002), *A modified score function for multinomial logistic regression*, Computational Statistics and dada Analysis 39, 57–74.

Firth, D.(1993), *Bias reduction of maximum likelihood estimates*, Biometrika 80(1), 27–38

Heinze, G. & Schemper, M. (2002), *A solution to the problem of separation in logistic regression*, Statistics in Medicine 21, 2109–2419

Le Cessie, S. & Van Houwelingen, J. (1992), *Ridge estimators in logistic regression*, Applied Statistics 41(1), 191–201.

Examples

```
data(HairColor)
data = data.matrix(HairColor)
G = NominalMatrix2Binary(data)
mca=afc(G,dim=2)
depVar = data[,1]
nomreg = polylogist(depVar,mca$RowCoordinates[,1:2],penalization=0.1)
nomreg
```

`RidgeMultinomialRegression`
Ridge Multinomial Logistic Regression

Description

Function that calculates an object with the fitted multinomial logistic regression for a nominal variable. It compares with the null model, so that we will be able to compare which model fits better the variable.

Usage

```
RidgeMultinomialRegression(y, x, penalization = 0.2,
cte = TRUE, tol = 1e-04, maxiter = 200, showIter = FALSE)
```

Arguments

`y`	Dependent variable.
`x`	A matrix with the independent variables.
`penalization`	Penalization used in the diagonal matrix to avoid singularities.
`cte`	Should the model have a constant?
`tol`	Value to stop the process of iterations.
`maxiter`	Maximum number of iterations.
`showIter`	Should the iteration history be printed?.

Value

An object that has the following components:

`fitted`	Matrix with the fitted probabilities
`cov`	Covariance matrix among the estimates
`Y`	Indicator matrix for the dependent variable
`beta`	Estimated coefficients for the multinomial logistic regression
`stderr`	Standard error of the estimates
`logLik`	Logarithm of the likelihood
`Deviance`	Deviance of the model
`AIC`	Akaike information criterion indicator
`BIC`	Bayesian information criterion indicator
`NullDeviance`	Deviance of the null model
`Difference`	Difference between the two deviance values
`df`	Degrees of freedom
`p`	p-value asociated to the chi-squared estimate

`CoxSnell`	Cox and Snell pseudo R squared
`Nagelkerke`	Nagelkerke pseudo R squared
`MacFaden`	MacFaden pseudo R squared
`PercentCorrect`	Percentage of correct classifications

Author(s)

Julio Cesar Hernandez Sanchez, Jose Luis Vicente-Villardon

Maintainer: Julio Cesar Hernandez Sanchez <juliocesar_avila@usal.es>

References

Albert,A. & Anderson,J.A. (1984),*On the existence of maximum likelihood estimates in logistic regression models*, Biometrika 71(1), 1–10.

Bull, S.B., Mak, C. & Greenwood, C.M. (2002), *A modified score function for multinomial logistic regression*, Computational Statistics and dada Analysis 39, 57–74.

Firth, D.(1993), *Bias reduction of maximum likelihood estimates*, Biometrika 80(1), 27–38

Heinze, G. & Schemper, M. (2002), *A solution to the problem of separation in logistic regression*, Statistics in Medicine 21, 2109–2419

Le Cessie, S. & Van Houwelingen, J. (1992), *Ridge estimators in logistic regression*, Applied Statistics 41(1), 191–201.

See Also

`polylogist`

Examples

```
data(HairColor)
data = data.matrix(HairColor)
G = NominalMatrix2Binary(data)
mca=afc(G,dim=2)
depVar = data[,1]
rmr = RidgeMultinomialRegression(depVar,mca$RowCoordinates[,1:2],penalization=0.1)
rmr
```

`summary.nominal.logistic.biplot` *Summary Method Function for Objects of Class 'nominal.logistic.biplot'*

Description

This function shows a summary of the principal results for the estimation for individuals and variables, like some Pseudo R-squared indices, the correct classification percentage of each regression, the logLikelihood and "Estimate coefficients", "Std. Error", "z value" or "Pr(>|z|)" values.

Usage

```
## S3 method for class 'nominal.logistic.biplot'
## S3 method for class 'nominal.logistic.biplot'
summary(object,completeEstim,coorInd,...)
```

Arguments

`object`	This parameter keeps the nominal logistic biplot object.
`completeEstim`	Boolean parameter to choose if the estimated coefficients will be printed on screen.Default value is FALSE.
`coorInd`	Boolean parameter to choose if the individual coordinates will be printed on screen.Default value is FALSE.
`...`	Additional parameters to summary.

Details

This function is a method for the generic function summary() for class "nominal.logistic.biplot". It can be invoked by calling summary(x) for an object x of the appropriate class.

Author(s)

Julio Cesar Hernandez Sanchez, Jose Luis Vicente-Villardon

Maintainer: Julio Cesar Hernandez Sanchez <juliocesar_avila@usal.es>

See Also

`NominalLogisticBiplot`

Examples

```
data(HairColor)
nlbo = NominalLogisticBiplot(HairColor,sFormula=NULL,
numFactors=2,method="EM",penalization=0.2,show=FALSE)
summary(nlbo)
```

Index

Apéndice H

Manual del paquete de R OrdinalLogisticBiplot

Package 'OrdinalLogisticBiplot'

January 16, 2015

Type Package

Title Biplot representations of ordinal variables

Version 0.4

Date 2015-15-01

Author Julio Cesar Hernandez Sanchez, Jose Luis Vicente-Villardon

Maintainer Julio Cesar Hernandez Sanchez <juliocesar_avila@usal.es>

Description Analysis of a matrix of polytomous items using Ordinal Logistic Biplots (OLB)
The OLB procedure extends the binary logistic biplot to ordinal (polytomous) data.
The individuals are represented as points on a plane and the variables are represented
as lines rather than vectors as in a classical or binary biplot, specifying the points
for each of the categories of the variable.
The set of prediction regions is established by stripes perpendicular to the line
between the category points, in such a way that the prediction for each individual is given
by its projection into the line of the variable.

License GPL (>=2)

Encoding latin1

Repository CRAN

Depends R (>= 2.15.1),mirt,MASS,NominalLogisticBiplot

LazyData yes

Archs i386, x64

NeedsCompilation no

R topics documented:

`OrdinalLogisticBiplot-package`
Ordinal Logistic Biplot representations for polytomous ordered data.

Description

Analysis of a matrix of polytomous ordered items using Ordinal Logistic Biplots (OLB). The OLB procedure extends the binary logistic biplot to ordinal (polytomous) data.

Details

Package:	OrdinalLogisticBiplot
Type:	Package
Version:	0.4
Date:	2015-01-16
License:	GPL (>=2)

Author(s)

Julio Cesar Hernandez Sanchez, Jose Luis Vicente-Villardon Maintainer: Julio Cesar Hernandez Sanchez <juliocesar_avila@usal.es>

See Also

`OrdinalLogisticBiplot`,`OrdinalLogBiplotEM`

Examples

```
data(LevelSatPhd)
olbo = OrdinalLogisticBiplot(LevelSatPhd,sFormula=NULL,numFactors=2,
method="EM",penalization=0.2,show=FALSE)
summary(olbo)
plot(olbo,PlotInd=TRUE,xlimi=-1,xlimu=1,ylimi=-1,ylimu=1,margin = 0.2,
ColorVar = c("red","green","black","blue","yellow"),CexVar = c(0.7),showIIC=FALSE)
```

`BiplotDensity` *Density plot of a data set with overlaid contours.*

Description

This function draws for a set of points a density contour lines plot. The densities can be calculated for the whole set of points or for the groups defined by a nominal variable.

Usage

```
BiplotDensity(X, y = NULL, nlevels = max(y), grouplabels = 1:nlevels,
  ncontours = 6, groupcols = 1:nlevels, img = TRUE, separate = FALSE,
  ncolors = 20, ColorType = 4, xliml = -1, xlimu = 1, yliml = -1,
  ylimu = 1, plotInd = FALSE)
```

Arguments

`X`	A matrix with the coordinates for the plane in which the the contour lines will be plotted.
`y`	Categorical variable used for defining clusters. If NULL, the density is calculated for the whole set of points.
`nlevels`	Number of clusters.
`grouplabels`	Set of labels for the centers of each cluster. It should be a vector with "nlevels" components.
`ncontours`	Number of contours that will be used in the representation.
`groupcols`	Vector whith a set of colors for the clusters.
`img`	Should the density be plotted (with different colors) together with the contour lines?. Default value is TRUE.
`separate`	Should the density for each cluster be represented on a different picture?. Default value is FALSE.
`ncolors`	Number of colors for the densities.
`ColorType`	Type of color schema for the density image. It should be a number between 1 and 5.
`xliml`	Minimum value on the x-axis.
`xlimu`	Maximum value on the x-axis.
`yliml`	Minimum value on the y-axis.
`ylimu`	Maximum value on the y-axis.
`plotInd`	Should the individuals be plotted?

Author(s)

Julio Cesar Hernandez Sanchez, Jose Luis Vicente-Villardon

Maintainer: Julio Cesar Hernandez Sanchez <juliocesar_avila@usal.es>

Examples

```
data(LevelSatPhd)
olbo = OrdinalLogisticBiplot(LevelSatPhd)
x = olbo$RowCoords[, 1]
y = olbo$RowCoords[, 2]
plot(x,y, cex = 0, xlim=c(-1,1),ylim=c(-1,1))
X = olbo$RowCoords
y = as.matrix(as.numeric(LevelSatPhd[,4]))
gcols = c("midnightblue","black","red","gray87")
BiplotDensity(X,y,groupcols = gcols)
```

`CheckDataSet` *Check a data set.*

Description

This function checks if a data set is a data frame or a matrix and it saves the data as a matrix of integers, and stores the names of rows, colums and levels for each variable as vectors to use them later.

Usage

```
CheckDataSet(datanom)
```

Arguments

`datanom`	It can be a data frame or a matrix.

Details

The function checks if some variable has NA values and it deletes the corresponding row. It also checks for missing categories and recodifies the variable keeping the original labels for levels.

Value

An object of class `"data.ordinal"`. This has components:

`datanom`	Matrix of integers with the values of the variables
`RowNames`	Vector with the names of the rows
`ColumNames`	Vector with the names of the variables
`LevelNames`	Levels of each variable

Author(s)

Julio Cesar Hernandez Sanchez, Jose Luis Vicente-Villardon

Maintainer: Julio Cesar Hernandez Sanchez <juliocesar_avila@usal.es>

Examples

```
data(LevelSatPhd)
dataChecked = CheckDataSet(LevelSatPhd)
```

`LevelSatPhd` *Data set extracted from the Careers of doctorate holders survey carried out by Spanish Statistical Office in 2008.*

Description

The sample data, as part of a large survey, corresponds to 100 people who have the PhD degree and it shows the level of satisfaction of the doctorate holders about some issues.

Usage

```
data(LevelSatPhd)
```

Format

This data frame contains 100 observation for the following 5 ordinal variables, with four categories each: (1= "Very Satisfied", 2= "Somewhat Satisfied",3="Somewhat dissatisfied", 4="Very dissatisfied")

Salary

Benefits

Job Security

Job Location

Working conditions

Source

Spanish Statistical Institute. Survey of PDH holders, 2006. URL: http://www.ine.es.

Examples

```
data(LevelSatPhd)
```

`OrdinalLogBiplotEM` *Alternated EM algorithm for Ordinal Logistic Biplots*

Description

This function computes, with an alternated algorithm, the row and column parameters of an Ordinal Logistic Biplot for ordered polytomous data. The row coordinates (E-step) are computed using multidimensional Gauss-Hermite quadratures and Expected *a posteriori* (EAP) scores and parameters for each variable or items (M-step) using Ridge Ordinal Logistic Regression to solve the separation problem present when the points for different categories of a variable are completely separated on the representation plane and the usual fitting methods do not converge. The separation problem is present in almost every data set for which the goodness of fit is high.

