Universidad de Salamanca

Departamento de Estadística

Doctorado en Estadística Multivariante Aplicada

Tesis Doctoral

Generalización del biplot logístico para dos o más matrices de datos

AUTOR: *Laura Vicente González– 70916732-G*

DIRECTOR: *José Luis Vicente Villardón*

2022

Departamento de Estadística

José Luis Vicente Villardón

Profesor Titular del Departamento de Estadística de la Universidad de Salamanca

CERTIFICA:

Que **Doña Laura Vicente González** ha realizado en la Universidad de Salamanca, bajo su dirección, el trabajo para optar al título del Doctorado en Estadística Multivariante Aplicada que presenta con el título de **Generalización del biplot logístico para dos o más matrices de datos**, autorizando expresamente su lectura y defensa.

Y para que conste, firma el presente certificado en Salamanca a 13 de diciembre de 2022.

Director: José Luis Vicente Villardón

Publicado en 13 de diciembre de 2022 por

Laura Vicente González

laura20vg@usal.es

Universidad de Salamanca

Agradecimientos

No se me ocurre mejor forma para hacer unos agradecimientos que utilizar a grandes científicas que han servido de inspiración y guía de tantas mujeres en el mundo de la ciencia, ya que, el que muchas mujeres podamos estar hoy aquí, se lo debemos en gran medida a ellas. Así que empezaré estos agradecimientos con el motivo que dio la empresaria Mary Kay Ash para hacerlos, *"Todo el mundo quiere ser apreciado, así que, si aprecias a alguien, no conviertas eso en un secreto"*.

Debo continuar dando el reconocimiento que merece a la Universidad de Salamanca y el Banco Santarder por darme la posibilidad de realizar esta tesis doctoral a través de los Programas de Contratos Predoctorales. De la misma forma, me gustaría incluir en este punto al grupo de investigación del CIALE con el que estamos trabajando, por permitir usar parte de sus bases dentro de esta tesis doctoral, y a Cursos Internacionales de la Universidad de Salamanca, cuyo contrato de investigación con mi tutor ha hecho posible dicha beca. No me puedo olvidar de agradecer a la Universidad Estatal de Milagro (UNEMI), y en especial a su rector el **Dr. Fabricio Guevara Viejó**, por la acogida y el aprendizaje que he realizado con ellos.

No hay mejor frase para agradecer al Departamento de Estadística, y en especial a su directora y coordinadora del programa de doctorado en Estadística Multivariante Aplicada, la **Dra. Mª Purificación Vicente Galindo**, por la acogida recibida desde que empecé en 2017 realizando las prácticas de grado allí e interesándome por este apasionante mundo de la ciencia, que con la frase de una de las científicas que podría definirse como una apasionada de la ciencia como fue Marie Curie, *"Me enseñaron que el camino*

del progreso no era rápido ni fácil". Gracias a cada uno de los profesores y compañeros que pacientemente me han resuelto dudas, dado sus opiniones y animado a continuar por este complejo camino.

A la primera persona en particular que quiero agradecerle el haber llegado hasta aquí, es a mi tío y tutor, el **Dr. José Luis Vicente Villardón**, que con gran esfuerzo y grandes momentos de desesperación por ambas partes, me ha enseñado, además de estadística, algo que ya dijo hace algún tiempo Margarita Salas, *"Una investigación básica de calidad es fundamental para un posterior desarrollo, porque de ella saldrán resultados no previsibles a priori"*. Gracias por desesperarme con tantos documentos que antes no había por donde coger y ahora parece que me empiezo a enterar de algo, y con el desorden de tu cabeza llena de conocimiento, tanto que es casi imposible seguirla la mitad de las veces. Gracias por la paciencia con mis múltiples actividades que hacen que te lleguen cosas a altas horas de la mañana, mi terrible escritura, mi catastrófico inglés y otras tantas y tantas cosas.

Linda B. Buck, cuando recibió el premio novel de Físiología o Medicina en 2004, dijo *"Como mujer y científica, espero sinceramente que el haber recibido el Nobel os envíe un mensaje a las mujeres jóvenes de todas partes: las puertas están abiertas para ellas y deben perseguir sus sueños"*, la persona a la que quiero darle las gracias en este párrafo no ha conseguido un Nobel, pero nos ha dado confianza y un lugar a todos los que estamos en el Departamento de Estadística gracias a perseguir sus sueños. Como mujer y como científica, ha tenido su peso en la familia Biplot, ha sido mi mami postiza durante el tiempo que estuvimos en Ecuador y la mami de muchos, incluida yo, en el mundo de la Estadística. **Dra. Purificación Galindo Villardón** gracias de corazón por todo lo que has luchado para que estemos hoy aquí.

¿Que mejor forma que utilizar la frase de una educadora para agradecer a quien me ha educado desde el minuto uno? Elizabeth Peabody decía algo que **mis padres** me han enseñado desde que era muy pequeña, *"No estudies para exhibir tus conocimientos, ni para ser admirado, ni para llamar la atención. Estudia por el placer derivado del sentimiento de*

energía que surge en la mente al ejercitar sus poderes en el razonamiento metafísico, científico o matemático". Gracias por enseñarme a buscar mi camino, animarme a estudiar aquello que me gusta y aguantar mis discursos sobre cosas que no sabíais ni que eran. Gracias por enseñarme el valor de la constancia, de la perseverancia, de la lucha, de la paciencia, y tantos otros que seguro que me dejo en el tintero.

La pareja de dos que lucha contra mis constantes comentarios sobre su carrera de *pinta y colorea* necesitarían un apartado largo para cada uno de ellos en este agradecimiento, pero hay una frase de María Teresa Ruiz que me recuerda a ambos, *"Para incentivar el gusto por la ciencia hay que devolverle la dignidad a los profesores"*. La pequeña de la casa, el terremoto, la pequeña maestrilla a la que no iba a querer nunca, pero que gracias a que está ahí, tengo como desestresarme a cualquier hora, tengo con quien compartir desesperaciones que terminan en risas, **Lucía**. Y la nueva incorporación a la familia, o ya no tan nueva, **Ri**, que llegando a Salamanca a través de la música, ya no le son extrañas palabrejas como PERMANOVA, RDA, PLS, o STATIS (no digo más que os hago spoiler). Gracias por aguantarme, por animarme, por ayudarme a desconectar y a reconectar, por tanto… simplemente, GRACIAS.

No puedo olvidar al resto de la familia, los que están en el mundo de la ciencia y los que no lo están, pues también tengo mucho que agradecerles. Seguro que no les importa que les robe su trocito en este apartado para centrarme en los abuelos, **Tina, José Manuel, Maru, Luis**, ¿qué haríamos sin las tan valiosas enseñanzas de los más mayores de la casa? Les dejaré una frase de Audrey Hepburn, *"Vivir es como avanzar por un museo: luego es cuando empiezas a entender lo que has visto"*, cada una de sus vivencias nos hacen ricos a todos los que nos las cuentan. Gracias por hacerme tan, tan rica.

Tengo que dejar un trocito de estos agradecimientos para mis compañeros de doctorado, a pesar de que nos ha pillado un momento en el que no hemos podido vivir en la biblioteca como generaciones anteriores. Marie Curie decía *"Uno nunca se da cuenta de lo que se ha hecho; uno solo puede ver lo que queda por hacer."*, ellos son los que te ayudan

a ver todo el camino que ya has recorrido. Gracias a todos, no podré nombres para no dejarme a nadie importante.

No quiero dejar de darle las gracias a los amigos y compañeros que han estado todo este tiempo a mi lado. Gracias a los que creen que ya soy doctora, incluso antes de haber empezado a escribir, los que confían sus trabajos en mis manos, los que me cubren cuando no llego y los que lo primero que preguntan es ¿qué más tienes de la tesis?. Gracias a quien vale una quedada cada varios meses para dar fuerza, como **Joni**, o quien aguanta mis desvelos en cada cerveza del martes noche, como **María**. Gracias a ese magnífico grupo de tiempo libre que me ayuda a desconectar y volver a empezar, a los que acaban de llegar y a los que llevan desde el principio aguantando, en especial a **Nacho, Nuria, Fati y Nando** quienes no se olvidan nunca de preguntar, *"Nunca estarás tan ocupado como para no pensar en los demás"* (Madre Teresa de Calcuta).

Por último, me voy a quedar con los que me acompañan y confían más en mí que yo misma, esos a los que les puedo decir lo mismo que May-Britt Moser, *"No siempre fui la mejor alumna con las calificaciones más altas, pero mis maestros vieron algo en mí y trataron de alentarlo"*, profesores y compañeros, gracias por alentarlo, el llegar hasta aquí también es gracias a vosotros.

Gracias.

Índice general

Índice de figuras **14**

Índice de tablas **17**

1. Introducción **1**

2. Objetivos **17**

3. Representaciones Biplot **21**

3.1. Introducción 22

3.2. Biplots clásicos para datos continuos 24

3.2.1. Biplots basado en la Descomposición en Valores Singulares . . . 24

3.2.2. HJ-Biplot 26

3.2.3. Biplot general de predicción 27

3.2.4. Biplot de interpolación 30

3.3. Biplot Logístico 32

3.3.1. Biplot Logísitico Clásico 33

3.4. Software biplot 37

3.4.1. Software de uso general 37

3.4.2. Software: Paquetes Comerciales 38

3.4.3. Software Paquetes Libres 41

3.4.4. R 44

3.5. Ejemplo biplot logístico 47

3.5.1. ¿Hay diferencias de género en los cambios provocados por la pandemia? . . . 47

4. MANOVA basado en distancias 65

4.1. Introducción . . . 66
4.2. Modelos Lineales . . . 69
4.2.1. Modelo Lineal General Multivariante . . . 70
4.2.2. Modelo Lineal Generalizado . . . 74
4.3. MANOVA . . . 75
4.3.1. MANOVA con un factor de variación . . . 76
4.3.2. Matrices para las combinaciones lineales . . . 81
4.3.3. Diseños más complejos . . . 84
4.4. Distancias . . . 85
4.4.1. Distancias para variables continuas . . . 86
4.4.2. Distancias para variables binarias . . . 87
4.4.3. Distancias para variables categóricas . . . 90
4.4.4. Distancias para variables de diferentes tipos . . . 91
4.5. PERMANOVA . . . 93
4.6. BOOTMANOVA . . . 94
4.6.1. Diseños con un factor de variación . . . 95
4.6.2. Diseños generales . . . 105
4.7. Representaciones gráficas . . . 107
4.7.1. Análisis de Coordenadas Principales . . . 107
4.7.2. Análisis de Coordenadas Principales de la matriz de medias . . . 109
4.7.3. Regiones de confianza bootstrap para los centroides . . . 111
4.8. software . . . 115
4.8.1. PRIMER-e (Clarke *et al.*, 2017) . . . 116
4.8.2. PAST (Hammer *et al.*, 2001) . . . 116
4.8.3. R (R Core Team, 2021) . . . 117
4.9. Ejemplo MANOVA basado en distancias . . . 118

4.9.1. *Colletotrichum graminicola* 119
4.9.2. Ejemplo 2. Proyecto HapMap 131

5. RDA para datos binarios **145**

5.1. Introducción 146
5.2. RDA 147
5.3. RDA para datos binarios 150
5.4. Biplot 152
5.4.1. Representaciones biplot asociadas a Análisis de la Redundancia para datos continuos 153
5.4.2. Representaciones biplot asociadas al Análisis de la Redundancia para datos binarios 154
5.5. Software 157
5.5.1. Paquetes comerciales 157
5.5.2. Paquetes de R 159
5.6. Ejemplo PDA 160
5.6.1. Arañas 160

6. PLS-BLR **179**

6.1. Introducción 180
6.2. PLSR 183
6.2.1. Algoritmo NIPALS 184
6.3. PLS-BLR 189
6.3.1. Componentes Separadas para Respuestas Binarias 191
6.3.2. Algoritmo PLS-BLR 194
6.3.3. Modelo de Regresión Logística 196
6.4. Biplot 197
6.4.1. Biplot PLS-R 197
6.4.2. Biplot PLS-BLR 198
6.5. Software 200
6.5.1. Software comercial 200

6.5.2. Paquetes de R 202
6.6. Ejemplo PLS-BLR 204
6.6.1. Vinos 204
6.6.2. Arañas 216

7. STATIS Tetracórico Dual **227**
7.1. Introducción 228
7.2. STATIS Dual para una matriz de datos continuos 229
7.2.1. Análisis de cada ocasión 229
7.2.2. STATIS Dual 231
7.3. STATIS Dual para una matriz de datos binarios: Tetra-STATIS dual 235
7.3.1. Análisis de cada ocasión 236
7.3.2. Tetra-STATIS Dual 243
7.4. Software 246
7.4.1. Paquetes comerciales 246
7.4.2. Paquetes de R 248
7.5. Ejemplo STATIS tetracórico 249
7.5.1. Estudio de la evolución de la opinión de los residentes en España sobre los efectos y las consecuencias del coronavirus 250

Conclusiones **261**

Bibliografía **263**

Anexos **283**

Anexo I: Encuestas CIS **I**
Efectos y consecuencias del Coronavirus (IV) I
Efectos y consecuencias del Coronavirus (V) XXI

Índice de figuras

3.1. Biplot de predicción con marcadores de escala para datos del consumo de proteínas en los países Europeas 28

3.2. Biplot de interpolación con tres variables 31

3.3. Representación de biplot logístico típico con escalas graduadas para variables binarias. Los datos usados para el biplot serán los del ejemplo de la sección 6.6 35

3.4. Representación del biplot logístico típico con flechas para variables binarias. Los datos usados para este biplot es la matriz de datos binarios utilizados en los ejemplos de capítulos posteriores (5.6 y 6.6) 36

3.5. Biplot Logístico para los hábitos antes y después de la pandemia 52

3.6. Biplot Logístico para los hábitos antes y después de la pandemia en el caso de las mujeres 55

3.7. Biplot Logístico para los hábitos antes y después de la pandemia en el caso de los hombres 58

4.1. Sumas de cuadrados de las distancias 97

4.2. Cálculo de las distancias. (a) Matriz de datos brutos. (b) Matriz simétrica que contiene las distancias. (c) Sumas de cuadrados totales. (d) Sumas de cuadrados dentro de los grupos. 99

4.3. Análisis de Coordenadas Principales de los tipos de cepas de maíz 126

4.4. Representación gráfica del PERMANOVA en a) las dimensiones 1 y 2, b) las dimensiones 1 y 3, c) las dimensiones 1 y 4 y d) las dimensiones 2 y 3. 127

4.5. Regiones bootstrap para los tipos de cepas de maíz 128

4.6. Dimensiones 1 y 2 del Biplot Logístico por el método del Descenso del Gradiente . . . 129
4.7. Dimensiones 2 y 3 y 2 y 4 del Biplot Logístico por el método del Descenso del Gradiente . . . 130
4.8. Primer plano principal del ACoP para los datos completos . . . 136
4.9. Dimensiones de la 3 a la 10 del ACoP para los datos completos . . . 137
4.10. Dimensiones de la 3 a la 10 del ACoP sobre los centroides para los datos completos . . . 139
4.11. Dimensiones de la 3 a la 10 del ACoP sobre los centroides para los datos completos . . . 140
4.12. Regiones de confianza bootstrap del HapMap. Dimensiones 1 y 2. . . . 142
4.13. Regiones de confianza bootstrap del HapMap. Dimensiones de la 3 a la 10. 143

5.1. Esquema de la realización del Análisis de la Redundancia para datos binarios . . . 150
5.2. Biplot sin restricciones de las especies de arañas con las variables ambientales . . . 164
5.3. Biplot restringido de las especies de arañas con las variables ambientales 165

6.1. Esquema de la realización de la Regresión de Mínimos Cuadrados Parciales 183
6.2. Triplot de Predicción PLS para los datos de los vinos . . . 209
6.3. Triplot PLS-BLR para los datos del vino que muestra sólo las variables con una predicción superior a 0,6 . . . 210
6.4. Triplot PLS-BLR para los datos del vino con grupos representados como convex hulls . . . 212
6.5. Regiones de predicción para cada variable por separado . . . 214
6.6. Regiones de predicción para la combinación de ambas variables . . . 214
6.7. Triplot PLS-BLR para los datos de las arañas con escalas para las variables 223
6.8. Triplot PLS-BLR para los datos de las arañas con flechas . . . 224
6.9. Regiones de predicción para cada especie . . . 225

7.1. Representación típica de la *Interestructura en STATIS* . . . 233

7.2. Esquema de la realización del STATIS Dual 236
7.3. Representación de la *Interestructura* 256
7.4. Representación de las *Estructura*: Correlaciones compromiso 257
7.5. Representación de la *Estructura*: Contribuciones del compromiso 258
7.6. Representación de las *Estructura*: Biplot STATIS del compromiso 259

Índice de tablas

4.1. Tabla de contingencia para el cruce de dos individuos 88
4.2. Resultados PERMANOVA de las especies del tipo de cepa de maíz 122
4.3. Resultados BOOTMANOVA de las especies del tipo de cepa de maíz . . . 122
4.4. Resultados de los contrastes para cada tipo de cepa de maíz 123
4.5. Contrastes a Posteriori de los tipos de cepas del maíz 125
4.6. Poblaciones de la fase III del proyecto HapMap con la codificación realizada en la aplicación práctica. 133
4.7. Varianza explicada por las coordenadas principales de los datos de Hapmap 135
4.8. PERMANOVA y BOOTMANOVA global para los datos de HAPMAP . . . 138
4.9. Contrastes del PERMANOVA y BOOTMANOVA para el proyecto HapMap 138
4.10. Comparaciones por parejas de los grupos de HapMap 144

5.1. Especies de arañas con la codificación realizada en la aplicación práctica. 162
5.2. Resumen de la independencia entre las presencias y ausencias de las diferentes tipos de arañas 167
5.3. Ajuste de las columnas en los modelos sin restricciones y restringido. Estas medidas de relación entre las variables observadas como función de las puntuaciones del sitio de muestreo para las primeras dos dimensiones 174
5.4. Porcentaje de varianza de las variables ambientales explicada por la ordenación 175

6.1. Descripción de las variables 206
6.2. Varianza de los predictores explicada en la dimensión reducida 208
6.3. Medidas de ajuste para cada respuesta 209

6.4. Calidad acumuladas de las columnas (porcentaje de la variabilidad de la variable correspondiente a la primera dimensión y a la suma de las dos primeras) 211
6.5. Especies de arañas utilizadas en el ejemplo 216
6.6. Datos Arañas: Especies de Araña Lobo 218
6.7. Datos Arañas: Variables ambientales 219
6.8. Varianza Explicada 221
6.9. Calidad de las columnas 221
6.10. Medidas de ajuste para los datos de las arañas 222

7.1. Correlaciones entre ocasiones 255

Capítulo 1

Introducción

Como se intuye en el título de este trabajo, "*Generalización del biplot logístico para dos o más matrices de datos*", la parte más importante del mismo es la propuesta de representaciones gráficas asociadas a cada uno de los métodos que se desarrollarán, es decir, incidiremos en los aspectos exploratorios de las técnicas a través de representaciones visuales que nos ayuden en la interpretación de la estructura de uno o varios conjuntos datos. Dentro de este tipo de representaciones, centraremos nuestro trabajo en los denominados métodos biplot, que son representaciones conjuntas de las filas y las columnas de la matriz de datos.

A lo largo de la historia, los dibujos y/o gráficos han sido fundamentales para plasmar aquello que se consideraba importante en la época, se ha llegado incluso a narrar historias a través de estas representaciones gráficas.

En el campo de las matemáticas y de la estadística no iba a ser diferente, se han utilizado gráficos desde hace siglos. El Instituto Nacional de Estadística (INE) (2019) publicó un trabajo que contenía la historia de los gráficos estadísticos. Las primeras representaciones gráficas en el campo de las matemáticas están datadas de mediados del siglo VI a.C. con los Teoremas de Tales, aunque no es hasta mediados del siglo XVIII donde se construyen los primeros gráficos estadísticos. Este tipo de representaciones han ido evolucionando a lo largo de la historia para complementar en todo momento diversas

técnicas que han sido desarrolladas por los estadísticos.

El documento emitido por el INE en 2019 en el que se recoge el desarrollo de los gráficos a lo largo de la historia, finaliza con los gráficos boxplot y no incluye las representaciones biplot. Los biplots son coetáneos de boxplot, sin embargo han tenido una menor difusión y, por lo tanto, en la actualidad son menos conocidos.

Los grandes avances en las técnicas de recogida de datos hacen que cada vez sea más frecuente encontrar bases de datos con un gran número de variables medidas en una muestra de individuos. En cualquier estudio experimental dispondremos, en general, de una matriz de datos cuyas filas se corresponden con un conjunto de individuos y cuyas columnas contienen características medidas sobre ellos. Las variables pueden ser de diversos tipos, desde cuantitativas a binarias, nominales a ordinales y, normalmente, todas del mismo tipo.

Es claro que, en este contexto, necesitamos de las técnicas multivariantes ya que las técnicas tradicionales de la Estadística Univariante, repetidas para cada variable, no son suficientes para recoger la estructura de los datos. El principal objetivo de los estudios multivariantes es el de establecer las similitudes y diferencias entre individuos, las posibles relaciones entre variables y las conexiones entre ambos, y todo ello, usando simultáneamente toda la información disponible. La mayor parte de las técnicas multivariantes utilizadas en este trabajo serán desarrolladas desde un punto de vista exploratorio si bien, en algunas de ellas, podemos incluir aspectos inferenciales como, por ejemplo, el caso en que tenemos los individuos divididos en grupos, queremos compararlos y necesitamos una significación estadística.

La mayor parte de las técnicas multivariantes desarrolladas en la literatura clásica trabajan con matrices de datos continuos en las que el número de individuos es mucho mayor que el de variables medidas sobre ellos. Por ejemplo, cuando se dispone de una única matriz, el método más popular es el Análisis de Componentes Principales (ACP/P-

CA) o su forma biplot. El ACP fue propuesto por Pearson (1901) hace más de 120 años y posteriormente formalizado por Hotelling (1933, 1936). A pesar de su antigüedad, aun sigue siendo una técnica popular, hasta el punto de que solamente en Google Scholar aparecen más de tres millones de referencias, más de cincuenta mil de las cuales son solamente en el último año. En sus inicios, el ACP se ha utilizado fundamentalmente en matrices con muchos más individuos que variables, incluso ha sido tratado como una forma de estimar las componentes en la población o como una posible solución del Análisis Factorial (AF) en la que el número de variables no es muy elevado. Más recientemente se ha utilizado de una forma más general, como una técnica exploratoria que permite incluso que el número de variables sea mayor que el de individuos.

En el año 1971, K. R. Gabriel (Gabriel, 1971) propone una representación conjunta de las filas y las columnas de la matriz de datos, directamente relacionada con el ACP y el AF. Especialmente en el caso exploratorio, el biplot añade herramientas de interpretación al ACP. A pesar de las claras ventajas que introduce en la exploración de los datos, los biplots no han sido tan populares como cabría esperar, solamente se encuentran 99600 referencias de las cuales 6730 son en el último año. Desde el grupo de trabajo en el que nos integramos, entendemos que este tipo de técnicas de tipo visual, podrían complementar, de forma satisfactoria, todos aquellos estudios en los que se utiliza un ACP, un AF o incluso muchas otras técnicas relacionadas.

Este tipo de métodos permite la visualización de la estructura de los datos, más allá de la típica lista de p-valores asociados a modelos estadísticos difíciles de entender y de interpretar, y que proporcionan la mayor parte de los investigadores. Si bien, también pueden entenderse, en algunas ocasiones, como la forma de visualizar los modelos complejos o incluso como métodos de diagnóstico de posibles modelos inferenciales que pueden ajustarse adecuadamente a los datos.

Comenzaremos nuestro trabajo revisando los conceptos fundamentales de los métodos biplot, ya que es la técnica en la que se basan muchas de las propuestas del mismo.

En el capítulo 3 desarrollaremos con mayor profundidad los fundamentos del biplot y sus principales características, ya que será el hilo conductor y punto fundamental de este trabajo.

Cuando en lugar de variables continuas, las variables medidas son binarias, nominales u ordinales, las representaciones clásicas para datos continuos no son adecuadas. En la literatura hay muchas propuestas entre las que cabe destacar el sistema denominado GIFI que se detalla en el libro de Gifi (1990) o en Michailidis y De Leeuw (1998). El sistema se basa en la idea de cuantificación de las variables categóricas para representarlas finalmente mediante biplots.

Recientemente se han desarrollado representaciones Biplot, basadas en respuestas logísticas en lugar de en las cuantificaciones, para una única matriz de datos binarios Vicente-Villardon *et al.* (2006); Demey *et al.* (2008), para datos nominales Hernández-Sánchez y Vicente-Villardón (2017), para datos ordinales Vicente-Villardón y Hernández Sánchez (2014) o incluso para datos mixtos con varios tipos de variables Vicente-Villardón y Hernández-Sánchez (2020). Las propuestas se basan en desarrollos anteriores para modelos bilineales generalizados que pueden encontrase en Gabriel (1998). En el resto del trabajo, utilizaremos esta aproximación basada en relaciones logísticas entre las dimensiones o componentes y nos centraremos en la aplicación a datos binarios. En la Sección 3.3 describiremos el biplot logístico para datos binarios como base de desarrollos posteriores para este tipo de datos.

Si nos centramos en las variables, es posible que nuestra base de datos tenga dos o más matrices de datos que, en algunos casos, pueden tener papeles no simétricos, por ejemplo, que una de las matrices contenga un conjunto de variables predictoras y la otra esté formada por un conjunto variables respuesta. Cuando nos encontramos en este caso, generalmente buscaremos modelos que permitan predecir las respuestas a partir de los predictores. En cada uno de los dos conjuntos podemos tener variables de cualquiera de los tipos mencionados antes.

Veamos algunos ejemplos que trataremos en este trabajo:

- Tenemos una matriz de datos (continuos o categóricos) y los individuos están divididos en grupos procedentes de diversas poblaciones o como resultado del uso de diferentes tratamientos en el diseño experimental. Si definimos una matriz de datos binarios con los indicadores de los grupos, la matriz de predictores es binaria y la de datos contiene las respuestas. Este es el caso típico en el que se aplica el Modelo Lineal General (MLG). Deseamos establecer las diferencias entre grupos mediante una significación estadística y además, representar las posibles diferencias mediante un gráfico. Actualmente hay muchas situaciones en las que el número de variables es muy elevado e incluso, mayor que el de individuos, por ejemplo, en los estudios genómicos en los que se mide la expresión o el comportamiento de un gran número de genes en un grupo reducido de individuos. En estos casos particulares es necesario adaptar algunas de las técnicas multivariantes clásicas que no permiten que el número de variables sea mayor que el de individuos. En el capítulo siguiente, estudiaremos métodos basados en el Modelo Lineal General tratando de realizar diferentes propuestas para evitar alguna de sus limitaciones en relación al tipo de datos o al número de variables. Describimos los Modelos Lineales y sus extensiones y proponemos la utilización de métodos bootstrap para encontrar la significación y la representación gráfica de los grupos mediante Coordenadas Principales. Presentamos esta parte en el Capítulo 4.

- Disponemos de respuestas binarias y de una matriz de predictores continuos de rango completo. Deseamos estudiar la relación entre ambos conjuntos y representarla en un gráfico (biplot). Proponemos el que hemos denominado Análisis de la Redundancia con respuestas binarias en el que reducimos la dimensión de las respuestas, realizamos una representación mediante un biplot logístico y proyectamos sobre el mismo las variables predictoras para explorar la relación entre ambos conjuntos. Presentamos esta parte en el Capítulo 5.