Usage

```
OrdinalLogBiplotEM(x,dim = 2, nnodos = 15, tol = 0.001, maxiter = 100,
                   penalization = 0.2,show=FALSE,initial=1,alfa=1)
```

Arguments

x	Matrix with the ordinal data. The matrix must be in numerical form.
dim	Dimension of the solution.
nnodos	Number of nodes for the multidimensional Gauss-Hermite quadrature.
tol	Value to stop the process of iterations.
maxiter	Maximum number of iterations in the process of solving the regression coefficients.
penalization	Penalization used in the diagonal matrix to avoid singularities.
show	Boolean parameter to specify if the user wants to see every iteration.
initial	Method used to choose the initial ability in the algorithm. Default value is 1.
alfa	Optional parameter to calculate row and column coordinates in Simple correspondence analysis if the initial parameter is equal to 1.

Value

An object of class "ordinal.logistic.biplot.EM".This has components:

RowCoordinates	Coordinates for the rows or individuals
ColumnParameters	List with information about the Ordinal Logistic Models calculated for each variable including: estimated parameters with thresholds, percents of correct classifications,and pseudo-Rsquared
loadings	factor loadings
LogLikelihood	Logarithm of the likelihood
r2	R squared coefficient
Ncats	Number of the categories of each variable

Author(s)

Jose Luis Vicente-Villardon, Julio Cesar Hernandez Sanchez

Maintainer: Julio Cesar Hernandez Sanchez <juliocesar_avila@usal.es>

References

Bock,R. & Aitkin,M. (1981),*Marginal maximum likelihood estimation of item parameters: Aplication of an EM algorithm*, Phychometrika 46(4), 443-459.

See Also

pordlogist

Examples

```
data(LevelSatPhd)
dataSet = CheckDataSet(LevelSatPhd)
datanom = dataSet$datanom
olb = OrdinalLogBiplotEM(datanom,dim = 2, nnodos = 10,
      tol = 0.001, maxiter = 100, penalization = 0.2)
olb
```

`OrdinalLogisticBiplot` *Ordinal Logistic Biplot for ordered polytomous data*

Description

Function that calculates the parameters of the Ordinal Logistic Biplot.

Usage

```
OrdinalLogisticBiplot(datanom,sFormula=NULL,numFactors=2,
method="EM",rotation="varimax",metfsco="EAP",
nnodos = 10, tol = 1e-04, maxiter = 100,
penalization = 0.1,cte=TRUE, show=FALSE,ItemCurves = FALSE,initial=1,alfa=1)
```

Arguments

`datanom`	The data set; it can be a *matrix* with integers or a *data frame* with factors. All variables have to be ordinal.
`sFormula`	This parameter follows the unifying interface for selecting variables from a data frame for a plot, test or model. The most common formula is of type y ~ x1+x2+x3. It has a default value of NULL if it is not specified.
`numFactors`	Number of dimensions of the solution. It should be lower than the number of variables. It has a default value of 2.
`method`	This parameter can be: "EM" or "MIRT". Method to compute the row coordinates.
`rotation`	Rotation method to used with "MIRT" option in "coordinates". No effect fror other options.
`metfsco`	Calculation method for the fscores with "MIRT" option in "coordinates". No effect for other options.
`nnodos`	Number of nodes for gauss quadrature in the EM algorithm.
`tol`	Tolerance for the EM algorithm.
`maxiter`	Maximum number of iterations in the EM algorithm.
`penalization`	Penalization for the ridge regression for each variable.
`cte`	Include constant in the logistic regression model. Default is TRUE.
`show`	Show intermediate computations. Default is FALSE.
`ItemCurves`	Show item information curves. Default is FALSE.
`initial`	Method used to choose the initial ability in the EM algorithm. Default value is 1.
`alfa`	Optional parameter to calculate row and column coordinates in Simple correspondence analysis if the initial parameter is equal to 1. Default value is 1.

Details

The general algorithm used is essentially an alternated procedure in which parameters for rows and columns are computed in alternated steps repeated until convergence. Parameters for the rows are calculated by expectation (E-step) and parameters for the columns are computed by maximization (M-step), i. e., by Ordinal Logistic Regression.

There are several options for the computation:

1.- Using the package **mirt** to obtain the row scores, i. e. using a solution obtained from a latent trait model. The column (item) parameters should be directly used by our biplot procedure but, because of the characteristics of the package that performs a default rotation after parameter estimation, we have to reestimate the item parametes to be coherent to the scores.

2.- Using our implementation of the EM algorithm alternating expected a porteriori scores and Ridge Ordinal Logistic Regression for each variable. We use here a Cumulative link model ,that is, a logistic regression model for cumulative logits.

Equations defining the set of probability response surfaces for the cumulative probabilities are sigmoidal as in the binary case (Vicente-Villardon et al.2006) and then they share its geometry. All categories have a different constant but the same slopes, that means that the prediction direction is common to all categories and just the prediction markers are different. The representation subspace can be divided into prediction regions, for each category, delimited by parallel straight lines.

Value

An object of class `"ordinal.logistic.biplot"`. This has some components:

`dataSet`	Data set of study with all the information about the name of the levels and names of the variables and individuals
`RowCoords`	Coordinates for the rows in the reduced space
`NCats`	Number of categories of each variable from the data set
`estimObject`	Object with all the estimated information using EM alternated algorithm or MIRT procedure
`Fitting`	matrix with the percentage of correct clasifications and pseudo R squared valued for each variable
`coefs`	matrix with the estimated coefficients
`thresholds`	matrix with the estimated intercept limits
`NumFactors`	Number of dimensions selected for the study
`Coordinates`	Type of coordinates to calculate the row positions
`Rotation`	Type of rotation if we have chosen mirt coordinates
`Methodfscores`	Method of calculation of the fscores in mirt process
`NumNodos`	Number of nodes for the gauss quadrature in EM algorithm
`tol`	Cut point to stop the EM-algorithm
`maxiter`	Maximum number of iterations in the EM-algorithm
`penalization`	Value for the correction of the ridge regression
`cte`	Boolean value to choose if the model for each variable will have independent term
`show`	Boolean value to indicate if we want to see the results of our analysis
`ItemCurves`	Boolean value to specify if item information curves will be plotted
`LogLik`	Logarithm of the likelihood
`FactorLoadingsComm`	Factor loadings and communalities

Author(s)

Julio Cesar Hernandez Sanchez, Jose Luis Vicente-Villardon

Maintainer: Julio Cesar Hernandez Sanchez <juliocesar_avila@usal.es>

References

Vicente-Villardon, J., Galindo, M.P & Blazquez-Zaballos, A. (2006), *Logistic biplots*,Multiple Correspondence Analysis and related methods pp. 491–509.

Demey, J., Vicente-Villardon, J. L., Galindo, M.P. & Zambrano, A. (2008) *Identifying Molecular Markers Associated With Classification Of Genotypes Using External Logistic Biplots*. Bioinformatics, 24(24), 2832-2838.

Baker, F.B. (1992): Item Response Theory. Parameter Estimation Techniques. Marcel Dekker. New York.

Gabriel, K. (1971), *The biplot graphic display of matrices with application to principal component analysis.*, Biometrika 58(3), 453–467.

Gabriel, K. R. (1998), *Generalised bilinear regression*, Biometrika 85(3), 689–700.

Gabriel, K. R. & Zamir, S. (1979), *Lower rank approximation of matrices by least squares with any choice of weights*, Technometrics 21(4), 489–498.

Gower, J. & Hand, D. (1996), *Biplots, Monographs on statistics and applied probability*. 54. London: Chapman and Hall., 277 pp.

Chalmers,R,P (2012). *mirt: A Multidimensional Item Response Theory Package for the R Environment*. Journal of Statistical Software, 48(6), 1-29. URL http://www.jstatsoft.org/v48/i06/.

See Also

`OrdinalLogBiplotEM`

Examples

```
data(LevelSatPhd)
olbo = OrdinalLogisticBiplot(LevelSatPhd)
summary(olbo)
```

`plot.ordinal.logistic.biplot`
Graphical representation of an Ordinal Logistic Biplot.

Description

This function plots an Ordinal Logistic Biplot. There are parameters related to the way in which the biplot is plotted. All the posible parameters have default values.

Usage

```
## S3 method for class ordinal.logistic.biplot
plot(x, planex = 1, planey = 2,
AtLeastR2 = 0.01, xlimi = -1.5, xlimu = 1.5, ylimi = -1.5,
ylimu = 1.5, margin = 0, ShowAxis = TRUE, PlotVars = TRUE,
PlotInd = TRUE, LabelVar = TRUE, LabelInd = TRUE, CexInd = NULL,
CexVar = NULL, ColorInd = NULL, ColorVar = NULL, PchInd = NULL,
PchVar = NULL, showIIC = FALSE, iicxi = -1.5, iicxu = 1.5,
legendPlot = FALSE,PlotClus = FALSE,Clusters=NULL,
chulls = TRUE,centers = TRUE,colorCluster = NULL,
ConfidentLevel=NULL,addToExistingPlot=FALSE,...)
```

Arguments

x	An object of the class ordinal.logistic.biplot.
planex	Dimension for X axis.
planey	Dimension for Y axis.
AtLeastR2	It establishes the cutting value to plot a variable attending to its Nagelkerke pseudo R squared value. A variable is plotted if its pseudo R squared is higher than this value.
xlimi	Minimum value on the x-axis.
xlimu	Maximum value on the x-axis.
ylimi	Minimum value on the y-axis.
ylimu	Maximum value on the y-axis.
margin	This value stablishes the space between the plotted items and the border of the window.
ShowAxis	Should the axis be shown?
PlotVars	Should the variables (items) be plotted?
PlotInd	Should the individuals be plotted?
LabelVar	Should the variable labels be shown?
LabelInd	Should the individual labels be shown?
CexInd	Size of the individual points. It can be an array with the cex information for each row.
CexVar	Size of the category points. It can be an array with the cex information for each variable.
ColorInd	Color of the individual points. It can be an array with the color information for each row.
ColorVar	Color for the variables. It can be an array with the color information for each variable.
PchInd	Symbol for the individuals. It can be an array with the pch information for each row.
PchVar	Symbol for the variables. It could be an array with the pch information for each variable.
showIIC	Boolean parameter to decide if the user wants to see the item information curves for each variable. Default value is FALSE.
iicxi	Lower limit for the X-axis when plotting item information curves.

`iicxu`	Upper limit for the X-axis when plotting item information curves.
`legendPlot`	Boolean parameter to show the legend of the plot. Default value is FALSE.
`PlotClus`	Boolean parameter to show the clusters studied. Default value is FALSE.
`Clusters`	Variable with the cluster asociated for each item. Default value is NULL.
`chulls`	Boolean parameter to specify if convex hulls figures will be plotted . Default value is FALSE.
`centers`	Boolean parameter to plot the centers of each cluster. Default value is NULL.
`colorCluster`	Color for every cluster. It can be an array with the color information for each cluster. Default value is NULL.
`ConfidentLevel`	Value between 0 and 1 to avoid extreme values for the plot. Default value is NULL.
`addToExistingPlot`	Boolean parameter to decide if the plotted items will be added to an existing plot or not. Default value is FALSE.
`...`	Additional parameters to plot.

Details

The function without parameters plots the ordinal.logistic.biplot object with labels in the original data and default values for colors, symbols and sizes for points and lines. Other values of colors, symbols and sizes can be supplied. A single value applies to all the points but an array with different values can be used to improve the undestanding of the plot.-

Author(s)

Julio Cesar Hernandez Sanchez, Jose Luis Vicente-Villardon

Maintainer: Julio Cesar Hernandez Sanchez <juliocesar_avila@usal.es>

See Also

`OrdinalLogisticBiplot`

Examples

```
data(LevelSatPhd)
olbo = OrdinalLogisticBiplot(LevelSatPhd,penalization=0.2)
plot(olbo,PlotInd=TRUE,xlimi=-1.5,xlimu=1.5,ylimi=-1.5,ylimu=1.5,
  margin = 0.2, ColorVar = c("red","green","black","blue","yellow"),
  CexVar = c(0.7),showIIC=FALSE)
```

`PlotClusters` *Graphical representation of clusters of individuals.*

Description

This function uses a nominal variable to represent groups or clusters of individuals. The clusters can be the result of a clustering algorithm or the groups defined by a external nominal variable. The centroids and convex hulls for each cluster can be represented.