- Disponemos de respuestas binarias y de una matriz de predictores continuos que

no es de rango completo porque hay variables relacionadas o el número es muy elevado. Deseamos estudiar la relación entre ambos conjuntos y representarla en un gráfico (biplot). El Análisis de la Redundancia del punto anterior no es adecuado porque necesita de predictores de rango completo. Proponemos una generalización de la técnica de Mínimos Cuadrados Parciales (PLS) que permita la inclusión de respuestas binarias. La representación visual final es una combinación de un biplot logístico binario y uno clásico que nos permite estudiar las relaciones entre los dos tipos de variables. Presentamos esta parte en el Capítulo 6.

- Disponemos de varias matrices binarias con las mismas variables en todas ellas (y probablemente distintos individuos) y deseamos buscar una estructura factorial común en las variables que tenga en cuenta las correlaciones entre las mismas, evitando la posible estructura espúrea producida por diferencias en la localización. Proponemos una adaptación de los métodos STATIS a datos binarios usando las matrices de correlaciones tetracóricas como objetos representantes de cada matriz. La represntación final será un biplot logístico consenso para todas las matrices. Presentamos esta parte en el Capítulo 7.

La lista de posibilidades expuesta no pretende ser exhaustiva, ni en la forma de las matrices ni en las soluciones posibles a los distintos problemas; se trata simplemente de los casos que hemos tratado en este trabajo y que deben entenderse como parte de un proyecto general del equipo de trabajo que se irá extendiendo en el futuro a otras situaciones y que resultará en nuevas publicaciones y trabajos de tesis. Sería imposible en un único trabajo desarrollar todas las posibilidades para distintos tipos de datos y diversas posibles soluciones a los problemas planteados.

A continuación, describiremos con más detalle lo que hemos incluido en cada uno de los capítulos para más de una matriz de datos.

Muchas de las técnicas desarrolladas en este trabajo se fundamentan en los Modelos Lineales, por ello se han desarrollado en la Sección 4.2. Esta sección estará dividida

en dos subsecciones de gran importancia para el desarrollo del trabajo. En la subsección 4.2.1 se desarrollan los Modelos Lineales Generales. Cuando el objetivo de nuestro trabajo es establecer las diferencias entre una serie de grupos o tratamientos, tradicionalmente la técnica más utilizada es el Análisis de la Varianza (ANOVA), basado en el Modelo Lineal General Univariante. Un gran número de investigaciones se limitan a emplear únicamente estas técnicas univariantes para cada una de las variables por separado, sin embargo, este tipo de prácticas incrementa el riesgo Tipo I e ignora las dependencias entre variables. La alternativa multivariante más extendida para la comparación de grupos es el Análisis Multivariante de la Varianza (MANOVA). Esta técnica, de forma análoga al ANOVA, se fundamenta en el Modelo Lineal General Multivariante y estudia la variabilidad entre los grupos buscando las diferencias significativas entre ellos.

Para poder aplicar los Modelos Lineales Generales, tanto el univariante como el multivariante, se requiere la verificación de una serie de hipótesis de partida. En primer lugar, los datos deben seguir una distribución normal (o normal multivariante). En segundo lugar los grupos deben ser homocedásticos, es decir, las varianzas de todos los grupos que se comparen deben ser iguales, debe existir igualdad de las matrices de covarianzas en el caso multivariante. Por último, se debe tener en cuenta que el número de individuos sea mayor que el número de variables respuesta empleadas, ya que en caso contrario no podrían realizarse los cálculos matriciales del ML General.

En los datos genómicos, por ejemplo, las variables se caracterizan por ser marcadamente asimétricas, se decir, no suelen seguir distribuciones normales, y el número de individuos en muchos casos es menor que el número de las variables. Sin embargo, la mayor parte de investigadores siguen empleando técnicas paramétricas y univariantes que no son óptimas. Los métodos multivariantes empleados en la búsqueda de significación no han tenido una gran aceptación, es posible que se deba a su complejidad o a sus condiciones de aplicación.

Las hipótesis básicas que deben cumplirse para poder emplear ML General Multiva-

riante en la población de estudio, no se cumplen en un gran número de situaciones, por ello existen técnicas alternativas que permitan el estudio de este tipo de conjuntos de datos. En casi todas las propuestas y análisis que se desarrollan en este trabajo, el ámbito mayoritario de aplicación es el de la Ecología, en el que algunas de estas técnicas fueron desarrolladas, y que mantienen la aplicación de este tipo de métodos no paramétricos.

Cuando los datos de aplicación no siguen una distribución normal, existen diversas alternativas, emplear técnicas de remuestreo para estimar la distribución del estadístico de contraste o utilizar Modelos Lineales Generalizados entre ellas. En este trabajo emplearemos ambos métodos para construir las técnicas no paramétricas que se adapten a los datos binarios, antes de asociarlas a sus representaciones biplot. Por este motivo, en la sección 4.2.2, se describen los Modelos Lineales Generalizados.

Para el caso en que el número de individuos de la muestra es menor que el de variables respuesta se ha estudiado con menor profundidad, sin embargo, es posible encontrar diferentes técnicas que permiten realizar este tipo de estudios, como por ejemplo la prueba de Mantel, ANOSIM (Clarke, 1993) o el Análisis Multivariante de la Varianza basado en distancias y permutaciones (PERMANOVA) (Anderson, 2001; McArdle y Anderson, 2001).

En el capítulo 4 recogeremos dos métodos que permiten realizar este tipo de análisis. En primer lugar, en la sección 4.5, desarrollaremos el PERMANOVA que consiste en combinar el Análisis de Permutaciones y el Análisis Multivariante de la Varianza. A continuación, en la sección 4.6 desarrollaremos de forma teórica una propuesta alternativa al PERMANOVA a la que hemos denominado BOOTMANOVA o Análisis Multivariante de la Varianza basado en distancias y bootstrap. El BOOTMANOVA se presenta como una técnica similar a las mencionadas anteriormente, pero empleando técnicas Bootstrap para el remuestreo. Los métodos bootstrap fueron propuestos por Efron (1979); Efron y Tibshirani (1986, 1994) y se basan fundamentalmente en el remuestreo con reposición. La principal diferencia con el Análisis de Permutaciones propuesto por Neyman y S.

(1923) es que las técnicas bootstrap emplean el muestreo con reposición en lugar de las permutaciones.

Cuando empleamos un gran número de observaciones, el análisis de permutaciones presenta más dificultades ya que, para realizar el análisis completo sería necesaria la realización de todas las posibles permutaciones de los datos, que implicaría un gran coste computacional. Generalmente no es posible su realización y, de forma alternativa, se coge una muestra lo suficientemente grande como para ser representativa de los datos reales. Por el contrario, en el caso de las técnicas bootstrap no se presenta esta problemática, ya que, en todo momento, es el investigador el encargado de elegir el número de remuestreos con reposición que desea realizar.

Existen multitud de trabajos en la bibliografía que utilizan tanto las técnicas bootstrap como el análisis de permutaciones, en algunos de ellos de forma complementaria y en otros realizando una comparación entre ellas (ter Braak, 1992; Præstgaard, 1995; Cheng y Palmer, 2013).

En ese mismo capítulo (4), incluiremos una representación gráfica, similar al Análisis Canónico que tradicionalmente ha acompañado al MANOVA en su forma paramétrica, para ilustrar los resultados obtenidos tanto en el PERMANOVA como en el BOOTMANOVA. El Análisis Canónico sobre las medias llevará asociadas regiones de confianza que se calcularán empleando métodos de remuestreo bootstrap dentro de cada grupo. También incluiremos los software para desarrollar las técnicas descritas y un par de ejemplos para ilustrar la aplicabilidad de estas técnicas.

Para los casos en que los datos no sigan una distribución normal, utilizaremos Modelos Lineales Generalizados, desarrollando dos técnicas diferentes basadas en respuestas logísticas. La primera será el Análisis de la Redundancia para datos de respuesta binaria que será desarrollada en el capítulo 5, la segunda será la Regresión de Mínimos Cuadrados Parciales para datos de respuesta binaria que se describirá en el capítulo 6.

Ambos casos se caracterizan porque partimos de dos matrices de datos con papeles no simétricos, una de ellas contendrá los predictores de carácter continuo y la otra las respuestas que serán de tipo binario. También, en ambos casos se buscará el mismo objetivo, explicar la matriz de respuestas a partir de las variables explicativas.

Tradicionalmente, cuando tenemos una matriz de predictores y una de respuestas, podemos emplear la Regresión Lineal Multivariante (MLR) para explicar las respuestas a partir de los predictores, sin embargo, cuando no se cumplen las condiciones de aplicación, como por ejemplo uno de los dos conjuntos tiene variables binarias, es necesario buscar alternativas válidas para poder realizar este tipo de análisis. Con este fin surgen las dos técnicas que detallaremos en estos capítulos.

Cuando la base de datos con la que estamos trabajando tiene más de una variable respuesta y, por lo tanto, no puede emplearse la MLR, surge en la literatura el Análisis de la Redundancia para datos continuos, que fue propuesto por Rao (1964) y más tarde redescubierto por van den Wollenberg (1977) como una alternativa al Análisis Canónico de Correlaciones (CCA). Esta técnica también se conoce como Análisis de Componentes Principales Restringido o Análisis de Componentes Principales con información externa.

La idea original del Análisis de la Redundancia para datos continuos (RDA) se ha desarrollado en la sección 5.2. De forma resumida podríamos afirmar que esta técnica busca la mejor combinación lineal de las variables predictoras que maximicen la varianza explicada de las variables respuesta. Empleará una combinación de la Regresión Lineal Multivariante y el Análisis de Componentes Principales.

El objetivo principal del capítulo 5 reside en la generalización del algoritmo del Análisis de la Redundancia para datos continuos a datos de respuesta binaria. Este nuevo algoritmo propuesto se encuentra recogido en la sección 5.3 y ha sido publicado por Vicente-Villardon y Vicente-Gonzalez (2021). De forma análoga al caso anterior, se em-

pleará la Regresión Logística y el Análisis de Componentes Principales.

También se incluirá en ese mismo capítulo, una representación biplot asociada a cada uno de los Análisis de la Redundancia, el software con el que se pueden desarrollar y un ejemplo que permita ilustrar la utilidad de la técnica descrita.

Por otro lado, cuando existen algunas condiciones de aplicación que no se cumplen, por ejemplo hay un bajo número de predictores o existen efectos redundantes, y por lo tanto no es posible la realización de la Regresión Lineal Multivariante, la técnica más utilizada es la Regresión de Mínimos Cuadrados Parciales (PLS-R). Uno de los casos más interesantes en los que se aplica la PLS-R es cuando el número de individuos es mucho menor que el número de variables y, por lo tanto, no es posible realizar la estimación de los parámetros de la MLR.

La Regresión de Mínimos Cuadrados Parciales ha sido desarrollada con mayor profundidad en la sección 6.2. Igual que la técnica anterior, empleará la Regresión Lineal Multivariante y las Componentes Principales, sin embargo, en este caso se buscarán las componentes de la matriz de predictores que sean más relevantes en la predicción de la matriz de respuestas reduciendo la dimensionalidad de ambas matrices de forma simultanea utilizando el algoritmo NIPALS.

En el caso de las respuestas binarias, generalmente se emplea el Análisis Discriminante PLS, que podría asemejarse a una Regresión PLS con variables dummy, en la que las respuestas se utilizan de forma lineal.

Sin embargo, los Modelos Lineales Generales no son los más adecuados, podría mejorarse utilizando la transformación logit a través de los Modelos Lineales Generalizados. Una posibilidad sería emplear la Regresión Logística Multivariante, pero presentaría las mismas limitaciones que presentaba la Regresión Lineal Multivariante en el caso continuo.

Por ello, en la sección 6.3, donde se encuentra la parte fundamental de ese capítulo (6), proponemos lo que hemos denominado Regresión Logística Binaria PLS que generaliza el algoritmo NIPALS incluyendo respuestas binarias.

Igual que en los casos anteriores, finalizaremos el capítulo con una revisión de los software con los que se pueden realizar estas técnicas, y un ejemplo que muestre la utilidad práctica de las mismas.

Cuando el número de matrices es mayor que dos, se debe recurrir a otro tipo de técnicas. En este trabajo nos centraremos en el caso en el que tenemos un conjunto de matrices de datos con las mismas variables, pero que no necesariamente están medidas en los mismos individuos. El objetivo será describir la estructura común a todas ellas teniendo en cuenta las correlaciones. Con esta finalidad encontramos en la literatura el método STATIS-Dual, que ha sido desarrollado en la sección 7.2, explicado partiendo de la matriz de correlaciones. Se incluirá en esta misma sección la representación biplot para datos continuos.

Estas técnicas han sido desarrolladas tradicionalmente para datos continuos, sin embargo, si las matrices son de tipo binario, los métodos tradicionales no son adecuados.

Es posible encontrar en la literatura algunas alternativas para datos binarios, por ejemplo, si se trata de los métodos STATIS tradicionales, en los que todas las matrices tienen los mismos individuos, es posible utilizar un STATIS basado en distancias, de la misma forma que realizábamos los MANOVAs basados en distancias, este método será denominado DISTATIS (Abdi *et al.*, 2005).

En este trabajo proponemos utilizar un coeficiente que, en lugar de distancia o similaridad entre individuos, establezca relación entre variables, por ello se propone la utilización de la matriz de correlaciones tetracóricas. En la sección 7.3 se recoge el méto-

do desarrollado, STATIS tetracórico Dual, incluyendo todas las adaptaciones necesarias. Del mismo modo que en el caso continuo, la sección contendrá la representación biplot para el STATIS tetracórico Dual.

Finalizaremos este último capítulo, de forma análoga a los anteriores, incluyendo los software con los que realizar este tipo de técnicas y un ejemplo que muestre la utilidad de estos métodos.