Usage

```
PlotClusters(A, Groups = ones(c(nrow(A), 1)),
        colors = NULL, chulls = TRUE, centers = TRUE, ConfidentLevel = 0.95)
```

Arguments

A	A matrix with the coordinates of each point. It should have only two columns.
Groups	Clustering variable: the cluster for each observation.
colors	It is a vector used to specify the color for each cluster.
chulls	Should convex hulls regions for each cluster be plotted?
centers	Should centroids of each cluster be plotted?
ConfidentLevel	Numerical value between 0 and 1. If it's value is 0.95, five percent of the points with higher distances to the center of each cluster will not be used to calculate centroids and convex hulls.

Author(s)

Julio Cesar Hernandez Sanchez, Jose Luis Vicente-Villardon

Maintainer: Julio Cesar Hernandez Sanchez <juliocesar_avila@usal.es>

Examples

```
data(LevelSatPhd)
olbo = OrdinalLogisticBiplot(LevelSatPhd)
x = olbo$RowCoords[, 1]
y = olbo$RowCoords[, 2]
plot(x,y, cex = 0.8, pch=17, xlim=c(-2,2),ylim=c(-2,2))
GroupsF = as.factor(LevelSatPhd[,4])
PlotClusters(olbo$RowCoords, Groups = GroupsF,
    colors = c(1,2,3,4),chulls = TRUE,centers = TRUE,ConfidentLevel=NULL)
```

`plotOrdinalFittedVariable`

Function that gives the posibility for the user for plotting in the reduced space an ordered and fitted categorical variable.

Description

Graphical representation of a polytomous ordered variable previously fitted in the reduced space, according to the Ordinal Logistic Biplot theory. It can be choosen some parameters related to the way in which the variable is plotted.

Usage

```
plotOrdinalFittedVariable(nameVariable, coeffic, D,numFactors, planex = 1, planey = 2,
xi = -3.5, xu = 3.5, yi = -3.5, yu = 3.5, margin = 0,
CexVar = 0.7, ColorVar = "blue",
PchVar = 0.7, addToPlot = FALSE, showIIC = TRUE,
iicxi = -2.5, iicxu = 2.5)
```

Arguments

`nameVariable`	Name of the variable the user wants to plot.
`coeffic`	Vector with the estimated coefficients and the thresholds in this order.
`D`	Parameter of the graded response model. In case of coefficients have been estimated by Mirt this parameter should be 1.702. In other cases it should be 1.
`numFactors`	Number of dimensions of the solution
`planex`	Dimension for X axis.
`planey`	Dimension for Y axis.
`xi`	Minimum value on the x-axis.
`xu`	Maximum value on the x-axis.
`yi`	Minimum value on the y-axis.
`yu`	Maximum value on the y-axis.
`margin`	This value stablishes the space between the plotted items and the border of the window.
`CexVar`	Size of the category points. It can be an array with the cex information for each variable.
`ColorVar`	Color for the variables. It can be an array with the color information for each variable.
`PchVar`	Symbol for the variables. It could be an array with the pch information for each variable.
`addToPlot`	Boolean parameter to decide if the user wants to add the ordinal variable representation to an existing plot.
`showIIC`	Boolean parameter to decide if the user wants to see the item information curves for each variable. Default value is FALSE.
`iicxi`	Lower limit for the X-axis when plotting item information curves.
`iicxu`	Upper limit for the X-axis when plotting item information curves.

Author(s)

Julio Cesar Hernandez Sanchez, Jose Luis Vicente-Villardon

Maintainer: Julio Cesar Hernandez Sanchez <juliocesar_avila@usal.es>

Examples

```
data(LevelSatPhd)
olbo = OrdinalLogisticBiplot(LevelSatPhd,sFormula=NULL,
      numFactors=2,method="EM",penalization=0.2)
nameVariable="Salary"
coeffic = c(olbo$coefs[1,],olbo$thresholds[1,])
plotOrdinalFittedVariable(nameVariable,coeffic,D=1,numFactors = 2)
```

`plotOrdinalVariable` *This function plots in the reduced space an ordered categorical variable.*

Description

Graphical representation of a polytomous ordered variable in the reduced space, according to Vicente-Villardon & Hernandez-Sanchez(2014) methodology. It can be choosen some parameters related to the way in which the variable is plotted.

Usage

```
plotOrdinalVariable(ordinalfVar,nameVariable,estimRows,planex = 1,planey = 2,
  xi=-3.5,xu=3.5,yi=-3.5,yu=3.5,margin=0,CexVar=0.7,ColorVar="blue",
  PchVar=0.7,addToPlot=FALSE,showIIC = TRUE,iicxi=-2.5,iicxu=2.5,
  tol = 1e-04, maxiter = 100, penalization = 0.1)
```

Arguments

`ordinalfVar`	The ordinal variable. It must be an ordered factor.
`nameVariable`	Name of the variable that the user wants to represent.
`estimRows`	Matrix with the estimated coordinates for the individuals in the reduced dimension.
`planex`	Dimension for X axis.
`planey`	Dimension for Y axis.
`xi`	Minimum value on the x-axis.
`xu`	Maximum value on the x-axis.
`yi`	Minimum value on the y-axis.
`yu`	Maximum value on the y-axis.
`margin`	This value stablishes the space between the plotted items and the border of the window.
`CexVar`	Size of the category points. It can be an array with the cex information for each variable.
`ColorVar`	Color for the variables. It can be an array with the color information for each variable.
`PchVar`	Symbol for the variables. It could be an array with the pch information for each variable.
`addToPlot`	Boolean parameter to decide if the user wants to add the ordinal variable representation to an existing plot.
`showIIC`	Boolean parameter to decide if the user wants to see the item information curves for each variable. Default value is FALSE.
`iicxi`	Lower limit for the X-axis when plotting item information curves.
`iicxu`	Upper limit for the X-axis when plotting item information curves.
`tol`	Tolerance for the iterations.
`maxiter`	Maximum number of iterations.
`penalization`	Penalization used to avoid singularities.

Author(s)

Julio Cesar Hernandez Sanchez, Jose Luis Vicente-Villardon

Maintainer: Julio Cesar Hernandez Sanchez <juliocesar_avila@usal.es>

References

Vicente-Villardon, J. L. & Hernandez, J. C. & (2014) Logistic Biplots for ordinal data with an application to job satisfaction of doctorate degree holders in Spain. Preprint available at arXiv.

Examples

```
data(LevelSatPhd)
olbo = OrdinalLogisticBiplot(LevelSatPhd,sFormula=NULL,
  numFactors=2,method="EM")
ordinalfVar = factor(LevelSatPhd[,1],ordered=TRUE)
levels(ordinalfVar) = c("VS","SS","SD","VD")
estimRows = olbo$RowCoords
nameVariable = "Salary"
plotOrdinalVariable(ordinalfVar,nameVariable,estimRows,planex = 1,
  planey = 2,xi=-1.5,xu=1.5,yi=-1.5,yu=1.5,
  margin=0.2,CexVar=0.7,showIIC = TRUE)
```

`pordlogist` *Ordinal logistic regression with ridge penalization*

Description

This function performs a logistic regression between a dependent ordinal variable y and some independent variables x, and solves the separation problem using ridge penalization.

Usage

```
pordlogist(y, x, penalization = 0.1, tol = 1e-04, maxiter = 200, show = FALSE)
```

Arguments

`y`	Dependent variable.
`x`	A matrix with the independent variables.
`penalization`	Penalization used to avoid singularities.
`tol`	Tolerance for the iterations.
`maxiter`	Maximum number of iterations.
`show`	Should the iteration history be printed?.

Details

The problem of the existence of the estimators in logistic regression can be seen in Albert (1984); a solution for the binary case, based on the Firth's method, Firth (1993) is proposed by Heinze(2002). All the procedures were initially developed to remove the bias but work well to avoid the problem of separation. Here we have chosen a simpler solution based on ridge estimators for logistic regression Cessie(1992).

Rather than maximizing $L_j(\mathbf{G}|\,\mathbf{b}_{j0},\mathbf{B}_j)$ we maximize

$$L_j(\mathbf{G}|\,\mathbf{b}_{j0},\mathbf{B}_j) - \lambda\left(\|\mathbf{b}_{j0}\|^2 + \|\mathbf{B}_j\|^2\right)$$

Changing the values of λ we obtain slightly different solutions not affected by the separation problem.

Value

An object of class `"pordlogist"`. This has components:

`nobs`	Number of observations
`J`	Maximum value of the dependent variable
`nvar`	Number of independent variables
`fitted.values`	Matrix with the fitted probabilities
`pred`	Predicted values for each item
`Covariances`	Covariances matrix
`clasif`	Matrix of classification of the items
`PercentClasif`	Percent of good classifications
`coefficients`	Estimated coefficients for the ordinal logistic regression
`thresholds`	Thresholds of the estimated model
`logLik`	Logarithm of the likelihood
`penalization`	Penalization used to avoid singularities
`Deviance`	Deviance of the model
`DevianceNull`	Deviance of the null model
`Dif`	Diference between the two deviances values calculated
`df`	Degrees of freedom
`pval`	p-value of the contrast
`CoxSnell`	Cox-Snell pseudo R squared
`Nagelkerke`	Nagelkerke pseudo R squared
`MacFaden`	Nagelkerke pseudo R squared
`iter`	Number of iterations made

Author(s)

Jose Luis Vicente-Villardon, Julio Cesar Hernandez Sanchez

Maintainer: Julio Cesar Hernandez Sanchez <juliocesar_avila@usal.es>

References

Albert,A. & Anderson,J.A. (1984),*On the existence of maximum likelihood estimates in logistic regression models*, Biometrika 71(1), 1–10.

Bull, S.B., Mak, C. & Greenwood, C.M. (2002), *A modified score function for multinomial logistic regression*, Computational Statistics and dada Analysis 39, 57–74.

Firth, D.(1993), *Bias reduction of maximum likelihood estimates*, Biometrika 80(1), 27–38

Heinze, G. & Schemper, M. (2002), *A solution to the problem of separation in logistic regression*, Statistics in Medicine 21, 2109–2419

Le Cessie, S. & Van Houwelingen, J. (1992), *Ridge estimators in logistic regression*, Applied Statistics 41(1), 191–201.

See Also

`OrdinalLogBiplotEM`,`CheckDataSet`

Examples

```
data(LevelSatPhd)
dataSet = CheckDataSet(LevelSatPhd)
datanom = dataSet$datanom
olb = OrdinalLogBiplotEM(datanom,dim = 2, nnodos = 10,
            tol = 0.001, maxiter = 100, penalization = 0.2)
model = pordlogist(datanom[, 1], olb$RowCoordinates, tol = 0.001,
        maxiter = 100, penalization = 0.2)
model
```

`summary.ordinal.logistic.biplot`
Summary Method Function for Objects of Class 'ordinal.logistic.biplot'

Description

This function shows a summary of the principal results for the estimation for individuals and variables, like some Pseudo R-squared indices, the percent of correct classifications for each regression, the logLikelihood and "Estimate coefficients", "Std. Error", "z value" or "Pr(>|z|)" values.

Usage

```
## S3 method for class ordinal.logistic.biplot
summary(object,data = FALSE,rowCoords = FALSE,
coefs = FALSE,loadCommun = FALSE,...)
```

Arguments

object	This parameter keeps the ordinal logistic biplot object
data	Boolean parameter to show the number of observations. Default value is FALSE.
rowCoords	Boolean parameter to show the coordinates of the individuals. Default value is FALSE.
coefs	Boolean parameter to show the coefficients of the object. Default value is FALSE.
loadCommun	Boolean parameter to show the factor loadings and communalities. Default value is FALSE.
...	Additional parameters to summary.

Details

This function is a method for the generic function summary() for class "ordinal.logistic.biplot". It can be invoked by calling summary(x) for an object x of the appropriate class.