En resumen, podemos decir que, tras la introducción (capítulo 1) y los objetivos (capítulo 2), el primer capítulo de este trabajo resumirá los fundamentos teóricos necesarios para el desarrollo de las técnicas posteriores, los fundamentos del Biplot (capítulo 3).

La parte central de la tesis estará constituida por los capítulos 4, 5, 6 y 7, que contienen los MANOVAs basados en distancias, el Análisis de la Redundancia para datos de respuesta binaria, la Regresión Logística Binaria de Mínimos Cuadrados Parciales y el STATIS tetracórico Dual respectivamente. Para todas ellas se han desarrollado los fundamentos teóricos de las nuevas aportaciones teóricas y de las representaciones gráficas multivariantes asociadas que ayudan en su interpretación. Se han revisado los software que pueden ser utilizados para la realización de los cálculos y de las representaciones gráficas de todos los métodos expuestos en este trabajo. Por último, en cada uno de ellos se presentan ejemplos que pretenden demostrar la utilidad de las técnicas descritas en aplicaciones de datos reales que no siguen las condiciones tradicionales de aplicación.

Todos los capítulos contienen una sección de Notación previa al desarrollo del capítulo (capítulos 3, 4, 5, 6 y 7). Estas secciones resumirán la notación que se empleará a lo largo de los capítulos, ya que cada uno de ellos ha sido construido de forma independiente pensando en su publicación de manera individual.

Las bases utilizadas en este trabajo son, mayoritariamente, bases públicas de diversos proyectos previos realizados por la comunidad científica, y que están disponibles para

su utilización, o por los equipos de investigación con los que se ha trabajado previamente.

El primer ejemplo será para ilustrar las técnicas biplot y el STATIS tetracórico Dual, y se emplearán los datos de una encuesta realizada por el CIS entre 2020 y 2021 para estudiar la situación provocada por la pandemia del COVID-19 (sección 3.5 y 7.5). La segunda base que se utilizará para los ejemplos será sobre el estudio de la antracnosis de las plantas del maíz, será extraída de un proyecto realizado por el CIALE en el que se está colaborando (sección 4.9). La tercera base utilizada será la del proyecto público HapMap, que recogen datos binarios de los polimorfismos mononucleotídicos y los grupos vienen determinados por la población en la que se ha recogido la muestra (sección 4.9). La cuarta base de datos pertenece también a un ejemplo público que estudia la presencia o ausencia de diversos tipos de arañas lobo en una zona de dunas de los Países Bajos (secciones 5.6 y 6.6). Por último, la quinta base de datos es la más antigua de las utilizadas y permitirá crear un ejemplo simple que ilustre la BLR-PLS, será una base de datos de vinos españoles de Ribera de Duero y Toro de dos cosechas diferentes (sección 6.3).

La utilización de un gran número de bases de datos para ilustrar estas técnicas permite ver la versatilidad de las mismas y su utilidad en diversos ámbitos.

Cada uno de los capítulos del documento estará estructurado en una primera parte teórica, en la que se detalla y desarrolla la técnica a la que se le asociará la representación gráfica correspondiente, y una segunda parte del capítulo en la que se desarrollarán los cálculos necesarios para la realización de dicha representación gráfica. Por último, en todos los casos se mostrará el software desarrollado y un ejemplo para mostrar su aplicabilidad y clarificar su interpretación.

Al inicio de cada capítulo hemos incluido la notación que se utilizará después. Cada capítulo se ha concebido para que pueda leerse de forma independiente, con su propia notación, debido a que es muy difícil unificarla debido a la gran cantidad de entidades

diferentes a denotar. No obstante se ha tratado de que la notación sea lo más uniforme posible.

Capítulo 2

Objetivos

El objetivo general de nuestro estudio es avanzar en el desarrollo y propuesta de métodos multivariantes y de minería de datos que permitan trabajar con matrices de datos categóricos en general y binarios en particular, especialmente cuando se dispone de dos o más matrices.

Objetivo 1. Estudiar los algoritmos de reducción de la dimensión para una única matriz de datos categóricos desarrollando un marco general para la obtención de variables latentes relacionadas, mediante respuestas logísticas, con las variables observadas y proponer una posibles algoritmos para datos binarios.

Objetivo 1.1. Revisar las representaciones gráficas biplot, tanto para datos continuos como para datos binarios, de una única matriz de datos y sus aplicaciones a datos reales.

Objetivo 1.2. Estudiar métodos iterativos para realizar el cálculo de las Componentes Principales de una única matriz a través del algoritmo NIPALS y proponer una generalización para datos binarios.

Objetivo 2. Desarrollar alternativas basadas en los Modelos Lineales Multivariantes, cuando la matriz de respuestas contiene agrupaciones de individuos y no se veri-

fican las condiciones de aplicación de las técnicas clásicas. Este caso se puede considerar como la extensión a dos matrices cuando las respuestas son numéricas y los predictores binarios, por una parte, y cuando tenemos respuestas binarias usando distancias.

Objetivo 2.1. Realizar una revisión de los fundamentos de los Modelo Lineales Multivariantes y sus aplicaciones tradicionales.

Objetivo 2.2. Presentar alternativas basadas en distancias y métodos de remuestreo que se encuentran en la literatura para el caso en el que las condiciones de aplicación no se verifican y sus representaciones gráficas asociadas

Objetivo 2.3. Presentar diferentes paquetes de software que permiten realizar este tipo de análisis y construir un paquete propio para la realización de las nuevas técnicas propuestas que será colocado en el repositorio CRAN.

Objetivo 2.4. Aplicar dichas técnicas a conjuntos de datos genómicos reales para mostrar su aplicabilidad.

Objetivo 3. Presentar técnicas de integración de dos matrices de datos con papeles no simétricos basadas en modelos de respuesta logística con reducción de la dimensión de la matriz de respuestas.

Objetivo 3.1. Estudiar los métodos de Análisis de la Redundancia desarrollados en la literatura y su posible aplicación en datos con diferentes tipos de variables respuesta.

Objetivo 3.2. Extender el Análisis de la Redundancia para datos continuos al caso en que la matriz de respuestas está compuesta por variables binarias.

Objetivo 3.3. Elaborar las funciones necesarias para realizar los cálculos del Aná-

lisis de la Redundancia para datos de respuesta binaria y sus representaciones biplot asociadas.

Objetivo 3.4. Estudiar la utilidad del Análisis de la Redundancia y sus representaciones gráficas asociadas en datos biológicos con respuestas binarias.

Objetivo 4. Extender los modelos compuestos por dos matrices de datos con papeles no simétricos basados en modelos de Regresión Logística en las que se reduce la dimensión de ambas matrices de forma simultánea.

Objetivo 4.1. Investigar sobre el uso de la Regresión de Mínimos Cuadrados Parciales en la actualidad y su utilidad en datos con respuestas dicotómicas.

Objetivo 4.2. Presentar una alternativa para la Regresión de Mínimos Cuadrados Parciales cuando la matriz de variables respuesta está formada por datos binarios.

Objetivo 4.3. Realizar un conjunto de funciones que permitan calcular los resultados de la Regresión de Mínimos Cuadrados Parciales para datos de respuesta binaria y sus representaciones biplot asociadas.

Objetivo 4.4. Observar las ventajas de la Regresión de Mínimos Cuadrados Parciales para datos binarios en varios ejemplos de datos reales.

Objetivo 5. Ampliar el estudio de la estructura común de varias (más de dos) matrices de datos cuando se ha medido el mismo conjunto variables binarias.

Objetivo 5.1. Realizar una breve descripción de las técnicas STATIS en su versión dual para variables continuas y sus representaciones asociadas.

Objetivo 5.2. Ampliar las técnicas STATIS en su versión dual cuando las matrices a integrar son de tipo binario.

Objetivo 5.3. Desarrollar las funciones para realizar los cálculos y representaciones gráficas de las técnicas desarrolladas para el STATIS Dual de matrices de datos binarios.

Objetivo 5.4. Presentar a través de un ejemplo con datos reales la utilidad de las técnicas desarrolladas para más de dos matrices de datos binarias

Capítulo 3

Representaciones Biplot

Notación

I : Número de individuos de la matriz de estudio.

J : Número de variables sometidas a estudio.

$\mathbf{X}_{(I\times J)}$: Matriz de datos transformados que queremos representar con I filas y J columnas.

R : Rango de la matriz $\mathbf{X}$.

S : Rango reducido de la matriz $\mathbf{X}$.

$\mathbf{A}_{(I\times s)}$: Matriz de marcadores filas con I filas y S columnas.

$\mathbf{B}_{(J\times s)}$: Matriz de marcadores columna con J filas y S columnas.

$\mathbf{E}_{(I\times J)}$: Matriz de residuales con I filas y J columnas.

$\mathbf{U}_{(I\times R)}$: Matriz de vectores singulares por la derecha ($\mathbf{u}_r$) con I filas y R columnas, que corresponden con el rango máximo de $\mathbf{X}$.

$\mathbf{V}_{(J\times R)}$: Matriz de vectores singulares por la izquierda ($\mathbf{v}_r$) con I filas y R columnas, que corresponden con el rango máximo de $\mathbf{X}$.

$\boldsymbol{\lambda}_r$: Valores singuales no negativos decrecientes de $\mathbf{X}^T\mathbf{X}$ y $\mathbf{X}\mathbf{X}^T$.

ρ^2 : Bondad de ajuste de la aproximación en dimensión reducida para la matriz completa.

π_{ij} : Probabilidad esperada que el individuo i tiene en la variable j.

3.1. Introducción

Desde tiempos remotos los dibujos o gráficos han sido fundamentales para plasmar aquello que se consideraba importante, incluso se ha llegado a narrar la historia a través de ellos.

En el campo de las matemáticas y la estadística, los primeros gráficos están datados de mediados del siglo VI a.C con los Teoremas de Tales. En el caso de la estadística las representaciones gráficas comenzaron a medidos del siglo XVIII y han ido evolucionando a lo largo de la historia para complementar los análisis realizados por los estadísticos. La historia de los gráficos estadísticos fue recogida por Instituto Nacional de Estadística (INE) (2019), aunque ellos finalizan este repaso gráfico con los boxplot y en nuestro caso hemos decidido emplear el biplot.

El biplot es contemporáneo al boxplot, pero menos extendido o conocido. Este gráfico nos permite, de forma análoga a la representación de dos variables en un diagrama de dispersión, representar tres o más variables sobre el plano. Esto va a hacer que los gráficos biplot sean de gran interés ya que, en espacios de dimensión reducida, van a permitir visualizar la distribución de variables e individuos.

Las representaciones biplot fueron introducidas por Gabriel (1971, 1972). El prefijo *bi* hace referencia a la representación de variables e individuos de forma simultánea en el mismo plano. A partir de su propuesta, a los que denominamos biplots clásicos, se han ido añadiendo diversas versiones que permiten ampliar este tipo de representaciones, desde el HJ-biplot (Galindo-Villardón, 1986) hasta las propuestas que se recogen en este documento. Una descripción exhaustiva de los biplots clásicos puede encontrarse en el primer libro enteramente dedicado al método (Gower y Hand, 1995) y más recientemente en Gower *et al.* (2011).

Vicente-Villardon *et al.* (2006) realizó una adaptación de los biplots clásicos cuando

la matriz de datos a representar contiene solamente variables binarias, proponiendo un biplot lineal generalizado en el que las respuestas a lo largo de las dimensiones son logísticas. En las propuestas de Demey *et al.* (2008); Babativa-Márquez y Vicente-Villardón (2021) se encuentra ampliaciones o mejoras del biplot logístico tradicional para adaptarlo a algunas situaciones concretas. También es posible encontrar una adaptación del biplot logístico para datos nominales en Hernández-Sánchez y Vicente-Villardón (2017) y para datos ordinales en Vicente-Villardón y Hernández Sánchez (2014).

Este capítulo contendrá una primera sección que incluye los biplots clásicos propuestos por Gabriel (1971) (sección 3.2.1). A continuación, dedicaremos un apartado al biplot propuesto por Galindo-Villardón (1986), el HJ-Biplot (sección 3.2.2). Las secciones 3.2.3 y 3.2.4 contendrán el desarrollo de los Biplot de predicción y de interpolación que serán de utilidad para la aplicación a las técnicas centrales de este documento. Los fundamentos del Biplot Logístico se recogerán en la sección 3.3.

En toda técnica estadística es necesario conocer el software que permite aplicarla, por ello se han recogido algunos de los software que permiten realizar representaciones biplot en la sección 3.4. Terminaremos el capítulo con un ejemplo que ilustre algunos de los biplots (sección 3.5).

3.2. Biplots clásicos para datos continuos

En esta sección describiremos los biplots clásicos para datos continuos.

3.2.1. Biplots basado en la Descomposición en Valores Singulares

Un biplot (Gabriel, 1971) es una representación gráfica de datos multivariantes. De la misma manera que un diagrama de dispersión muestra la distribución conjunta de dos variables, un biplot representa tres o más (Gabriel y Odoroff, 1990).

El biplot aproxima la distribución de una muestra multivariante en un espacio de dimensión reducida, normalmente de dimensión dos, y superpone sobre la misma representaciones de las variables sobre las que se mide la muestra (Gower y Hand, 1995). Las representaciones de las variables son normalmente vectores, y coinciden con las direcciones en las que mejor se muestra el cambio individual de cada variable.

Como ya habíamos mencionado en la introducción, el prefijo "bi" se refiere a la superposición, en la misma representación, de individuos y variables.

De acuerdo con Gabriel (1971), las representaciones biplot se basan en la reducción de la dimensión a través de la factorización de la matriz de partida en el producto escalar de otras dos.

Si $\mathbf{X}$ es la matriz de datos (transformados adecuadamente) que queremos representar, con I filas y J columnas, la factorización puede escribirse como:

$$\mathbf{X} = \mathbf{A}\mathbf{B}^T + \mathbf{E} \tag{3.1}$$

donde las filas de $\mathbf{A}$ definen un conjunto de puntos que usaremos como marcadores fila, las filas de $\mathbf{B}$ como marcadores columna y $\mathbf{E}$ la matriz de errores o residuales.