Author(s)

Julio Cesar Hernandez Sanchez, Jose Luis Vicente-Villardon

Maintainer: Julio Cesar Hernandez Sanchez <juliocesar_avila@usal.es>

See Also

OrdinalLogisticBiplot

Examples

```
data(LevelSatPhd)
olbo = OrdinalLogisticBiplot(LevelSatPhd,sFormula=NULL,numFactors=2,
method="EM",penalization=0.2,show=FALSE)
summary(olbo)
```

`summary.pordlogist` *Summary Method Function for Objects of Class 'pordlogist'*

Description

This function shows a summary of the principal results for the estimation for individuals and variables, like number of observations, the number of iterations, the covariances matrix, some Pseudo R-squared indices with the correct classification percentage of each regression and the logLikelihood with "Estimate coefficients", "Std. Error", "z value" or "Pr(>|z|)" values.

Usage

```
## S3 method for class pordlogist
summary(object,...)
```

Arguments

object	This parameter keeps 'pordlogist' object for a variable.
...	Additional parameters to summary.

Details

This function is a method for the generic function summary() for class "pordlogist". It can be invoked by calling summary(x) for an object x of the appropriate class.

Author(s)

Julio Cesar Hernandez Sanchez, Jose Luis Vicente-Villardon

Maintainer: Julio Cesar Hernandez Sanchez <juliocesar_avila@usal.es>

See Also

pordlogist,CheckDataSet,OrdinalLogBiplotEM

Examples

```
data(LevelSatPhd)
dataSet = CheckDataSet(LevelSatPhd)
datanom = dataSet$datanom
olb = OrdinalLogBiplotEM(datanom,dim = 2, nnodos = 10, tol = 0.001,
  maxiter = 100, penalization = 0.2)
model = pordlogist(datanom[, 1], olb$RowCoordinates, tol = 0.001,
  maxiter = 100, penalization = 0.2)
summary(model)
```

Index

Apéndice I

Manual del paquete de R BiplotForCategoricalVariables

Package 'BiplotForCategoricalVariables'

October 20, 2014

Type Package

Title Biplot representations of categorical data

Version 0.1

Date 2014-10-20

Author Julio Cesar Hernandez Sanchez, Jose Luis Vicente-Villardon

Maintainer Julio Cesar Hernandez Sanchez <juliocesar_avila@usal.es>

Description Analysis of a matrix of polytomous ordered or unordered items. The individuals are represented as points on a plane and this type of variables are represented as convex prediction regions rather than vectors as in a classical or binary biplot.
Using the methods from Computational Geometry, the set of prediction regions is converted to a set of points in such a way that the prediction for each individual is established by its closest
``category point". Then interpretation is based on distances rather than on projections.
The variables with ordered values are represented as lines rather than vectors as in a classical or binary biplot, specifying the points for each of the categories of the variable. The set of prediction regions is established by stripes perpendicular to the line between the category points, in such a way that the prediction for each individual is given by its projection into the line of the variable.
In this package we use the implementation of the geometry of such representations given by NominalLogisticBiplot and OrdinalLogisticBiplot packages and we adapt the computational algorithms for the treatment of any data set with categorical variables.

License GPL (>=2)

Encoding latin1

Repository CRAN

Depends R (>= 2.15.1),mirt,gmodels,MASS,NominalLogisticBiplot,OrdinalLogisticBiplot

LazyData yes

Archs i386, x64

NeedsCompilation no

R topics documented:

`BiplotForCategoricalVariables-package`
Biplot representations for categorical data.

Description

Analysis of a matrix of polytomous ordered or unordered items. The individuals are represented as points on a plane and this type of variables are represented as convex prediction regions rather than vectors as in a classical or binary biplot. Using the methods from Computational Geometry, the set of prediction regions is converted to a set of points in such a way that the prediction for each individual is established by its closest "category point". Then interpretation is based on distances rather than on projections. The variables with ordered values are represented as lines rather than vectors as in a classical or binary biplot, specifying the points for each of the categories of the variable. The set of prediction regions is established by stripes perpendicular to the line between the category points, in such a way that the prediction for each individual is given by its projection into the line of the variable. In this package we use the implementation of the geometry of such representations given by NominalLogisticBiplot and OrdinalLogisticBiplot packages and we adapt the computational algorithms for the treatment of any data set with categorical variables.

Details

Package:	BiplotForCategoricalVariables
Type:	Package
Version:	1.0
Date:	2014-10-17
License:	GPL (>=2)

Author(s)

Julio Cesar Hernandez Sanchez, Jose Luis Vicente-Villardon Maintainer: Julio Cesar Hernandez Sanchez <juliocesar_avila@usal.es>

See Also

`NominalLogisticBiplot package`, `OrdinalLogisticBiplot package`, `BiplotForCategoricalVariables`, `summary.mixed.logistic.biplot.EM`,

Examples

```
library("mirt")
library("gmodels")
library(MASS)
library(NominalLogisticBiplot)
library(OrdinalLogisticBiplot)
#We load the data set of categorical variables
data(PhD_categCyL)
datanomordPhD_mixedEM = PhD_categCyL[1:100,]

#The object from class mixed.logistic.biplot.EM
#is created with the alternated algorithm
xEM = BiplotForCategoricalVariables(datanomordPhD_mixedEM,
    itemtype=c(nominal,nominal,nominal, nominal,
    nominal,ordinal,ordinal,ordinal,ordinal,
    ordinal,ordinal,ordinal,ordinal,ordinal,
    ordinal,ordinal,nominal,nominal), dim = 2,
    nnodos = 10, tol = 0.0001,maxiter = 500,
    penalization = 0.3,initial=1,showResults=TRUE)

#We plot the representation using public functions for
#plotting nominal and ordinal variables from
#NominalLogisticBiplot and OrdinalLogisticBiplot packages
ColorVar = c("red","black","green","blue","orange","burlywood4",
        "chartreuse","chocolate3","chocolate4","azure4","black",
        "darkblue","darkolivegreen","darkmagenta","deeppink",
        "gray54","peru","salmon3")
plot(xEM,xlimi=-2,xlimu=2,ylimi=-2,ylimu=2,sepVarDifWindow=TRUE,
       PlotVars=TRUE,PlotInd=TRUE,LabelInd = TRUE,
       linesVoronoi = TRUE,ColorVar = ColorVar)
```

`BiplotForCategoricalVariables`

Alternated EM algorithm for Categorical Logistic Biplots

Description

This function computes, with an alternated algorithm, the row and column parameters of a Categorical Logistic Biplot for polytomous data. The row coordinates (E-step) are computed using multidimensional Gauss-Hermite quadratures and Expected *a posteriori* (EAP) scores and parameters for each variable or items (M-step)using Ridge Logistic Regression(ordinal or nominal regressions for ordered and unordered variables respectively) to solve the separation problem present when the points for different categories of a variable are completely separated on the representation plane and the usual fitting methods do not converge.

Usage

```
BiplotForCategoricalVariables(x,itemtype=NULL, dim = 2, nnodos = 10,
    tol = 1e-04, maxiter = 100, penalization = 0.2,initial=1,
    alfa=1,Plot=FALSE,showResults=FALSE)
```

Arguments

x	Matrix or Data Frame with the categorical variables.
itemtype	Vector that specifies the type of each variable. Posibles values are 'nominal' and 'ordinal'.
dim	Dimension of the solution
nnodos	Number of nodes for the multidimensional Gauss-Hermite quadrature
tol	Value to stop the process of iterations.
maxiter	Maximum number of iterations in the process of solving the regression coefficients.
penalization	Penalization used in the diagonal matrix to avoid singularities.
initial	Value to decide the method(1-Correspondence analysis, 2-Mirt) that calculates the initial abilities values for the individuals.
alfa	If initial parameter method is correspondence analysis, this parameter determines the weight for rows and columns.
Plot	Boolean parameter to plot the row coordinates.
showResults	Boolean parameter to show all the information about the iterations.

Value

An object of class "mixed.logistic.biplot.EM".This has components:

datanom	Matrix of integers with the values of the variables
RowNames	Vector with the names of the rows
ColumNames	Vector with the names of the variables
LevelNames	Levels of each variable
RowCoordinates	Coordinates for the individuals in the reduced space
dimensions	Number of dimensions retained
nnodos	Number of nodes for the multidimensional Gauss-Hermite quadrature
penalization	Penalization used to avoid singularities. It is a real number
itemtype	Vector that specifies the type of each variable. Posibles values for them are 'nominal' and 'ordinal'
ColumnModelsNominal	List with information about the Nominal Logistic Models calculated for each nominal variable including: estimated parameters with covariances and standard errors, log-likelihood, deviances, percents of correct classifications, pvalues and pseudo-Rsquared measures
ColumnModelsOrdinal	List with 3 matrices that keep the coefficients estimated, the thresholds and the goodness of fit for each ordinal variable.

Author(s)

Julio Cesar Hernandez Sanchez, Jose Luis Vicente-Villardon

Maintainer: Julio Cesar Hernandez Sanchez <juliocesar_avila@usal.es>

References

Bock,R. & Aitkin,M. (1981),*Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm*, Phychometrika 46(4), 443-459.

Gabriel, K. R. (1998). Generalised bilinear regression. Biometrika, 85(3), 689-700.

Vicente-Villardon, J. L., Galindo Villardon, M. P., & Blazquez Zaballos, A. (2006). Logistic biplots. Multiple correspondence analysis and related methods. London: Chapman & Hall, 503-521.

Gabriel, K. R., & Zamir, S. (1979). Lower rank approximation of matrices by least squares with any choice of weights. Technometrics, 21(4), 489-498.

Examples

```
library("mirt")
library("gmodels")
library(MASS)
library(NominalLogisticBiplot)
library(OrdinalLogisticBiplot)
data(PhD_categCyL)
datanomordPhD_mixedEM = PhD_categCyL[1:100,]
xEM = BiplotForCategoricalVariables(datanomordPhD_mixedEM,
        itemtype=c(nominal,nominal,nominal, nominal,
        nominal,ordinal,ordinal,ordinal,ordinal,
        ordinal,ordinal,ordinal,ordinal,ordinal,
        ordinal,ordinal,nominal,nominal), dim = 2,
        nnodos = 10, tol = 0.0001, maxiter = 500,
        penalization = 0.3,initial=1,showResults=TRUE)
```

`PhD_categCyL` *Data set from Survey on Human Resources in Science and Technology carried out by Spanish Statistical Office.*

Description

The sample data corresponds 681 answers from PhD holders, corresponding to people that in 2006 had a doctoral degree and with their residence in Castilla-Le\'on region in Spain. The data come from Survey on Careers of doctorate holders(CDH) carried out by Spanish Statistical Office in 2008.

Usage

```
data(PhD_categCyL)
```

Format

This data frame contains 18 variables. Some of them are nominal variables without any orden in their categories and the others are ordinal, rating the satisfaction with the principal job of the PhD holders.

MS Marital Status:(1:M(Married),2:MLR(Living in a marriage-like relationship), 3:SD (Separated or Divorced),4:SW(Widowed or Single)

SECT Sector of employment(1:BES(Business Enterprise Sector), 2:GS (Government Sector), 3: HES(Higher Education Sector), 4:PNP(Private Non Profit))

MIN Minimum education level required for the principal job: (1:mPD(Postdoc),2:mARQ(Advanced Research Qualification), 3:mPG(Post-graduate),4:mGL(Graduate or lower)

DES Desirable education level required for the principal job: (1:dPD(Postdoc),2:dARQ(Advanced Research Qualification), 3:dPG(Post-graduate),4:dGL(Graduate or lower)

PJREL Is your principal job related to your advanced research qualification degree: (1: H(High), 2: M(Medium), 3: L(Low))

NS1 Salary(1= "Very Satisfied", 2= "Somewhat Satisfied",3="Somewhat dissatisfied", 4="Very dissatisfied")

NS2 Benefits(1= "Very Satisfied", 2= "Somewhat Satisfied",3="Somewhat dissatisfied", 4="Very dissatisfied")

NS3 Job Security(1= "Very Satisfied", 2= "Somewhat Satisfied",3="Somewhat dissatisfied", 4="Very dissatisfied")

NS4 Job Location(1= "Very Satisfied", 2= "Somewhat Satisfied",3="Somewhat dissatisfied", 4="Very dissatisfied")

NS5 Working Conditions(1= "Very Satisfied", 2= "Somewhat Satisfied", 3="Somewhat dissatisfied", 4="Very dissatisfied")

NS6 Opportunities for advancement(1= "Very Satisfied", 2= "Somewhat Satisfied",3="Somewhat dissatisfied", 4="Very dissatisfied")

NS7 Intellectual challenge(1= "Very Satisfied", 2= "Somewhat Satisfied",3="Somewhat dissatisfied", 4="Very dissatisfied")

NS8 Level of responsability(1= "Very Satisfied", 2= "Somewhat Satisfied",3="Somewhat dissatisfied", 4="Very dissatisfied")

NS9 Degree of independence(1= "Very Satisfied", 2= "Somewhat Satisfied",3="Somewhat dissatisfied", 4="Very dissatisfied")

NS10 Contribution to society(1= "Very Satisfied", 2= "Somewhat Satisfied",3="Somewhat dissatisfied", 4="Very dissatisfied")

NS11 Social status(1= "Very Satisfied", 2= "Somewhat Satisfied",3="Somewhat dissatisfied", 4="Very dissatisfied")

FOSAT Field of science and technology (1: NS(Natural Sciences), 2: ET(Engineering and technology), 3: MH(Medical and health sciences), 4: AS(Agricultural sciences), 5: SS(Social Sciences), 6: H(Humanities))

SOF Principal source of financial support during your research studies (1: F(Fellowship), 2: T(Teaching), 3: OE(Other Employment), 4: R(Reimbursement), 5: LPSO(LoanPersonalSavingsOther)

Source

http://www.ine.es/prodyser/micro_recurciencia.htm , Spanish Statistical Office (Survey on Human Resources in Science and Technology, 2006)

Examples

```
data(PhD_categCyL)
```

`plot.mixed.logistic.biplot.EM`
Graphical representation of a Categorical Logistic Biplot.

Description

This function plots a biplot with nominal and ordinal variables. There are parameters related to the way in which the biplot is plotted. All the posible parameters have default values. This function uses the public functions 'plotNominalVariable' and 'plotOrdinalVariable' from 'NominalLogisticBiplot' and 'OrdinalLogisticBiplot' packages respectively.

Usage

```
## S3 method for class mixed.logistic.biplot.EM
plot(x,planex=1,planey=2,QuitNotPredicted=TRUE,
       sepVarDifWindow=TRUE,xlimi=-1.5,xlimu=1.5,ylimi=-1.5,ylimu=1.5,margin=0.1,
       linesVoronoi = TRUE, ShowAxis = TRUE, PlotVars = TRUE, PlotInd = TRUE,
       LabelVar = TRUE, LabelInd = TRUE,CexInd = NULL, CexVar = NULL,
       ColorInd = NULL, ColorVar = NULL, SmartLabels = FALSE, PchInd = NULL,
       PchVar = NULL,ShowResults=TRUE, showIIC = FALSE,penalOrd=0.1,
       penalNom=0.1,tol = 1e-04, maxiter = 100, iicxi=-1,iicxu=1,
       addToPlot=TRUE,legendBR=FALSE,...)
```

Arguments

`x`	An object of the class mixed.logistic.biplot.EM
`planex`	Dimension for X axis.
`planey`	Dimension for Y axis.
`QuitNotPredicted`	Should the non-predicted categories be represented on the graph?
`sepVarDifWindow`	Should each variable represented on a separate window?
`xlimi`	Minimum value on the x-axis.
`xlimu`	Maximum value on the x-axis.
`ylimi`	Minimum value on the y-axis.
`ylimu`	Maximum value on the y-axis.
`margin`	This value stablishes the space between the plotted items and the border of the window.
`linesVoronoi`	Should the tesselation be plotted.? Default is FALSE and only the category points are plotted for a better reading of the plot.
`ShowAxis`	Should the axis be shown?
`PlotVars`	Should the variables (items) be pplotted?
`PlotInd`	Should the individuals be plotted?
`LabelVar`	Should the variable labels be shown?
`LabelInd`	Should the individual labels be shown?

CexInd	Size of the individual points. It can be an array with the cex information for each row.
CexVar	Size of the category points. It can be an array with the cex information for each variable.
ColorInd	Color of the individual points. It can be an array with the color information for each row.
ColorVar	Color for the variables. It can be an array with the color information for each variable.
SmartLabels	Should the text labels be printed accordind to its position on the plot?.
PchInd	Symbol for the individuals. It can be an array with the pch information for each row.
PchVar	Symbol for the variables. It could be an array with the pch information for each variable.
ShowResults	Should the results of the proccess of calculating the prediction regions be shown?
showIIC	Boolean parameter to decide if the user wants to see the item information curves for each variable. Default value is FALSE.
penalOrd	Penalization used to avoid singularities in ordinal logistic regression.
penalNom	Penalization used to avoid singularities in multinomial logistic regression.
tol	Tolerance for the iterations.
maxiter	Maximum number of iterations.
iicxi	Lower limit for the X-axis when plotting item information curves.
iicxu	Upper limit for the X-axis when plotting item information curves.
addToPlot	Boolean parameter to decide if the user wants to add the ordinal or nominal variable representation to an existing plot.
legendBR	Boolean parameter to decide if the user wants to see the legend of the representation in case that all the variables are plotted in the same window.
...	Additional parameters to plot.

Details

The function plots the 'mixed.logistic.biplot.EM' object with labels in the original data and default values for colors, symbols and sizes for points and lines. Other values of colors, symbols and sizes can be supplied. A single value applies to all the points but an array with different values can be used to improve the undestanding of the plot. By default, each variable will be presented on a separate window.-

Author(s)

Julio Cesar Hernandez Sanchez, Jose Luis Vicente-Villardon

Maintainer: Julio Cesar Hernandez Sanchez <juliocesar_avila@usal.es>

See Also

plotOrdinalVariable,plotNominalVariable

Examples

```
library("mirt")
library("gmodels")
library(MASS)
library(NominalLogisticBiplot)
library(OrdinalLogisticBiplot)
data(PhD_categCyL)
datanomordPhD_mixedEM = PhD_categCyL[1:100,]
xEM = BiplotForCategoricalVariables(datanomordPhD_mixedEM,
        itemtype=c(nominal,nominal,nominal, nominal,
        nominal,ordinal,ordinal,ordinal,ordinal,
        ordinal,ordinal,ordinal,ordinal,ordinal,
        ordinal,ordinal,nominal,nominal), dim = 2,
        nnodos = 10, tol = 0.0001, maxiter = 500,
        penalization = 0.3,initial=1,showResults=TRUE)
plot(xEM,xlimi=-2,xlimu=2,ylimi=-2,ylimu=2,sepVarDifWindow=TRUE,
       PlotVars=TRUE,PlotInd=TRUE,LabelInd = TRUE,
       linesVoronoi = TRUE)
```

`summary.mixed.logistic.biplot.EM` *Summary Method Function for Objects of Class 'mixed.logistic.biplot.EM'*

Description

This function shows a summary of the principal results for the estimation for individuals and variables, like some Pseudo R-squared indices, the correct classification percentage of each regression, the logLikelihood and "Estimate coefficients", "Std. Error", "z value" or "Pr(>|z|)" values.

Usage

```
## S3 method for class mixed.logistic.biplot.EM
summary(object,summFitting, coorInd,
                          nominalsFitting, ordinalsFitting,...)
```

Arguments

`object`	This parameter keeps the categorical(also called mixed because it works with nominal and ordinal variables) logistic biplot object.
`summFitting`	Boolean parameter to choose if the user wants to see the goodness of fit of the variables. Default value is FALSE.
`coorInd`	Boolean parameter to choose if the individual coordinates will be printed on screen.Default value is FALSE.
`nominalsFitting`	Boolean parameter to show the complete information about the estimation for the nominal variables.Default value is FALSE.
`ordinalsFitting`	Boolean parameter to show the complete information about the estimation for the ordinal variables.Default value is FALSE.
`...`	Additional parameters to summary.

Details

This function is a method for the generic function summary() for class "mixed.logistic.biplot.EM". It can be invoked by calling summary(x) with an object x of the appropriate class.

Author(s)

Julio Cesar Hernandez Sanchez, Jose Luis Vicente-Villardon

Maintainer: Julio Cesar Hernandez Sanchez <juliocesar_avila@usal.es>

See Also

BiplotForCategoricalVariables

Examples

```
library("mirt")
library("gmodels")
library(MASS)
library(NominalLogisticBiplot)
library(OrdinalLogisticBiplot)
data(PhD_categCyL)
datanomordPhD_mixedEM = PhD_categCyL[1:100,]
xEM = BiplotForCategoricalVariables(datanomordPhD_mixedEM,
        itemtype=c(nominal,nominal,nominal, nominal,
        nominal,ordinal,ordinal,ordinal,ordinal,
        ordinal,ordinal,ordinal,ordinal,ordinal,
        ordinal,ordinal,nominal,nominal), dim = 2,
        nnodos = 10, tol = 0.0001, maxiter = 500,
        penalization = 0.3,initial=1,showResults=TRUE)
summary(xEM, summFitting = TRUE)
```

Index

Apéndice J

Encuesta sobre Recursos Humanos en Ciencia y Tecnología 2009

Instituto Nacional de Estadística

Encuesta sobre Recursos Humanos en Ciencia y Tecnología 2006

DOCUMENTO PROTEGIDO INE POR EL SECRETO ESTADÍSTICO

Identificación del doctor/a

Modificaciones en la identificación *(Cumplimentar **sólo** los apartados sujetos a variación)*

Nombre del informante _____ NIF _____

Domicilio (calle, plaza, paseo, avenida, ...) _____ Código postal _____

Municipio _____ Cód. Munic. _____ Provincia _____ Cód. provi. _____

Teléfono _____ Fax _____ e-mail _____ Teléfono móvil _____

Naturaleza, características y finalidad

Esta Encuesta se enmarca dentro del Plan general de estadísticas de ciencia y tecnología propugnado por la oficina de Estadísticas de la Unión Europea (Eurostat). El objetivo de la encuesta es cuantificar el nivel de investigación de los doctores en España, la actividad profesional que desarrollan y la movilidad nacional e internacional de los mismos.

Legislación

Secreto Estadístico

Serán objeto de protección y quedarán amparados por el **secreto estadístico** los datos personales que obtengan los servicios estadísticos, tanto directamente de los informantes como a través de fuentes administrativas (Art. 13.1 de la Ley de la Función Estadística Pública de 9 de mayo de 1989, LFEP). Todo el personal estadístico tendrá la obligación de preservar el secreto estadístico (Art. 17.1 de la LFEP).

Obligación de facilitar los datos

Esta encuesta forma parte del Plan Estadístico Nacional y por ello, de acuerdo con la Ley 13/1996 este cuestionario tiene el carácter de obligatorio.

Los servicios estadísticos podrán solicitar datos de todas las personas físicas y jurídicas, nacionales y extranjeras residentes en España (Art. 10.1 de la LFEP). Todas las personas físicas o jurídicas que suministren datos, tanto si su colaboración es obligatoria como voluntaria, deben contestar de forma veraz, exacta, completa y dentro del plazo a las preguntas ordenadas en la debida forma por parte de los servicios estadísticos (Art. 10.2 de la LFEP).

(Ley 12/1989, de la Función Estadística Pública).

Mod. RR.HH.-2006

Apéndice J. Encuesta sobre Recursos Humanos en Ciencia y Tecnología 2009

Instrucciones generales

Unidad de información: la información que se solicita en este cuestionario se refiere a personas físicas que sean doctores residentes en España.

Período de referencia: los datos deben referirse a 31 de diciembre de **2006**, salvo que en la pregunta se solicite información referida a otro periodo.

Estructura del cuestionario: el cuestionario se compone de 6 apartados y 4 anexos:
A. Características personales
B. Doctorado
C. Situación laboral
D. Desempleados e inactivos
E. Movilidad internacional
F. Experiencia profesional y productividad científica
Anexo I. Definiciones básicas y guía para la cumplimentación del cuestionario
Anexo I I. Clasificación 1. Campos de Ciencia y Tecnología
Anexo I I I. Clasificación 2. Ocupaciones ISCO-88
Anexo I V Ejemplo de tabla de Movilidad Internacional.

Forma de anotar los datos: cumplimente los datos claramente. Los datos económicos se solicitan en **euros.** No deben rellenarse las casillas sombreadas.