La versión original utiliza como factorización la denominada *Descomposición en Valores Singulares* (DVS) estrechamente relacionada con el Análisis de Componentes Prin-

cipales y el Análisis Factorial. La DVS puede definirse como:

$$\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T = \sum \lambda_r \mathbf{u}_r \mathbf{v}_r^T \tag{3.2}$$

donde $\mathbf{v}_r$ son los vectores singulares por la derecha contenidos en $\mathbf{V}$, es decir, los vectores propios de $\mathbf{X}^T\mathbf{X}$; u_r son los vectores singulares por la izquierda recogidos en la matriz $\mathbf{U}$, es decir, los vectores propios de $\mathbf{X}\mathbf{X}^T$, y λ_r los valores singulares no negativos ordenados de forma decreciente. Los cuadrados de los valores singulares λ_r^2 son también los valores propios no nulos de $\mathbf{X}^T\mathbf{X}$ y $\mathbf{X}\mathbf{X}^T$, que coinciden.

Para la matriz $\mathbf{X}$ cuyo rango es $R \leq min(I, J)$, es posible obtener una aproximación ($\hat{\mathbf{X}}$) de bajo rango ($S < R$) tomando los primeros S términos de la DVS.

$$\mathbf{X} = \hat{\mathbf{X}} + \mathbf{E} = \mathbf{U}_{(S)}\mathbf{\Lambda}_{(S)}\mathbf{V}_{(S)}^T + \mathbf{U}_{(-S)}\mathbf{\Lambda}_{(-S)}\mathbf{V}_{(-S)}^T = \sum_{r=1}^{S} \lambda_r \mathbf{u}_r \mathbf{v}_r^T + \sum_{r=s+1}^{R} \lambda_l \mathbf{u}_r \mathbf{v}_r^T \tag{3.3}$$

siendo (S) las S primeras columnas y $(-S)$ el resto de las columnas.

Es posible representar esta factorización con un biplot $\hat{\mathbf{X}}$, en dos o tres dimensiones, si puede considerarse que tiene una bondad de ajuste adecuada. Para calcular la bondad de ajuste de la aproximación a bajo rango se calcula el cociente entre la suma de cuadrados de los primeros S valores singulares y la suma de cuadrados de todos. Generalmente se presenta mediante porcentajes.

$$\frac{\sum_{r=1}^{S} \lambda_r^2}{\sum_{r=1}^{R} \lambda_r^2} * 100 \tag{3.4}$$

Volviendo a la factorización inicial (ecuación 3.1), una forma general para el biplot es tomar como marcadores fila a la matriz $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}^\gamma$ y como marcadores columna a $\mathbf{B} = \mathbf{V}\mathbf{\Lambda}^{1-\gamma}$ con γ comprendida entre 0 y 1. En función del valor γ seleccionado se obtienen los diferentes tipos de biplot clásicos:

JK – Biplot También denominado RMP-Biplot, con $\gamma = 1$. Los marcadores de las filas pueden definirse como $\mathbf{J}_{(S)} = \mathbf{U}_{(S)}\mathbf{\Lambda}_{(S)}$ y coinciden, en el espacio de las componentes

principales, con las coordenadas de los individuos. Los marcadores columna serán las proyecciones de los ejes originales en este mismo espacio y se pueden definir como: $\mathbf{K}_{(S)} = \mathbf{V}_{(S)}$. La calidad de representación en este tipo de gráfico es mejor en filas que en columnas.

GH – Biplot Con $\gamma = 0$, y también denominado CMP-Biplot. Los marcadores de las filas pueden definirse como $\mathbf{G}_{(S)} = \sqrt{(I-1)}\mathbf{U}_{(S)}$, cuya distancia se aproxima a la distancia de Mahalanobis en el espacio multidimensional con baja calidad de representación. Las coordenadas de las columnas $\mathbf{H}_{(S)} = \frac{1}{\sqrt{(I-1)}}\mathbf{\Lambda}_{(S)}\mathbf{V}_{(S)}$ son las cargas de un modelo de Análisis Factorial si los datos están estandarizados. Las varianzas y covarianzas entre las variables se aproximan de forma que las correlaciones entre las variables correspondan con los cosenos de los ángulos entre ellas y la variabilidad a través de la longitud del vector.

SQRT – Biplot Cuando $\gamma = 1/2$ el biplot obtenido no está relacionado con las técnicas más conocidas. Los marcadores fila son $\mathbf{U}_{(S)}\mathbf{\Lambda}_{(S)}^{1/2}$ y los marcadores columna como $\mathbf{V}_{(S)}\boldsymbol{\lambda}_{(S)}^{1/2}$. La calidad de representación de las entradas de la matriz de datos se mantiene.

3.2.2. HJ-Biplot

Los biplots clásicos permiten la representación simultánea de las filas y las columnas de una matriz, sin embargo, la calidad de representación de filas y columnas no es la misma ya que cambia en función del tipo de biplot que se esté utilizando. Si se busca que la representación simultánea mantenga la calidad de individuos y variables, Galindo-Villardón (1986) propone el HJ-Biplot, que es una representación simétrica tal y como se define en el Análisis de Correspondencias.

Para una matriz de datos $\mathbf{X}$, un HJ-Biplot es una representación gráfica multivariante cuyos marcadores fila serán identificados con la matriz $\mathbf{J}$ y los marcadores columna por $\mathbf{H}$.

Partiendo de la descomposición en valores y vectores propios de la matriz $\mathbf{X}$ como en 3.2, los marcadores se definen de la siguiente forma:

$$\mathbf{J}_{(S)} = \mathbf{U}_{(S)}\mathbf{\Lambda}_{(S)}, \tag{3.5}$$

$$\mathbf{H}_{(S)} = \mathbf{V}_{(S)}\mathbf{\Lambda}_{(S)}. \tag{3.6}$$

Como tanto los marcadores fila como los marcadores columna comparten los valores propios, se realizará la representación sobre el mismo sistema de referencia.

3.2.3. Biplot general de predicción

Es de sobra conocido que una factorización dada de una matriz $\mathbf{X} \approx \mathbf{A}\mathbf{B}^T$ define un biplot para $\mathbf{X}$ (Gabriel, 1971; Gower y Hand, 1995). Normalmente, la factorización está relacionado con las PCA y la DVS de forma que la factorización, en dimensión reducida (dos o tres), produce la mejor aproximación de $\mathbf{X}$ de bajo rango (Gabriel, 1971; Gower y Hand, 1995), aunque sería posible utilizar factorizaciones distintas de la DVS. Las filas de $\mathbf{A}$ y $\mathbf{B}$ se pueden usar como marcadores para los individuos y las variables de $\mathbf{X}$, respectivamente, en una representación gráfica (generalmente en dos o tres dimensiones) en la que cada elemento de la matriz de datos se puede aproximar como el producto interno de los marcadores

$$x_{ij} \approx \mathbf{a}^T{}_i\mathbf{b}_j = Proj\left(\mathbf{a}_i/\mathbf{b}_j\right) \otimes \parallel \mathbf{b}_j \parallel, \tag{3.7}$$

en la representación biplot, donde $\otimes$ es la proyección con signo.

Fundamentalmente, la interpretación consiste en proyectar cada punto fila sobre la dirección del vector que representa una variable. Para facilitar la interpretación, las variables se pueden complementar con escalas graduadas para obtener los valores aproximados de las entradas de la matriz (x_{ij}). Las proyecciones de todas las filas sobre las variables darán los valores esperados así como una aproximación del orden de todos los valores de las filas en la variable. Los marcadores para las escalas graduadas son fáciles de

obtener: para los marcadores para un valor particular de μ, en la dirección de $\mathbf{b}_j$, buscaremos un punto (x, y) que prediga μ en dicha dirección y que verifique que $\mu = b_{j1}x + b_{j2}y$.

Luego obtendremos

$$x = \frac{\mu\ b_{j1}}{b_{j1}^2 + b_{j2}^2}; \qquad y = \frac{\mu\ b_{j2}}{b_{j1}^2 + b_{j2}^2}. \tag{3.8}$$

La aproximación biplot permite la exploración de las principales características de los datos. Las propiedades de la aproximación de la matriz no depende de la elección de la factorización de **A** y **B**, aunque las propiedades separadas de cada conjunto de marcadores sí lo hacen, como se ha descrito anteriormente.

Una representación biplot típica con escalas para cada variable se puede observar en la figura 3.1. Se han utilizado los datos de los trabajos Hand *et al.* (1993); Gabriel (1981).

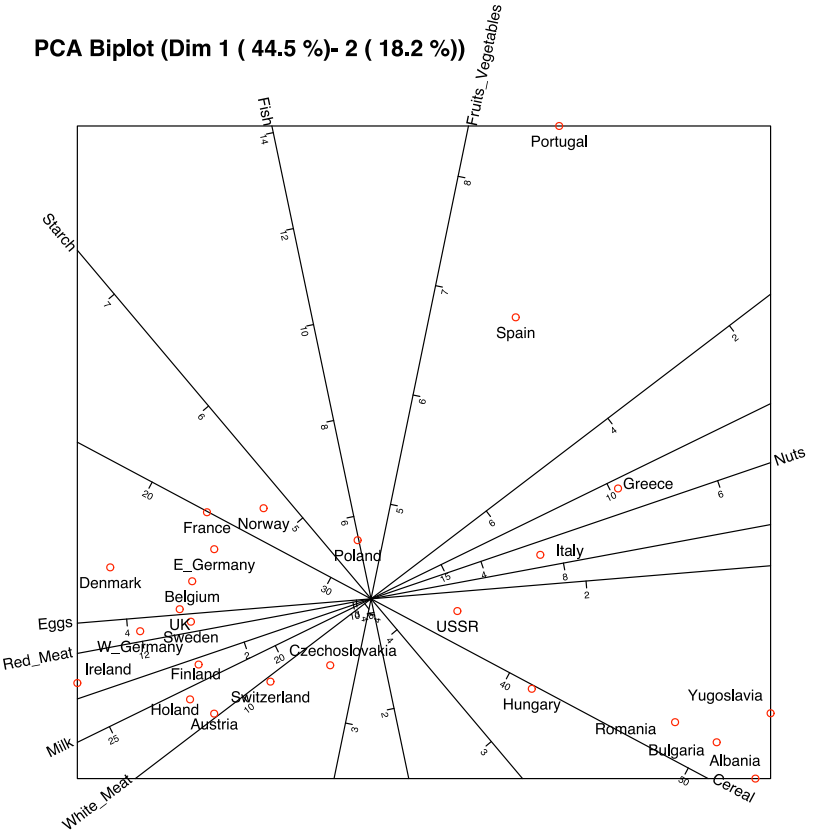

Figura 3.1: Biplot de predicción con marcadores de escala para datos del consumo de proteínas en los países Europeas

Si $\hat{\mathbf{X}}$ son los valores esperados del biplot en dimensión reducida $\hat{\mathbf{X}} = \mathbf{AB}^T$, la bondad de ajuste para la matriz completa se mide con el porcentaje de variabilidad recogida por la predicción, es decir

$$\rho^2 = \frac{\text{tr}\left(\hat{\mathbf{X}}^{\text{T}}\hat{\mathbf{X}}\right)}{\text{tr}\left(\mathbf{X}^{\text{T}}\mathbf{X}\right)}. \tag{3.9}$$

La bondad del ajuste medida de esta forma coincide con la mostrada antes en el caso de que el biplot se haya obtenido a partir de la DVS.

Incluso para los casos que tienen un buen ajuste general, algunas filas o columnas pueden no tener un buen ajuste, su variabilidad no está bien recogida en el biplot.

El ajuste para las columnas se encuentra en el vector

$$\rho^2_{(C)} = \text{diag}\left(\hat{\mathbf{X}}^{\text{T}}\hat{\mathbf{X}}\right) \div \text{diag}\left(\mathbf{X}^{\text{T}}\mathbf{X}\right), \tag{3.10}$$

donde $\div$ significa la operación elemento a elemento. El vector $\rho^2_{(C)}$ contiene el R^2 de la regresión para cada variable en $\mathbf{X}$ en las dimensiones de $\mathbf{A}$. Esto será denominado *calidad de representación* para las columnas como en los trabajos de Benzécri (1973); Greenacre (1984). Gardner-Lubbe *et al.* (2008) lo denominan *predictividad* de la columna. La calidad (o predictividad) contiene el porcentaje de la variabilidad recogida en las dimensiones (componentes) del biplot y serán usados para identificar las variables más relevantes para las dimensiones o aquellos cuya información se conserva en el biplot.

La bondad de ajuste para las filas es

$$\rho^2_{(R)} = \text{diag}\left(\hat{\mathbf{X}}\hat{\mathbf{X}}^{\text{T}}\right) \div \text{diag}\left(\mathbf{X}\mathbf{X}^{\text{T}}\right). \tag{3.11}$$

El vector $\rho^2_{(R)}$ contiene los cosenos de los ángulos formados por los vectores representando a un individuo en el espacio de alta dimensión y su proyección en baja dimensión.

Los cosenos al cuadrado también pueden ser interpretados como la calidad de representación de las filas y usarlo para identificar que dimensiones son útiles para diferencias a un individuo (fila) del resto. Los individuos con valores bajos generalmente se encontrarán cerca del origen.

Hasta ahora se ha descrito al denominado *"biplot de predicción"*, ya que pretende aproximar o predecir las entradas originales de la matriz de partida. Los biplots clásicos basados en la DVS son los ejemplos más comunes de biplot de predicción.

3.2.4. Biplot de interpolación

Existe otro tipo de biplot de interés en este documento, el *biplot de interpolación*. Este tipo de gráficos permitirá la proyección de nuevos individuos adicionales en el biplot empleando un conjunto de valores de los predictores. Será de gran utilidad para los capítulos 5 y 6 en los que buscaremos predecir la matriz $\mathbf{Y}$ a partir de los predictores de $\mathbf{X}$. A partir de las puntuaciones del biplot, será posible interpolar un nuevo punto y luego predecir las respuestas.

Volvemos a partir de una factorización dada de la matriz $\mathbf{X} \approx \mathbf{A}\mathbf{B}^T$ que define un biplot para $\mathbf{X}$. Si $\mathbf{B}^T\mathbf{B} = \mathbf{I}$, tendremos que

$$\mathbf{A} = \mathbf{X}\mathbf{B}. \tag{3.12}$$

Supongamos que tenemos una nueva observación $\mathbf{x} = (x_1, \dots, x_J)$. Podemos proyectar la nueva observación sobre el biplot con

$$\mathbf{a} = \mathbf{x}^T\mathbf{B} = \sum_{j=1}^{J} x_j \mathbf{b}_j. \tag{3.13}$$

Es una suma ponderada de los vectores $\mathbf{b}_j$ usando los valores observados x_j como ponderación. La interpretación gráfica de la interpolación se puede observar en la figura 3.2. La suma de los vectores se calculará multiplicando el centroide por el numero de puntos. Se puede ver un ejemplo en Gower y Hand (1995).

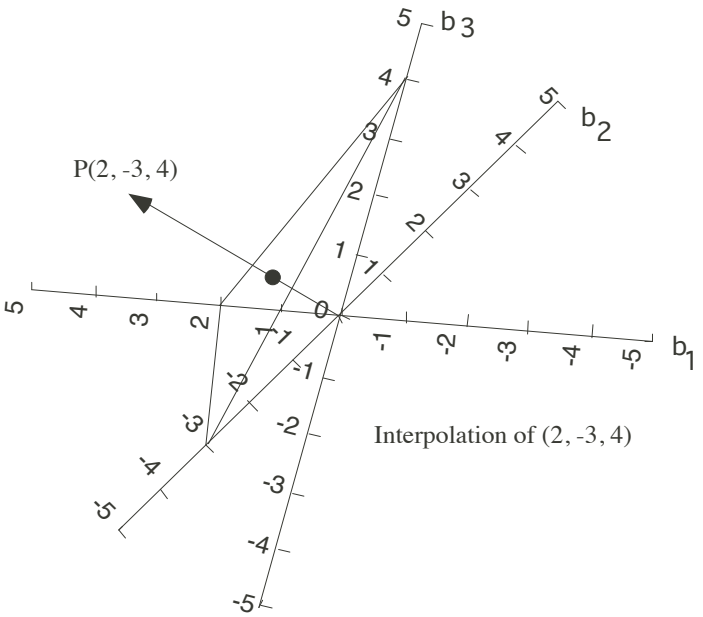

Figura 3.2: Biplot de interpolación con tres variables

Las direcciones serán las mismas que en el biplot de predicción, pero las escalas serán diferentes. Los marcadores para un valor fijado μ, estará ahora en el punto (x, y), con

$$x = \mu\ b_{j1}; \qquad y = \mu\ b_{j2}. \tag{3.14}$$

El **JK-Biplot** descrito antes sería especialmente últil para colocar este tipo de escalas de interpolación.

3.3. Biplot Logístico

Tanto los biplots clásicos, descritos en la sección 3.2.1, como el HJ-biplot, descrito en la sección 3.2.2, suponen implícitamente que la relación entre las variables observadas y las componentes es lineal. Si las variables observadas son de tipo binario, 0 cuando la característica se encuentra ausente y 1 cuando está presente, estas representaciones no son óptimas de la misma manera que una regresión lineal no captura bien la relación con una respuesta binaria.

Diversos autores han desarrollado alternativas para el análisis de datos binarios basándose en regresiones alternadas. Eeuwijk (1995) propone un biplot empleando regresiones bilineales generalizadas; Gabriel (1998) y de Falguerolles (1998) realizan modelos bilineales sobre tablas de contingencia de dos vías; en ese mismo año, Blazquez-Zaballos (1998) propone la utilización de regresiones generalizadas alternadas para estimar las respuestas dicotómicas. Schein *et al.* (2003) o Vicente-Villardon *et al.* (2006) proponen usar un Modelo Lineal Generalizado para las Componentes Principales para datos binarios. Ninguna de estas propuestas ha sido aplicada ampliamente ni sus propiedades estudiadas exhaustivamente.

La técnica más extendida para el análisis de datos categóricos (en particular, binarios) es el Análisis de Correspondencias Múltiples (MCA) y se considera como una forma particular de biplot para matrices binarias. La teoría completa del MCA en relación con los biplots fue desarrollada por Gabriel (1995) y Gower y Hand (1996), que propusieron las denominadas "regiones de predicción" como extensión de las lineas de predicción de los biplots clásicos. Los modelos de MCA se basan en la idea de homogeneidad o de cuantificación

En los denominados biplot logísticos, Vicente-Villardon *et al.* (2006) realiza una propuesta en la que la relación entre las variables observadas y las dimensiones es logística. Está relacionada con métodos psicométricos como la teoría de respuesta al ítem o los

rasgos latentes. En el artículo original se propone un algoritmo mediante regresiones generalizadas alternadas basado en el método de Newton-Raphson para maximizar la verosimilitud. El algoritmo presenta el mismo problema que la Regresión Logística en el caso de que haya separación, es decir, que para alguna de las variables binarias, los individuos que presentan presencia de la característica estén separados completamente (por un hiperplano) de los que presentan ausencia, en la representación final.

Demey *et al.* (2008) proponen un nuevo algoritmo que combina el Análisis de Coordenadas Principales, el Análisis de Cluster y la Regresión Logística para la realización de este tipo de biplots. El resultado es denominado *biplot logístico externo* ya que se trata de una aproximación en dos pasos (Coordenadas Principales para los individuos y Regresiones Logísticas para las variables) en lugar de obtener los marcadores de filas y columnas simultáneamente. Esta simplificación del procedimiento permitía evitar el problema de la separación a cambio de una pérdida en la bondad del ajuste.

Recientemente, Babativa-Márquez y Vicente-Villardón (2021) han propuesto un algoritmo que optimiza el ajuste de los parámetros empleando el algoritmo de gradiente conjugado no lineal o de mayorización-minimización.

En un artículo relacionado con esta tesis, Vicente-Gonzalez y Vicente-Villardon (2022) proponen una generalización del algoritmo NIPALS para datos binarios, más eficiente que el procedimiento original, en la que se obtienen las componentes de manera recursiva usando el método del gradiente.

3.3.1. Biplot Logísitico Clásico

Sea $\mathbf{X}$ una matriz $I \times J$ de datos binarios en la que se han medido J variables (o características) binarias en I individuos. Y sea $\pi_{ij} = E\left(x_{ij}\right)$ la probabilidad esperada que el individuo i tiene en la característica (variable) j. Si tenemos un modelo bilineal de la forma

$$\pi_{ij} = \frac{e^{\left(b_{j0}+\sum_s b_{js}a_{is}\right)}}{1 + e^{\left(b_{j0}+\sum_s b_{js}a_{is}\right)}}, \tag{3.15}$$

donde a_{is} y b_{js}, $(i = 1, \dots, I; j = 1, \dots, J; s = 1, \dots, S)$ son los parámetros del mismo, éstos se pueden usar como marcadores para las filas y las columnas de $\mathbf{X}$, respectivamente, de una forma similar a como hacíamos en la sección anterior. La ecuación (3.15) es una generalización del modelo bilineal utilizando la función de enlace $logit$

$$\text{logit}(\pi_{ij}) = b_{j0} + \sum_{s=1}^{S} b_{js}a_{ij} = b_{j0} + \mathbf{a'}_i\mathbf{b}_j.$$

En forma matricial,

$$\text{logit}(\mathbf{\Pi}) = \mathbf{1}_I\mathbf{b}_0^T + \mathbf{A}\mathbf{B}^T. \tag{3.16}$$

Excepto por el vector de constantes, tendremos un biplot en escala $logit$. Es necesario mantener esta constante porque la matriz de datos binarios no se puede centrar como ocurre en el caso continuo. Como tenemos un modelo generalizado, la geometría es muy similar al caso lineal. Los cálculos son similares a los realizados en los casos previos, pero añadiendo la constante. El marcador para cualquier valor de probabilidad p, es el punto (x, y) que predice p en la dirección de $\beta_j = \left(b_{j1}, b_{j2}\right)$, que es

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

La predicción también verifica que

$$\text{logit}(p) = b_{j0} + b_{j1}x + b_{j2}y.$$

Luego, obtenemos

$$x = \frac{\left(\text{logit}(p) - b_{j0}\right)b_{j1}}{b_{j1}^2 + b_{j2}^2}; \quad y = \frac{\left(\text{logit}(p) - b_{j0}\right)b_{j2}}{b_{j1}^2 + b_{j2}^2}.$$

El punto en la dirección β_j que predice $0,5$ ($logit\,(0,5) = 0$), es

$$x = \frac{-b_{j0}\,b_{j1}}{b_{j1}^2 + b_{j2}^2}; \quad y = \frac{-b_{j0}\,b_{j2}}{b_{j1}^2 + b_{j2}^2}.$$

El punto que predice 0,5 siempre sería el origen si restringimos la intersección a 0.

La interpretación es fundamentalmente la misma que el biplot lineal, excepto que los marcadores de probabilidades equidistantes no son necesariamente equidistantes en el gráfico.

Un biplot logístico típico con escalas graduadas para las variables se encuentra en la figura 3.3.

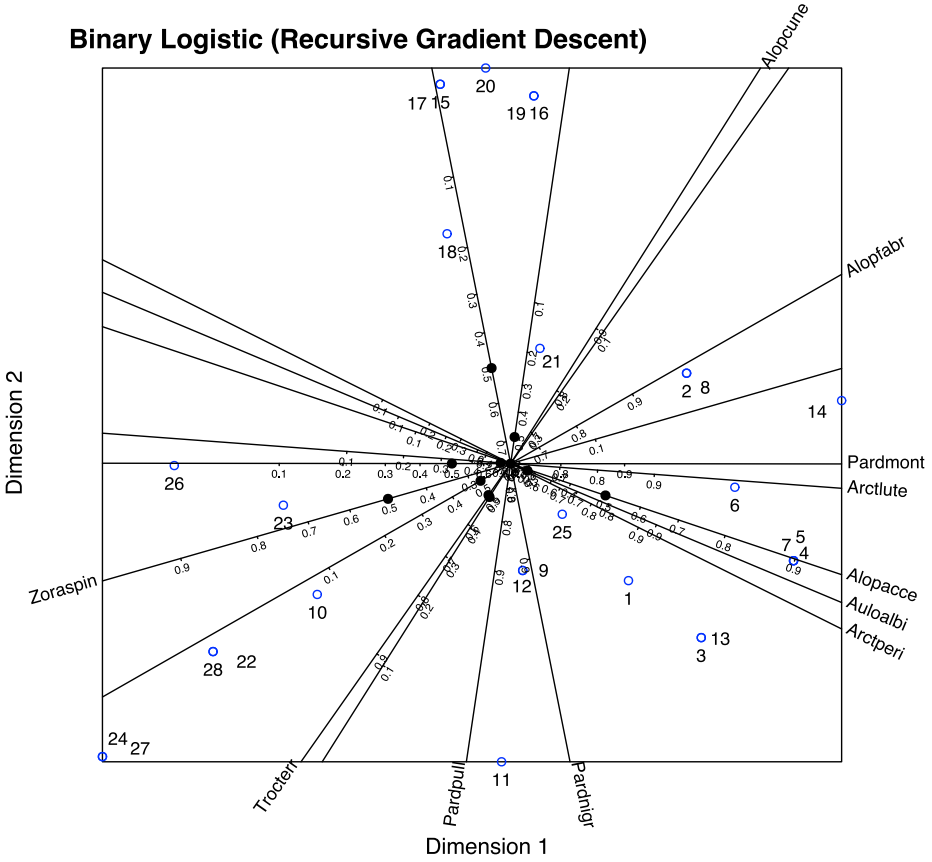

Figura 3.3: Representación de biplot logístico típico con escalas graduadas para variables binarias. Los datos usados para el biplot serán los del ejemplo de la sección 6.6

Para simplificar la representación, a veces usamos la versión reducida de las escalas de predicción situando una flecha desde el punto de predicción 0,5 hasta el punto de predicción 0,75. Proporciona la dirección de incremento de las probabilidades e información sobre la discriminación; las flechas cortas generalmente indican mayor poder de discriminación para explicar la variable representada. Una representación típica del biplot logístico con flechas, es posible observarla en la figura 3.4.