Plazo de remisión: este cuestionario cumplimentado con la información solicitada, debe ser devuelto en un plazo no superior a **10 días.**

A. Características personales

*Rellenar siempre los apartados A.1, A.2 y A.3 , **se tenga o no** la titulación de doctor*

A.1 Indique los siguientes datos

1. Fecha de nacimiento: día ☐☐ mes ☐☐ año ☐☐☐☐

2. Lugar de nacimiento: Municipio ..... ☐☐☐

Provincia ..... ☐☐ País ..... ☐☐☐

3. Nacionalidad
Española ☐ → Ir a la pregunta 6
Extranjera ☐
Española y otra ☐

4. País de la nacionalidad extrajera ..... ☐☐☐

NOTA: *si tiene más de 1 nacionalidad extranjera, anótelo en el campo de observaciones (última hoja del cuestionario)*

5. Indique la relación que le une con España *(Consulte en el anexo I las definiciones de residencia temporal y permanente)*
Residencia temporal ☐
Residencia permanente ☐
Refugiado ☐

6. Sexo
Varón ☐
Mujer ☐

7. Estado civil
1. Casado ☐ 3. Separado ☐ 5. Viudo ☐
2. Pareja de hecho ☐ 4. Divorciado ☐ 6. Soltero ☐

A.2 Residencia en 2006

1. ¿Residió en España en algún periodo de 2006?

SÍ ☐

NO ☐ → Ir al apartado A.3

2. Indique en qué provincias de España residió durante 2006 y el tiempo que permaneció en ellas:

Provincia 1: ..... ☐☐ Número de meses ☐☐

Provincia 2: ..... ☐☐ Número de meses ☐☐

Provincia 3: ..... ☐☐ Número de meses ☐☐

NOTA: *si ha vivido en más provincias anótelo en el campo de observaciones (última hoja del cuestionario)*

A.3 A 31 de diciembre de 2006, ¿tenía usted la titulación de doctor?

SÍ ☐

NO ☐ → **Fin del cuestionario**

B. Doctorado

El doctorado es un programa de tercer ciclo en el que se dirige un programa de investigación avanzada en algún tema sin desarrollar hasta el momento. Estos programas requieren, normalmente, la presentación de una tesis de una calidad publicable y que represente una nueva aportación al conocimiento. En muchos de los casos forma parte de la preparación para puestos de profesor en las Universidades o como investigador en la Administración Pública, la Industria, ...

B.1 Indique el centro en que obtuvo usted su título de Doctor

Departamento .....

Universidad .....

B.2 Indique el código correspondiente a sus estudios de doctorado *(ver Anexo II)*

Código ____ ☐☐☐

B.3 Indique el carácter de la investigación llevada a cabo durante su doctorado *(responda SI o NO en todas las opciones)*

	SÍ	NO
1. Investigación básica o fundamental	☐	☐
2. Investigación aplicada	☐	☐
3. Desarrollo experimental	☐	☐

Apéndice J. Encuesta sobre Recursos Humanos en Ciencia y Tecnología 2009

B.4 Periodo en el que se realizó el doctorado

1. ¿En qué año obtuvo su título de Doctor? _____ |_|_|_|_|

2. ¿Cuánto tiempo transcurrió desde que inició los cursos de doctorado hasta que obtuvo el titulo de Doctor? _____ Años |_|_| Meses |_|_|

B.5 Indique el centro en el que realizó los estudios universitarios de 2º ciclo previos al doctorado

Universidad .....

País ..... |_|_|_|

B.6 Indique el año en el que finalizó los estudios universitarios de 2º ciclo

Año _____ |_|_|_|_|

B.7 ¿Cuál fue la principal fuente de financiación que utilizó usted durante sus estudios de doctorado?
(marque sólo una opción)

1. Beca de la institución en la que realizó el doctorado _____ ☐
2. Beca de la Administración Pública (central, autonómica, ...) _____ ☐
3. Beca empresarial _____ ☐
4. Beca de una Institución Privada sin Fines de Lucro _____ ☐
5. Beca internacional _____ ☐
6. Trabajó como ayudante de investigación _____ ☐
7. Trabajó como ayudante de profesor _____ ☐
8. Otra ocupación a tiempo completo _____ ☐
9. Otra ocupación a tiempo parcial _____ ☐
10. Subvención reembolsada por el empleador _____ ☐
11. Préstamo _____ ☐
12. Ahorros personales _____ ☐
13. Ayuda familiar _____ ☐
14. Otras formas *(especificar)* ..... ☐

B.8¿Trabajó en algo relacionado con su doctorado una vez finalizado y antes de enero de 2007?

SÍ _____ ☐

NO _____ ☐ → Ir al apartado C

B.9 Indique el tiempo transcurrido desde la finalización de los estudios de doctorado hasta que empezó a trabajar en algo relacionado con estos

Años ⌊_⌋⌊_⌋ Meses ⌊_⌋⌊_⌋

C. Situación Laboral

C.1 Indique cuál era su situación laboral a 31 de diciembre de 2006 *(admite respuesta múltiple)*

1. Trabajador por cuenta propia ☐
2. Trabajador por cuenta ajena ☐
3. Profesor emérito ☐
4. Desempleado o inactivo ☐

*Si ha señalado **únicamente** la opción 4 "desempleado o inactivo" pase al apartado D*

C.2 Indique el número medio de horas que trabajaba a la semana en diciembre de 2006

Horas semanales ⌊_⌋⌊_⌋

C.3 Indique el tramo en que se encuentran sus ingresos brutos anuales de 2006 *(teniendo en cuenta todos los trabajos de 2006)*

1. Menos de 10.000 euros ☐
2. De 10.000 a 20.000 euros ☐
3. De 20.001 a 30.000 euros ☐
4. De 30.001 a 35.000 euros ☐
5. De 35.001 a 40.000 euros ☐
6. De 40.001 a 45.000 euros ☐
7. De 45.001 a 50.000 euros ☐
8. Más de 50.000 euros ☐

C.4 ¿Cuántas personas dependían económicamente de usted, total o parcialmente a 31 de diciembre de 2006? *(no debe incluirse usted mismo; si no tenía personas dependientes en alguno o en todos los grupos de edad, por favor ponga "0" en las opciones de respuesta)*

Menores de 5 años ⌊_⌋⌊_⌋
De 5 a 18 años ⌊_⌋⌊_⌋
Mayores de 18 años ⌊_⌋⌊_⌋

Apéndice J. Encuesta sobre Recursos Humanos en Ciencia y Tecnología 2009

C.5 Actividad laboral principal a 31 de diciembre de 2006

1. Indique el nombre de la organización o empresa en la que trabajaba: .....
En el caso de que sea usted autónomo rellene este apartado con la palabra "autónomo"

.....

Descripción de la actividad principal de la empresa:

..... CNAE-93

2. Indique la fecha en la que comenzó a trabajar en la organización mes ___ año ____

3. Indique la localización de la organización en la que trabajaba

Municipio ..... ____

Provincia ..... ___

País ..... ____

4. Indique el sector al que pertenece la organización en la que trabajaba *(marque sólo una opción)*

Empresas ☐

Administraciones Públicas ☐

Enseñanza Superior ☐

Instituciones privadas sin ánimo de lucro ☐

5. Indique, utilizando los códigos del **anexo III**, aquel que mejor se adapte a la actividad laboral que desarrollaba ___

5.1. Si el código anterior es 231, indique la categoría profesional o puesto de trabajo que ocupaba

Catedrático ☐

Profesor titular ☐

Profesor asociado, emérito, visitante, ayudante y similar ☐

Otro *(especificar)* ..... ☐

6. Indique el tipo de contrato al que estaba sujeto *(marque sólo una opción)*

Indefinido ☐

Temporal ☐

7. Indique cual es la jornada laboral que desempeñaba *(marque sólo una opción)*

A tiempo completo ☐ → Ir al apartado C.6

A tiempo parcial ☐

8. ¿Estaba usted buscando un trabajo a tiempo completo en diciembre de 2006?

SÍ ☐

NO ☐

C.6 Empleo y formación

1. Indique el nivel mínimo requerido para el puesto que tenía usted en su trabajo principal a 31 de diciembre de 2006 *(marque sólo una opción)*

Postdoctorado ☐
Doctor universitario ☐
Licenciado, arquitecto, ingeniero o similar ☐
Diplomado, arquitecto técnico, ingeniero técnico o similar ☐
Ciclos formativos de grado superior (formación profesional específica) ☐
Ciclos formativos de grado medio, título de bachiller y similares ☐

2. Indique el nivel que considera que es el adecuado para el puesto que tenía usted en su trabajo principal a 31 de diciembre de 2006 *(marque sólo una opción)*

Postdoctorado ☐
Doctor universitario ☐
Licenciado, arquitecto, ingeniero o similar ☐
Diplomado, arquitecto técnico, ingeniero técnico o similar ☐
Ciclos formativos de grado superior (formación profesional específica) ☐
Ciclos formativos de grado medio, título de bachiller y similares ☐

3. Indique el grado de relación entre el trabajo principal que desempeñaba a 31 de diciembre de 2006 y sus estudios de doctorado *(marque sólo una opción)*

Alto ☐
Medio ☐
Bajo ☐

4. Indique el nivel de satisfacción con los siguientes factores relacionados con su trabajo a 31 de diciembre de 2006

	Alto	Medio	Bajo	Ninguno
Salario	☐	☐	☐	☐
Beneficios económicos	☐	☐	☐	☐
Estabilidad	☐	☐	☐	☐
Localización	☐	☐	☐	☐
Condiciones laborales	☐	☐	☐	☐
Oportunidades para promocionar	☐	☐	☐	☐
Componente o reto intelectual	☐	☐	☐	☐
Nivel de responsabilidad	☐	☐	☐	☐
Grado de independencia	☐	☐	☐	☐
Contribución a la sociedad	☐	☐	☐	☐
Status social	☐	☐	☐	☐

C.7. Postdoctorado

El Postdoctorado es el trabajo realizado por una persona que ha obtenido el título de doctor en los últimos seis años y que trabaja como investigador no permanente. Este trabajo se remunera en forma de salario, estipendio o beca de investigación financiada en parte o en su totalidad por un organismo patrocinado por el gobierno, la universidad, organismo no gubernamental, empresa o entidad internacional.

1. ¿Era un puesto postdoctoral su principal empleo a 31 de diciembre de 2006?

SÍ ☐

NO ☐ → Ir al apartado E

2. Indique las razones que le llevaron a realizar ese trabajo de postdoctorado

	SÍ	NO
Realizar prácticas en su campo de estudio	☐	☐
Realizar prácticas en un campo diferente al suyo	☐	☐
Trabajar con una persona específica o en un lugar concreto	☐	☐
No pudo elegir otro empleo	☐	☐
Es el trabajo que se suele realizar en su campo de estudio	☐	☐

3. Indique el porcentaje de su tiempo que empleó como postdoctorado en:

Actividades de investigación ____ |_|_|_|

Actividades de docencia ____ |_|_|_|

Otras ____ |_|_|_|

Total ____ |1|0|0| **%**

4. Indique cual era la forma principal de financiación de ese postdoctorado *(marque sólo una opción)*

Empresas ☐

Administraciones públicas ☐

Centros de Enseñanza Superior ☐

Instituciones privadas sin ánimo de lucro ☐

Fundación privada ☐

Otras *(especificar):* ☐

→ Ir al apartado E

D. Desempleados e inactivos

D.1 Durante enero de 2007, ¿hizo alguna gestión para buscar empleo?

SÍ ☐

NO ☐ → Ir al apartado D.3

D.2 Indique cuál o cuáles de las siguientes gestiones realizó para buscar empleo *(admite respuesta múltiple)*

1. Ponerse en contacto con una empresa de trabajo temporal ☐
2. Ponerse en contacto directamente con las empresas que necesitan personal ☐
3. Revisar y responder anuncios de periódicos ☐
4. Buscar en oficinas de empleo ☐
5. Buscar ayuda entre amigos y parientes ☐
6. Buscar terrenos, maquinaria o equipamiento para establecer su propia empresa ☐
7. Hacer entrevistas de trabajo ☐
8. Solicitar permisos, licencias o recursos financieros ☐
9. Otras *(especificar):* ☐

→ Ir al apartado E

D.3 Indique los motivos por los que no buscó empleo *(admite respuesta múltiple)*

1. Enfermedad o discapacidad ☐
2. Cuidado de hijos ☐
3. Cuidado de otras personas dependientes ☐
4. Otras responsabilidades familiares ☐
5. Jubilación ☐
6. Asistencia a instituciones educativas ☐
7. Convicción de que no existe ningún trabajo disponible para usted ☐
8. No necesita o no quiere trabajar ☐
9. Otros *(especificar):* ☐

E. Movilidad internacional

SI SOLAMENTE VIVIÓ EN ESPAÑA EN EL PERIODO 1996-2006 PASE AL APARTADO E.4

E.1 Realice una lista de los países en los que vivió en el periodo 1996-2006 *(incluyendo España)*

COLOQUE LOS PERIODOS EN ORDEN CRONOLÓGICO INVERSO, DE LO MAS RECIENTE A LO MÁS LEJANO. (VER EJEMPLO EN EL ANEXO IV)

País	Cod. país	Periodo de residencia			
		Desde		Hasta	
		Mes	Año	Mes	Año
				1 2	2 0 0 6

NOTA: *Si la tabla resulta insuficiente anótelo en el campo de observaciones (última hoja del cuestionario)*

Apéndice J. Encuesta sobre Recursos Humanos en Ciencia y Tecnología 2009

E.2 Si se fue a vivir fuera de España en el periodo 1996-2006, indique cuáles fueron los motivos que le llevaron a tomar esa decisión *(admite respuesta múltiple)*

1. Finalizar el doctorado ☐
2. Finalizar el postdoctorado o un contrato de trabajo ☐
3. Otros factores relacionados con el empleo: traslado laboral por el mismo organismo o empresa, cambio de empleo, ... ☐
4. Factores académicos: desarrollo o continuidad de la tesis doctoral, creación equipo de investigación, ... ☐
5. Factores personales, económicos o políticos ☐
6. Otros motivos *(indique cuáles)* .....

..... ☐

E.3 Si vino a vivir a España en el periodo 1996-2006, indique cuáles fueron los motivos que le llevaron a tomar esa decisión (admite respuesta múltiple)

1. Finalizar el doctorado ☐
2. Finalizar el postdoctorado o un contrato de trabajo ☐
3. Otros factores relacionados con el empleo: traslado laboral por el mismo organismo o empresa, cambio de empleo, ... ☐
4. Factores académicos: desarrollo o continuidad de la tesis doctoral, creación equipo de investigación, ... ☐
5. Factores personales, económicos o políticos ☐
6. Otros motivos *(indique cuáles)* .....

..... ☐

E.4 En diciembre de 2006, ¿tenía previsto marcharse a vivir fuera de España?

Sí, permanentemente ☐
Sí, temporalmente ☐
No ☐ → Ir al apartado F

E.5 Indique el tiempo que consideraba que iba a transcurrir antes de abandonar España *(marque sólo una opción)*

1. Menos de 6 meses ☐
2. De 6 meses a 1 año ☐
3. De 1 a 2 años ☐
4. De 2 a 3 años ☐
5. De 3 a 4 años ☐
6. De 4 a 5 años ☐
7. De 5 a 10 años ☐
8. Más de 10 años ☐

E.6 Indique las razones que le han llevado a tomar esa decisión (admite respuesta múltiple)

1. Finalizar el doctorado ☐
2. Finalizar el postdoctorado o un contrato de trabajo ☐
3. Otros factores relacionados con el empleo: traslado laboral por el mismo organismo o empresa, cambio de empleo, ... ☐
4. Factores académicos: desarrollo o continuidad de la tesis doctoral, creación equipo de investigación, ... ☐
5. Factores personales, económicos o políticos ☐
6. Otros motivos *(indique cuáles)* ☐

E.7 ¿A qué país tenía usted previsto trasladarse?

País ____ | | |

F. Experiencia profesional y productividad científica

F.1 ¿Durante los años 2005 y 2006 trabajó para alguna institución educativa?

SÍ ☐
NO ☐ → Ir al apartado F.3

F.2 Del total de horas que usted dedicó a su actividad laboral durante los años 2005 y 2006 ¿Cuál fue la proporción de su tiempo que dedicó a la docencia?

Menos del 25% ☐
Del 25 al 50% ☐
Del 50 al 75% ☐
Más del 75% ☐

F.3 ¿Estaba realizando actividades de investigación a 31 de diciembre de 2006?

SÍ ☐ → Ir al apartado F.6
NO ☐

F.4 Indique los motivos por los que no estaba trabajando como investigador a 31 de diciembre de 2006 *(admite respuesta múltiple)*

1. No estaba interesado en investigar ☐
2. Oportunidades laborales muy limitadas en el campo de la investigación ☐
3. Baja remuneración ☐
4. Malas condiciones laborales ☐
5. Falta de reconocimiento público a la investigación ☐
6. Jubilación ☐
7. Otros *(especificar)* ☐

F.5 ¿Ha realizado alguna vez actividades de investigador entre enero de 2004 y diciembre de 2006?

SÍ ☐
NO ☐ → Ir al apartado F.10

F.6 Indique los motivos por los que se dedicaba a la investigación *(admite respuesta múltiple)*

1. Trabajo creativo e innovador ☐
2. Alta remuneración ☐
3. Promoción profesional ☐
4. Seguridad laboral ☐
5. Buenas condiciones laborales ☐
6. Contribución a la sociedad ☐
7. No pudo elegir otro empleo ☐
8. Otros *(especificar)* ☐

F.7 Indique el total de meses en los que ha llevado a cabo una labor investigadora a lo largo de su vida

Número de meses ____ |_|_|_|

F.8 ¿Cuántos libros o monografías (incluyendo colaboraciones) han sido publicados o aceptados para su publicación entre enero de 2004 y diciembre de 2006?

Número de trabajos ____ |_|_|_|

F.9 ¿Cuántos artículos (incluyendo colaboraciones) han sido publicados o aceptados para su publicación entre enero de 2004 y diciembre de 2006?

Número de trabajos ____ |_|_|_|

F.10 ¿Ha realizado algún invento o alguna aplicación para patentar entre enero de 2004 y diciembre de 2006?

SÍ ☐

NO ☐ → Ir al apartado F.13

F.11 Indique el número de patentes que ha registrado como investigador entre enero de 2004 y diciembre de 2006

Número de patentes ____ ⌊__⌋

F.12 Indique el número de sus patentes de productos o procesos que han sido comercializadas o han obtenido ya las licencias necesarias para ello entre enero de 2004 y diciembre de 2006

Número de patentes ____ ⌊__⌋

F.13 ¿Ha constituido una empresa entre enero de 2004 y diciembre de 2006?

SÍ ☐

NO ☐

F.14 ¿Ha dirigido algún Master o tesis doctoral entre enero de 2004 y diciembre de 2006?

SÍ ☐

NO ☐

F.15 ¿Ha cooperado con grupos de investigación extranjeros entre enero de 2004 y diciembre de 2005?

SÍ ☐

NO ☐

F.16 ¿Tiene intención de dedicarse a la investigación en el periodo 2007-2009?

SÍ ☐

NO ☐

Observaciones

Una vez finalizada la ENCUESTA, introduzca el cuestionario en el sobre de Respuesta Gratuita y deposítelo en un buzón de correos

El Instituto Nacional de Estadística le agradece su colaboración

Anexo I

1. Características personales

1.1 Definiciones básicas

➤ Se considera **nacionalidad** el vínculo que cada individuo tiene con su Estado, adquirido por nacimiento o nacionalización posterior, por declaración, opción, matrimonio u otros métodos de acuerdo con la legislación del país.

➤ La **residencia permanente** es la situación por la que un extranjero reside en España de forma indefinida y puede trabajar en igualdad de condiciones que los españoles.

➤ La **residencia temporal** es la situación por la que un extranjero reside en España por un período superior a 90 días e inferior a 5 años.

2. Doctorado

2.1 Definiciones básicas

➤ La **investigación básica** consiste en trabajos experimentales o teóricos que se emprenden fundamentalmente para obtener nuevos conocimientos acerca de los fundamentos de fenómenos y hechos observables, sin pensar en darles ninguna aplicación o utilización determinada.

➤ La **investigación aplicada** consiste también en trabajos originales realizados para adquirir nuevos conocimientos; sin embargo, está dirigida fundamentalmente hacia un objetivo práctico específico.

➤ El **desarrollo experimental** consiste en trabajos sistemáticos basados en los conocimientos existentes, derivados de la investigación y/o la experiencia práctica, dirigidos a la producción de nuevos materiales, productos o dispositivos; al establecimiento de nuevos procesos, sistemas y servicios, o a la mejora sustancial de los ya existentes.

3. Situación laboral

3.1 Definiciones básicas

➤ **Trabajador por cuenta propia** es la persona que durante el periodo de referencia realiza algún trabajo a cambio de algún beneficio o ganancia familiar.

➤ **Trabajador por cuenta ajena** es la persona que durante el periodo de referencia desarrolla alguna actividad laboral a cambio de una retribución o salario en efectivo o en especie.

➤ **Desempleado** es la persona que, encontrándose en edad activa durante el periodo de referencia no tiene trabajo, está disponible para trabajar y busca empleo activamente. Esta búsqueda de empleo puede consistir en hacer entrevistas, enviar currícula, contactar con Empresas de Trabajo Temporal, ... También pueden estar haciendo gestiones para crear su propia empresa.

➤ **Inactivo** es la persona que no forma parte de la población activa. Forman parte de este colectivo las personas dedicadas exclusivamente a los trabajos de su hogar, las que prestan asistencia en instituciones educativas, los jubilados, los enfermos y los discapacitados.

➤ **Contrato temporal** es el contrato que tiene una fecha de finalización.

➤ **Contrato indefinido** es el que no tiene fecha de finalización. Normalmente los trabajadores contratados por tiempo indefinido disfrutan de una mayor protección a nivel legal que los que tienen contratos temporales.

➤**Trabajadores a tiempo parcial** son los que desarrollan su actividad laboral menos de 30 horas a la semana.

➤ **Trabajadores a tiempo completo** son los que trabajan más de 30 horas a la semana.

➤ Los **Ingresos brutos anuales** se determinan sumando los salarios que ha percibido el trabajador (de uno o más trabajos) antes de restar deducciones e impuestos y sin tener en cuenta bonus, horas extras ni otras compensaciones adicionales.

➤ El sector **Empresas** comprende todas las empresas, organismos e instituciones cuya actividad principal consiste en la producción mercantil de bienes o servicios (exceptuando la enseñanza superior) para su venta al público, a un precio que corresponde al de la realidad económica. También comprende las instituciones privadas sin fines de lucro que están al servicio de las empresas.

➤ El sector de las **Administraciones Públicas** engloba todos los departamentos, oficinas y otros organismos que suministran, generalmente a título gratuito, servicios colectivos, excepto enseñanza superior, que no sería fácil ni rentable suministrar de otro modo y que además, administran los asuntos públicos y la política social de la colectividad. (Las empresas públicas se incluyen en el sector empresas). También se incluyen dentro del sector las IPSFL controladas y financiadas principalmente por la administración, con excepción de las administradas por el sector de la enseñanza superior.

➤ El sector de las **Instituciones Privadas sin Fines de Lucro** comprende las IPSFL que están fuera del mercado y al servicio de los hogares (es decir, del público), los particulares y los hogares. Forman parte de estas instituciones siempre que no se incluyan dentro de la Educación Superior. Las fundaciones de I+D dirigidas o controladas por doctores cuya financiación procede del Estado en más de un 50% se incluyen en el sector Administraciones Públicas.

➤ El sector **Enseñanza Superior** está formado por todas las universidades, institutos tecnológicos y otros centros post-secundarios, cualesquiera que sea el origen de sus recursos y su personalidad jurídica. Incluye también todos los institutos de investigación, estaciones experimentales y hospitales directamente controlados, administrados o asociados a centros de enseñanza superior.