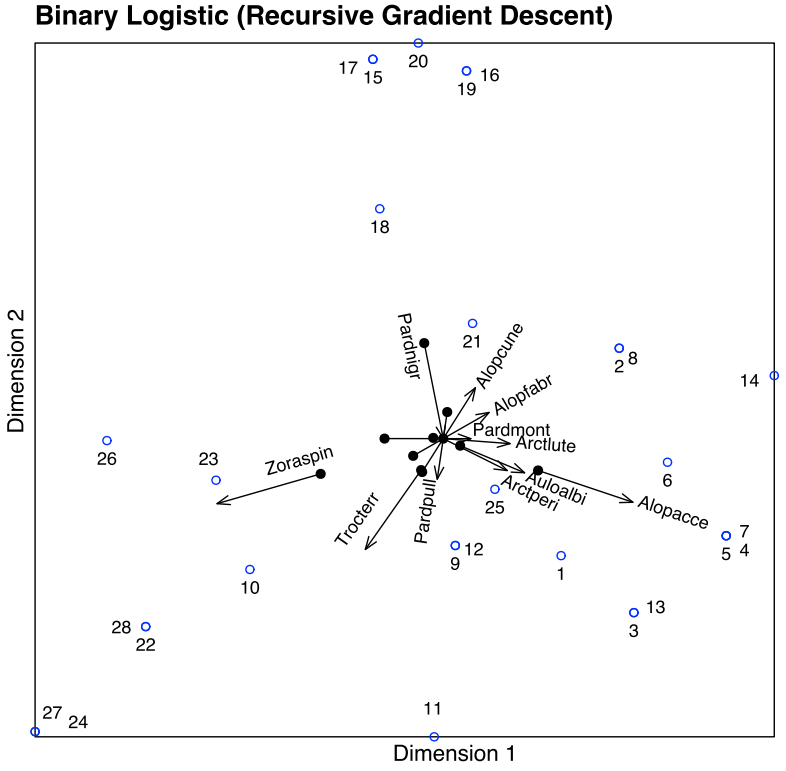

Figura 3.4: Representación del biplot logístico típico con flechas para variables binarias. Los datos usados para este biplot es la matriz de datos binarios utilizados en los ejemplos de capítulos posteriores (5.6 y 6.6)

Necesitamos una medida de predictividad (o calidad de representación) para cada variable binaria. Considerando que en el caso continuo usamos el recuento de la variabilidad de cada variable explicada por las dimensiones (el R^2 como las medidas de Cox-Snell o Nagelkerke). Es fácil ver si fijamos las coordenadas de la fila $\mathbf{a}$ y la variable j, la ecuación (3.15) define una Regresión Logística para esa variable.

Usaremos las probabilidades esperadas en la ecuación (3.15) para las predicciones en el caso habitual, es decir, predecimos presencia $\left(\hat{x}_{ij} = 1\right)$ si $\pi_{ij} \leq 0,5$ y ausencia si

$\left(\hat{x}_{ij} = 0\right)$. De otra forma, obtenemos una matriz binaria esperada $\hat{\mathbf{X}} = \left(\hat{x}_{ij}\right)$. Una medida de bondad de ajuste general podría ser el porcentaje de correctamente clasificados (predicciones). Calculando los porcentajes para cada fila o columna, tenemos medidas separadas de la calidad de individuos y variables.

3.4. Software biplot

Para que las técnicas descritas a lo largo del capítulo sean útiles, es necesario que exista al menos un software que permita realizar tanto los cálculos como las representaciones gráficas. El biplot es un técnica que, debido a su gran utilidad, su uso es cada vez mayor y por ello existen diversos paquetes y/o programas que han incluido una sección dedicada al biplot o que se han desarrollado específicamente para ello. En esta sección recogeremos algunos de los más importantes y que están siendo más utilizados en la actualidad.

3.4.1. Software de uso general

Dentro de los software de uso general que están ampliamente extendidos, SPSS, SAS Minitab16 o STATA contienen un apartado específico para realizar este tipo de técnicas, aunque en todos los casos contiene únicamente los biplots clásicos.

IBM SPSS Statistics (IBM Corp., 2021)

Para realizar un Análisis Biplot dentro de este software será necesario realizar en primer lugar el análisis que denominan *"Análisis de componentes principales categórico (CATPCA)"*. Este análisis se encontrará dentro del menú de *Reducción de dimensiones, Escalamiento óptimo*; dentro, deberá ser elegido el biplot en el apartado de gráficos, el *diagrama de dispersión biespacial.*

Statistical Analysis Software (SAS) (SAS Institute Inc., 2022)

Este programa contiene varias opciones para la realización de cuatro tipos de biplot, los tres biplots clásicos y el biplot de covarianza (biplot-COV). Las opciones que presenta este software son las siguientes:

- Una opción es usar el menú PROC PRINQUAL del software SAS/STAT para crear el biplot COV.
- Con la licencia para el software SAS/GRAPH y para el software SAS/IML es posible usar la macro %BIPLOT de Michael Friendly empleando la opción OUT = en ella. Es posible crear una versión más moderna empleando PROC SGPLOT.
- Otra opción es realizar los cálculos matriciales con SAS/IML para obtener las coordenadas de los marcadores y vectores. A continuación, se empleará el módulo de biplot para realizar la representación gráfica.
- Por último, es posible utilizar el módulo WriteBiplot para realizar los cálculos y PROC SGPLOT para crear el biplot.

Minitab16 (LLC, 2022)

En el caso de Minitab, igual que ocurría en SPSS, debemos realizar un Análisis de Componentes Principales, y a continuación, en el menú de gráficos que se encuentra dentro del análisis anterior, será necesario seleccionar la *gráfica de doble proyección.*

STATA (StataCorp, 2021)

En el software STATA la función biplot será utilizada para hacer este tipo de representaciones gráficas. La función se encuentra en el menú de estadísticos, dentro del apartado de estadística multivariante.

3.4.2. Software: Paquetes Comerciales

Debido a que el software de uso general no permiten realizar un gran número de tipos de biplots, se han creado recientemente otros paquetes comerciales que amplían el

número de representaciones gráficas que se pueden realizar.

Algunas de estas son GGE-Biplot, MVSP, Statgraphics, PC-ORD, CANOCO y Analyse-it, que serán descritos a continuación.

GGE-Biplot (Yan y Kang, 2006)

En este software, además de los biplots clásicos, se pueden construir otros biplots como el AMMI biplot, el GGE biplot (genotipo-ambiente), biplot de expresión génica, biplot QTL de mapeo, etc.

Este software está orientado al ámbito de la Agronomía, pero puede ser utilizado en cualquier ámbito.

MultiVariate Statistical Package (Kovach, 1999)

Este software se especializa en la realización de Componentes Principales, Coordenadas Principales y Análisis de Correspondencias. Dentro de sus funciones se encuentra la posibilidad de realizar representaciones biplot con los resultados obtenidos.

Para elaborar una representación biplot en este programa será necesario representarla tras haber obtenido los cálculos del PCA o del CCA.

El campo que utiliza este software por excelencia es el ámbito de la Ecología.

Statgraphics (Inc. Statgraphics Technologies, 2020)

Este programa ha realizado recientemente nuevas incorporaciones, entre ellas la conexión con R y con Python desde su interfaz gráfica.

Incorpora un gran número de técnicas univariantes y multivariantes entre sus menús. En este caso es de interés destacar el Análisis de Componentes Principales ya que a partir de ella realizaremos las representaciones gráficas que estamos estudiando.

Para construir el biplot en este software debemos seleccionar, dentro del menú Avanzado, los Métodos Multivantes, y a continuación Componentes Principales y las opciones avanzadas de los gráficos.

Este software dará la posibilidad de realizar gráficos en 3D.

PC-ORD (James Grace y Hatch, 2018)

Uno las características de este software es que solo se ha desarrollado para el sistema operativo Windows.

Es un paquete muy amplio que contiene multitud de técnicas multivariantes, sobre todo las más utilizadas dentro del campo de la Ecología. Entre las funciones disponibles dentro del paquete se encuentra un menú para realizar diferentes tipos de biplots. Este software también incluye la posibilidad de realizar GGE biplots.

CANOCO (Šmilauer, 2012)

Igual que en el caso anterior, este programa solo se encuentra disponible para el sistema operativo Windows.

CANOCO fue desarrollado para realizar el Análisis Canónico de Correspondencias, sin embargo en la actualidad contiene un gran número de análisis multivariantes que pueden realizarse, incluyendo la realización de biplots dentro de su ventana de visualización de los datos.

El campo en el que más se utiliza este software es el ámbito de la Ecología.

Analyse-it (Excel) (Ltd. Analyse-it Software, 2022)

El Software Analyse-it está implementado como un complemento de Excel, sin embargo no está disponible para las versiones de Excel de MacOs.

Este complemento aporta a las hojas de cálculo de Excel la posibilidad de realizar técnicas estadísticas más complejas, construcción de modelos, control de calidad o métodos de validación entre otras.

Las técnicas de mayor interés en este capítulo son las de Análisis Multivariante, concretamente las relacionadas con la realización de Análisis de Componentes Principales, ya que dentro de este menú se puede encontrar un menú que permite la representación biplot de las variables de estudio.

3.4.3. Software Paquetes Libres

Los paquetes comerciales, debido al costo, no siempre son las opciones óptimas para crear los biplots. Además, en muchos de ellos, están cerrados a la incorporación de nuevos métodos, y por ello, los biplots de reciente creación no aparecen tampoco en ellos.

De la misma forma que se crean los paquetes comerciales, se han ido desarrollando algunos paquetes libres que sí que contienen más alternativas, desde complementos de Excel, software propio o paquetes de R. A estos últimos les dedicaremos una sección diferenciada, la sección 3.4.4.

En esta sección describiremos los paquetes XLS-Biplot, ViSta, Brodgar y MultBiplot.

XLS-Biplot (Udina, 2005)

Se trata de un complemento para Excel en el que se incluye la construcción de este tipo de representaciones gráficas.

Para utilizarla será necesario instalar la macro correspondiente y abrir, dentro de los Complementos, la opción Biplot. A continuación, deberemos elegir la Descomposición en Valores Singulares, seleccionar los datos de análisis y sus correspondientes etiquetas, y elegir el método de cálculo y las transformaciones. Por último, aparecerá la ventana para la realización del Biplot que contiene las características del gráfico y, en el escalado, es posible elegir entre los biplots clásicos para realizar la representación.

Este software fue descrito en un artículo por Udina (2004) donde se puede encontrar más información.

ViSta The Visual Statistics System (Young, 1990)

Este software ha sido desarrollado como una forma de representar gráficamente datos creando visualizaciones dinámicas.

No se ha actualizado desde hace tiempo, sin embargo contiene un apartado que permite realizar biplots dentro de las representaciones gráficas de las Componentes Principales de los Análisis Multivariantes creados en el Software.

Brodgar (Highland Statistics Ltd., 2017)

Para el uso de este software será necesario tener instalado R ya que lo usará como compilador. Son programas separados, aunque existe una versión para R. Las funciones en ambos casos son mayoritariamente las mismas, sin embargo hay algunas más dentro del software externo.

No existen actualizaciones recientes y será un software libre hasta el año 2026, teniendo que renovar el código de licencia de forma anual.

Este paquete contiene un gran número de análisis, tanto univariantes como multivariantes, que se pueden realizar a través de su interfaz gráfica.

Por defecto, Brodgar realizará un Biplot de correlaciones, aunque es posible seleccionar el tipo de biplot que se desea representar.

MultBiplot: Multivariate Analysis using Biplots (Vicente-Villardon, 2020)

Este software está asociado a Matlab, sin embargo, a diferencia del caso anterior, no es necesario haber descargado Matlab para poder operar con él si se realiza la descarga completa. También existe una versión dentro de R y una versión integrada dentro de Matlab.

El paquete descargado, en este caso, contiene código menos actualizado y un menor número de técnicas que el paquete de R. Su interfaz gráfica está disponible tanto para el sistema operativa Windows como para MacOs.

Como en otros de los paquetes presentados hasta el momento, el paquete MultBiplot contiene un gran número de técnicas multivariantes y todas ellas presentadas en forma de Biplot, entre ellas es posible destacar el Análisis de Componentes Principales, el Análisis Factorial, el Unfolding o el STATIS-ACT.

De los programas presentados, este es el único que contiene los biplots logísticos como se han descrito en este capítulo.

3.4.4. Paquetes de R (R Core Team, 2021)

R es otro software donde podemos realizar un gran número de biplots, será el que emplearemos en los gráficos de este trabajo.

Al tratarse de un software libre, se encuentran multitud de paquetes que permiten realizar este tipo de representaciones gráficas. Los autores de cada paquete incluyen los biplots que son de interés con las técnicas que se han incluido en el mismo.

La mayoría de los paquetes presentados a continuación es posible encontrarlos en el repositorio CRAN.

stats

El paquete de estadística básico de R contiene una función ("`biplot`") que permite realizar los biplots clásicos sobre los resultados de unas Componentes Principales.

multibiplotGUI (Librero *et al.*, 2022)

Este paquete contiene una interfaz gráfica que permite realizar un biplot asociado al Análisis Factorial Multiple.

Es posible elegir el tipo de biplot utilizado entre los biplots clásicos y el HJ-biplot.

biplotbootGUI (Librero *et al.*, 2019)

En el paquete biplotbootGUI las autoras introducen métodos bootstrap a los biplots clásicos para obtener las regiones de confianza a través de una interfaz gráfica de R.

Con esta interfaz también es posible realizar los cálculos del Clustering Disjoint Bi-

plot así como el Análisis de Componentes Principales asociado a este tipo de biplot.

GGEBiplot (Dumble, 2022)

Este paquete contiene las funciones necesarias para construir y evaluar un biplot genotipo ambiente. Es posible realizar la comparación de dos genotipos en todos los ambientes sometidos a estudio, así como establecer la relación entre los ambientes, examinar por separado genotipos y ambientes, o construir un orden de cada uno de ellos.

Recientemente ha sido eliminada de CRAN la interfaz gráfica de este paquete por compatibilidad con las versiones más recientes.

ade4 (Dray *et al.*, 2022)

El paquete desarrollado por el Laboratorio de Biometría y Biología Evolutiva (UMR CNRS 5558) de la Universidad de Lyon contiene un conjunto de funciones que permiten analizar datos Ecológicos y Ambientales utilizando métodos de tres vías.

Entre la multitud de funciones que se pueden encontrar en este paquete, encontramos diversas representaciones gráficas, una de ellas son los biplots presentados en este capítulo. Utilizando la función "`scatter`" haremos una representación básica. Existen funciones dentro del paquete para realizar biplots asociados a los mapas factoriales de diferentes Análisis de Correspondencias, a un PLS con el algoritmo NIPALS o a un Análisis de Coordenadas Principales.

vegan (Oksanen *et al.*, 2017)

Este paquete también está diseñado para los análisis ecológicos. Contiene un gran número de técnicas de gran utilidad en este campo.

Este paquete contiene, igual que el anterior, varias funciones que permiten realizar representaciones biplot. Estos biplots estarán asociados fundamentalmente al Análisis Canónico de Correspondencias y al Análisis de la Redundancia.