4. Experiencia profesional y productividad científica

4.1 Definiciones básicas

➤ **Investigador** es el profesional encargado de la concepción o creación de nuevos conocimientos, procesos, métodos y sistemas, y también de los proyectos respectivos.

Anexo II

Clasificación 1. Campos de Ciencia y Tecnología

1. CIENCIAS NATURALES
101 Matemáticas
102 Informática y tecnologías de la información
103 Ciencias físicas
104 Ciencias químicas
105 Ciencias de la tierra y medio ambiente
106 Biología (excluyendo agricultura y ciencias médicas)
107 Otras ciencias naturales

2. INGENIERÍA Y TECNOLOGÍA
201 Ingeniería civil
202 Ingeniería eléctrica, electrónica y de telecomunicaciones
203 Ingeniería mecánica
204 Ingeniería química
205 Ingeniería de materiales
206 Ingeniería médica
207 Ingeniería medioambiental
208 Biotecnología medioambiental
209 Biotecnología industrial
210 Nanotecnología
211 Otras ingenierías y tecnologías (comida, bebida y otras)

3. CIENCIAS MÉDICAS
301 Medicina básica
302 Medicina clínica
303 Ciencias de la salud
304 Biotecnología médica
305 Otras ciencias médicas (forenses y otras ciencias médicas)

4. CIENCIAS DE LA AGRICULTURA
401 Agricultura, ciencias forestales y piscifactorías
402 Ciencias de los animales y de la leche
403 Veterinaria
404 Biotecnología agrícola
405 Otras ciencias de la agricultura

5. CIENCIAS SOCIALES
501 Psicología
502 Economía y empresas
503 Ciencias de la educación
504 Sociología
505 Derecho
506 Ciencias políticas
507 Geografía económica y social
508 Periodismo y comunicaciones
509 Otras ciencias sociales

6. HUMANIDADES
601 Historia y arqueología
602 Lenguaje y literatura
603 Filosofía, ética y religión
604 Arte (historia del arte, bellas artes y música)
605 Otras humanidades

Anexo III

Clasificación 2. Ocupaciones ISCO-88

Código	Título ISCO 88
100	Miembros del poder ejecutivo y de los cuerpos legislativos y personal directivo de la administración pública y de empresas
200	Profesionales científicos e intelectuales (*)
211	Físicos, químicos y profesionales afines
212	Matemáticos, estadísticos y profesionales afines
213	Profesionales de la informática
214	Arquitectos, ingenieros y profesionales afines
221	Profesionales en ciencias biológicas y otras disciplinas relativas a los seres orgánicos
222	Médicos y profesionales afines (excepto el personal de enfermería y partería)
223	Personal de enfermería y partería de nivel superior
231	Profesores de universidades y otros establecimientos de la enseñanza superior
232	Profesores de la enseñanza secundaria
233	Maestros de nivel superior de la enseñanza primaria y preescolar
234	Maestros e instructores de nivel superior de la enseñanza especial
235	Otros profesionales de la enseñanza
241	Especialistas en organización y administración de empresas y afines
242	Profesionales del derecho
243	Archiveros, bibliotecarios, documentalistas y afines
244	Especialistas en ciencias sociales y humanas
245	Escritores, artistas creativos y ejecutantes
300	Técnicos y profesionales de nivel medio
400	Empleados de oficina
500	Trabajadores de los servicios y vendedores de comercios y mercados
600	Agricultores y trabajadores calificados agropecuarios y pesqueros
700	Oficiales, operarios y artesanos de artes mecánicas y de otros oficios
800	Operadores de instalaciones y máquinas y montadores
900	Trabajadores no calificados
000	Fuerzas armadas

Fuente: Clasificación internacional uniforme de ocupaciones (ISCO-88).

(*) El código 200 se utilizará por los profesionales científicos e intelectuales que no se pueden clasificar en ninguno de los otros códigos.

Anexo IV

Ejemplo de tabla de Movilidad Internacional

Para responder a la pregunta E.1 debe completar la tabla de países en orden cronológico inverso, es decir, de lo mas reciente a lo más lejano.

Supongamos el caso de un doctor que residió en España al inicio del año 96 hasta Mayo del mismo año. Tras esta etapa viaja a Italia de junio de 1996 a febrero de 1997 por motivos laborales. En marzo del mismo año regresó a España dónde continuó viviendo hasta mayo de 2000. En junio de ese año se desplazó a Francia a trabajar y estuvo allí hasta septiembre del año siguiente (2001). Y en octubre de 2001 regresó a España donde continúa residiendo hasta el momento. El resultado de este proceso puede verse reflejado en la siguiente tabla.

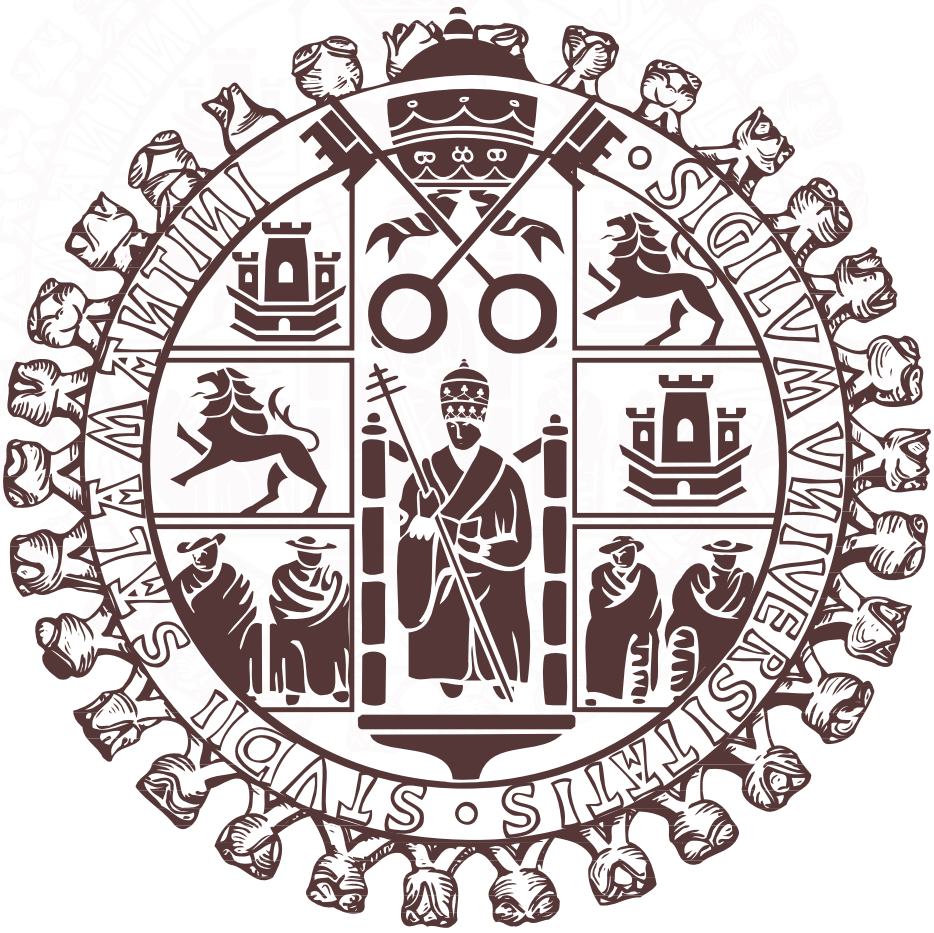
E. Movilidad internacional

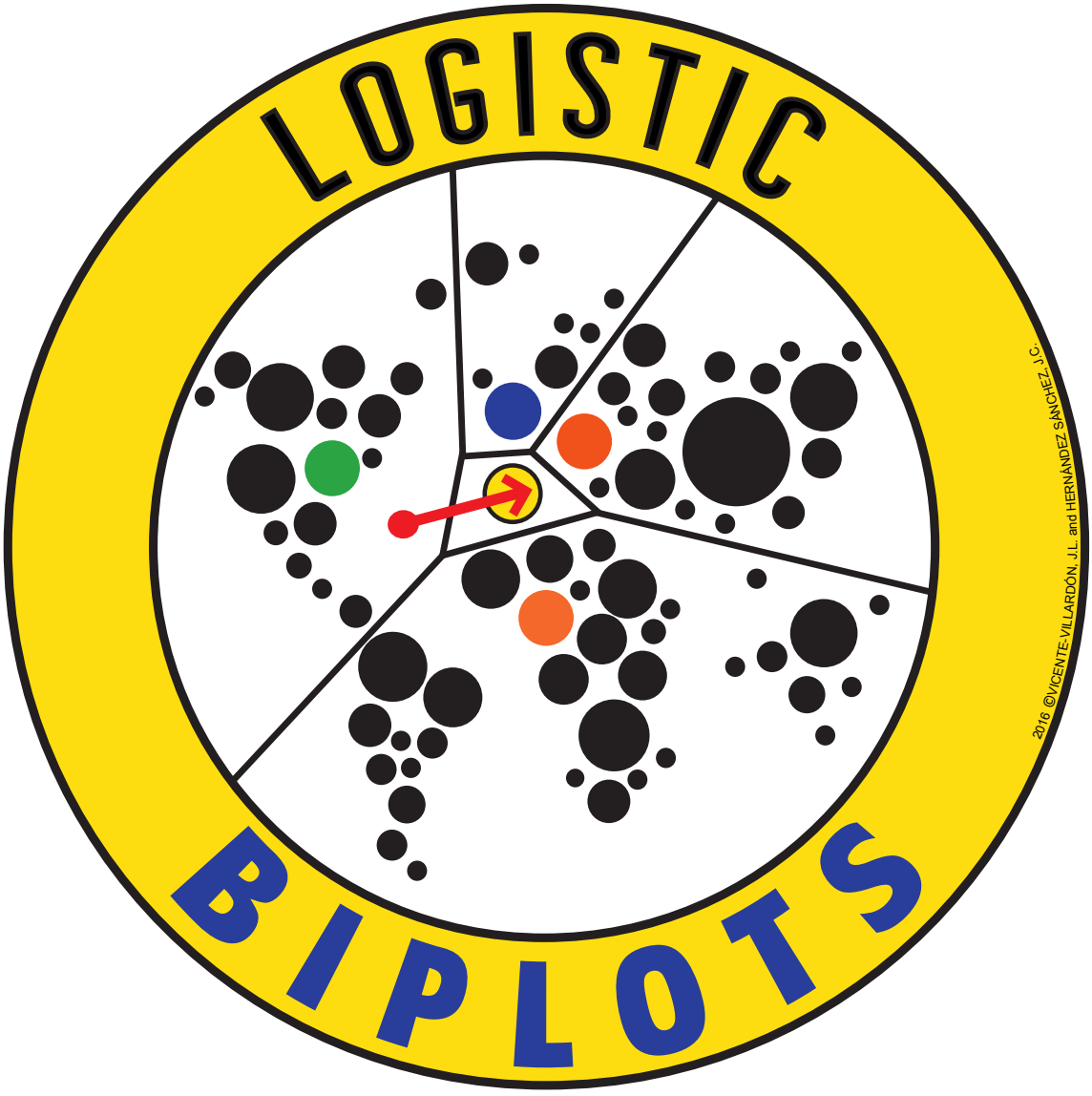
SI SOLAMENTE VIVIÓ EN ESPAÑA EN EL PERIODO 1996-2006 PASE AL APARTADO E.4

E.1 Realice una lista de los países en los que vivió en el periodo 1996-2006 *(incluyendo España)*

País	Código País	Periodo de residencia			
		Desde		Hasta	
		Mes	Año	Mes	Año
ESPAÑA		10	2001	12	2006
FRANCIA		06	2000	09	2001
ESPAÑA		03	1997	05	2000
ITALIA		06	1966	02	1997
ESPAÑA		01	1996	05	1996


INSTITUTO NACIONAL DE ESTADÍSTICA
INSTITUTO NACIONAL DE ESTADÍSTICA
INE


SIGILVM VNIVERSITATIS STVDII


LOGISTIC
BIPLOTS
2016 ©VICENTE-VILLARDÓN, J.L. and HERNÁNDEZ SÁNCHEZ, J.C.