BiplotML (Babativa-Márquez, 2020)

El paquete BiplotML, de reciente creación, será utilizado para realizar representaciones de biplots logísticos.

Aunque en este documento se ha descrito la forma tradicional de calcular los parámetros, este paquete implementa nuevas metodologías que optimizan la estimación de los parámetros del modelo y construye los Biplot Logísticos a partir de ellos.

MultBiplotR (Vicente-Villardon, 2021)

Como ya habíamos mencionado anteriormente, este paquete es una versión mejorada de la interfaz gráfica desarrollada en Matlab y contiene un mayor número de recursos que el paquete libre de partida.

Este paquete contiene varias técnicas multivariantes desde una perspectiva de biplot. Son muchos los análisis que se incluyen en este paquete, desde innovaciones hasta técnicas clásicas. Algunos de ellos son los Biplots clásicos, HJ-Biplot, Biplots canónicos, MANOVA Biplots, Análisis de Correspondencia, Análisis de Correspondencia Canónico, STATIS-ACT canónico, Biplots logísticos para datos binarios y ordinales, Despliegue multidimensional, Biplots externos para el Análisis de Coordenadas Principales o el Escalado Multidimensional, entre muchos otros.

3.5. Ejemplo biplot logístico

A lo largo de los capítulos restantes, como ya se indicaba en el titulo del documento "*Generalización del biplot logístico para dos o más matrices de datos*", el biplot logístico va a tener especial relevancia. Por ello, a pesar de que los contenidos desarrollados en este capítulo no contienen nuevas contribuciones, en esta sección se va a incluir un ejemplo que ilustre los conceptos desarrollados a lo largo del mismo sobre este tipo de representaciones gráficas.

3.5.1. ¿Hay diferencias de género en los cambios provocados por la pandemia?

Para las representaciones biplot presentaremos un ejemplo sobre una encuesta del CIS de 2021 para estudiar la existencia de diferencias de género en los cambios provocados por la pandemia.

En la actualidad, y desde hace ya algún tiempo, se ha destacado la discriminación de género en el ámbito laboral y han sido muchos los autores que durante los últimos años han profundizado sobre el tema, que en muchos casos se ha agravado debido a la pandemia del Coronavirus. Estos estudios se han realizado en diversos campos y cabe destacar que en el ámbito de la estadística y el análisis de datos existen un gran número de mujeres con vocaciones tempranas en este tipo de ámbitos.

En este ejemplo proponemos el análisis de la situación actual, provocada por la pandemia del COVID-19, en la vida diaria de las personas españolas que han cumplido la mayoría de edad.

Base de datos

Utilizaremos una encuesta del CIS (Centro de Investigaciones Sociológicas) llamada "Efectos y consecuencias del coronavirus". Inicialmente se planteo utilizar la encuesta número III realizada entre el 11 y el 16 de diciembre de 2020, pero finalmente se ha decidido utilizar datos más recientes, los resultados de la encuesta número IV realizada entre el 14 y el 29 de mayo de 2021.

Se ha encuestado a 3200 personas mayores de edad de toda España, utilizando un muestreo estratificado en función del número de habitantes de las comunidades y ciudades autónomas obteniendo una muestra final de 3008 individuos. La encuesta se realizó de manera telefónica a todos ellos.

Se obtienen 189 variables de resultados reales y más de 80 variables añadidas por el encuestados por fallos y omisiones detectadas durante la encuesta. Por ello, vamos a elegir un subconjunto de variables para presentar en el ejemplo, cabe destacar que es un estudio preliminar en el que se continuará investigando, como observaremos en capítulos sucesivos.

Las variables seleccionadas son de tipo binario, es decir, presencia o ausencia de la realización de un evento. En este caso han sido 5 correspondientes a los hábitos de las personas entrevistadas sobre ellas mismas, y otros 8 hábitos en relación con las personas mayores que tienen en su familia y con las que no conviven. Se han medido dos momentos diferentes, aunque se han realizado en la misma encuesta, antes de la pandemia y desde que tenemos la pandemia.

Las variables que se utilizarán serán:

ComidasFestivos ¿Habitualmente solía/suele Usted comer o cenar en días festivos con familiares?

Cumpleaños ¿Habitualmente solía/suele Usted asistir a cumpleaños o "santos" de fa-

miliares?

Eventos ¿Habitualmente solía/suele Usted participar en otras celebraciones familiares como comuniones, bodas o similares?

Ocio ¿Habitualmente solía/suele Usted asistir a actividades culturales, deportivas y de ocio?

Videollamadas ¿Habitualmente solía/suele Usted comunicarse por videollamada con más frecuencia que antes?

VisitasM Con los parientes mayores con los que no convive (al menos uno de ellos) ¿Habitualmente solía/suele Usted ir a visitarles?

CompraM Con los parientes mayores con los que no convive (al menos uno de ellos) ¿Habitualmente solía/suele Usted hacerles la compra?

GestionesM Con los parientes mayores con los que no convive (al menos uno de ellos) ¿Habitualmente solía/suele Usted acompañarles a hacer gestiones?

TareasM Con los parientes mayores con los que no convive (al menos uno de ellos) ¿Habitualmente solía/suele Usted hacerles tareas domésticas (limpiar, hacer la comida)?

OcioM Con los parientes mayores con los que no convive (al menos uno de ellos) ¿Habitualmente solía/suele Usted salir con ellos a actividades de entretenimiento, ocio o similares?

MedicoM Con los parientes mayores con los que no convive (al menos uno de ellos) ¿Habitualmente solía/suele Usted acompañarles al médico, pruebas o análisis?

TelefonoM Con los parientes mayores con los que no convive (al menos uno de ellos) ¿Habitualmente solía/suele Usted hablar regularmente por teléfono?

VideollamadasM Con los parientes mayores con los que no convive (al menos uno de ellos) ¿Habitualmente solía/suele Usted comunicarse con ellos regularmente por videollamada?

Las variables correspondientes a los datos previos a la pandemia estarán precedidas por una "A_", mientras que los del momento en el que se realiza la encuesta que diremos que son los posteriores serán precedidas por una "D_".

La encuesta completa que se ha realizado se encuentra en el Anexo I.

Se presentarán solo los resultados de los biplots de interés, aunque se han aplicado algunas de las técnicas expuestas en los capítulos posteriores con resultados satisfactorios.

Objetivos del ejemplo

Con este ejemplo se plantea:

Objetivo 1. Estudiar la relación entre las variables medidas antes y después de la pandemia para todos los casos de forma conjunta.

- **1.1.** Analizar la relación entre las variables asociadas con los hábitos personales.
- **1.2.** Presentar la relación entre las variables de los hábitos con las personas mayores con las que no convive el encuestado.
- **1.3.** Identificar relaciones generales entre todas las variables.

Objetivo 2. Exponer la relación de las variables medidas antes y después de la pandemia para las mujeres del estudio.

- **2.1.** Analizar la relación entre las variables asociadas con los hábitos personales.
- **2.2.** Presentar la relación entre las variables de los hábitos con las personas mayores con las que no convive el encuestado.
- **2.3.** Identificar relaciones generales entre todas las variables.

Objetivo 3. Estudiar las respuestas previas y posteriores a la pandemia de los varones en las variables seleccionadas de la encuesta.

3.1. Analizar la relación entre las variables asociadas con los hábitos personales.

3.2. Presentar la relación entre las variables de los hábitos con las personas mayores con las que no convive el encuestado.

3.3. Identificar relaciones generales entre todas las variables.

Objetivo 4. Comparar el comportamiento de hombres y mujeres en función de las variables seleccionadas para el análisis de los hábitos antes y después de la pandemia.

4.1. Analizar la relación entre las variables asociadas con los hábitos personales.

4.2. Presentar la relación entre las variables de los hábitos y de las personas mayores con las que no convive el encuestado.

Metodología

Emplearemos tres biplots logísticos calculados a través del método del gradiente.

- En el primero utilizaremos todos los individuos disponibles, emplearemos como constante de la penalización Ridge 0,5.
- En el segundo usaremos únicamente las mujeres del estudio, utilizaremos 1 para la penalización Ridge.
- Y en el tercero, y último, solo los hombres con una penalización Ridge de 2.

Todos los análisis serán realizados con el software estadístico R (R Core Team, 2021), con el paquete MultBiplotR (Vicente-Villardon, 2021).

Resultados

En primer lugar, se ha realizado el biplot logístico utilizando el método del descenso del gradiente para todos los individuos de forma conjunta. Este gráfico se encuentra en la figura 3.5, las variables previas a la pandemia están representadas en color aguamarina mientras las posteriores se han reflejado en color salmón.

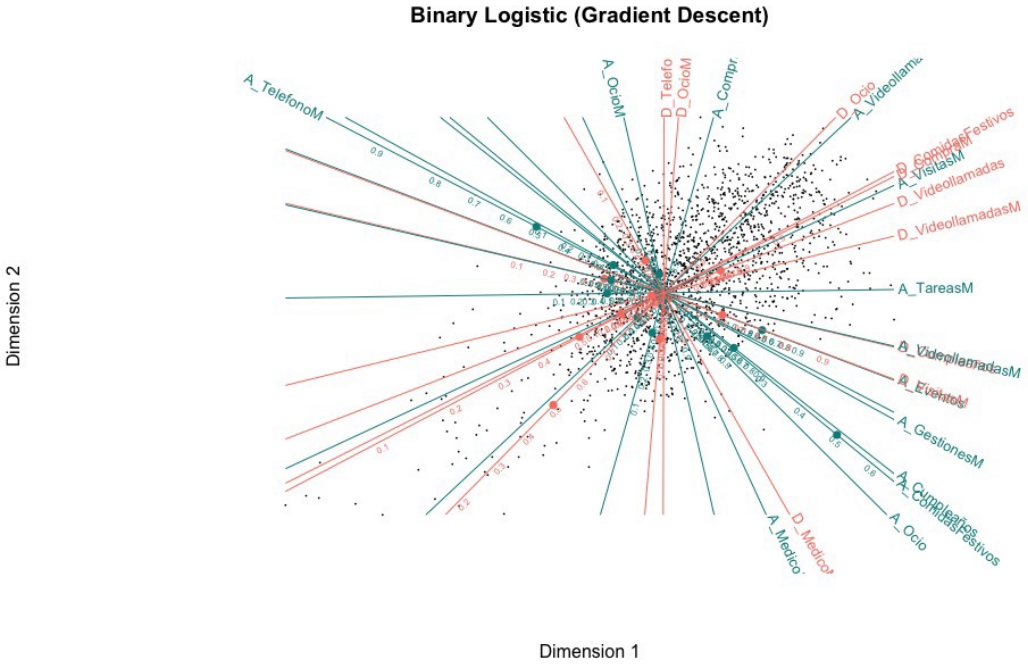

Figura 3.5: Biplot Logístico para los hábitos antes y después de la pandemia

Para el caso general, podemos destacar que, en cuanto a los hábitos con ellos mismos, antes de la pandemia las respuestas sobre acudir a las comidas en días festivos con familiares, los cumpleaños y el ocio estaban muy relacionadas. La asistencia a eventos podría incluirse en este grupo, aunque la relación es un poco menor. La realización de videollamadas con frecuencia es independiente al resto de las variables de este grupo.

Después de la pandemia, la asistencia a cumpleaños es practicamente independiente de la participación en actividades culturales, deportivas y de ocio. Este tipo de actividades está más relacionado con las comidas en días festivos y las videollamadas, que están

estrechamente relacionadas entre ellas.

El ocio antes y después de la pandemia no se encuentran relacionados, al igual que las comidas en días festivos. Sin embargo, la realización de videollamadas y los cumpleaños sí que está relacionados, aunque las relaciones no son tan fuertes como habíamos visto en los parrafos anteriores.

En cuanto a los hábitos en relación con los mayores que no residen en el mismo domicilio que los encuestados, a nivel general, antes de la pandemia las llamadas telefónicas y la ayuda en sus gestiones a los mayores se encuentran inversamente relacionadas. La relación de estas llamadas también está inversamente relacionada con las videollamadas y la realización de tareas para los mayores. Estas llamadas, además, no tienen relación con si se acompaña a los mayores o no a realizar las compras necesarias. Esta última variable tampoco está relacionada con las videollamadas a los mayores, las gestiones o las tareas realizadas. Sin embargo, la variable de las visitas está relacionada tanto con la variable de hacerles la compra como la de hacerles las tareas domésticas, y no está relacionada con si acompañan a los mayores a las consultas medicas. El asistir a actividades de entretenimiento u ocio con los mayores se encuentra inversamente relacionado con la asistencia a las consultas médicas y directamente relacionado con hacer la compra a los mayores.

En el momento en que se realizó la encuesta, a nivel general, se observa que las llamadas telefónicas y el ocio con los mayores están estrecha y directamente relacionados. Sin embargo, ambas tienen escasa relación con la realización de videollamadas o visitas a los mayores, y tienen una relación inversa con la asistencia a consultas médicas. Estas visitas al médico o pruebas relacionadas son independientes de la realización de las compras para los mayores y prácticamente independiente de si se realizan videollamadas con ellos; sí que tienen relación con las realización de visitas a los familiares mayores.

En la relación entre antes y después observamos que las visitas al médico están muy

relacionadas en ambos casos. Casi todas las variables tienen una relación entre el antes y el después, excepto las llamadas telefónicas y las visitas que son prácticamente independientes.

En cuanto a la relación entre los hábitos personales y los hábitos con las personas mayores que no conviven con los encuestados, es posible destacar que, antes de la pandemia, el ocio, cumpleaños y comidas familiares estaban relacionadas directamente con la ayuda a las gestiones de los mayores y acompañarles al médico. Estas gestiones están también relacionadas con la asistencia a eventos, las videollamadas a los mayores y la realización de tareas domésticas, pero no existe relación con las compras o el uso general de las videollamas. Las llamadas de teléfono también son independientes de estas dos variables.

Después de la pandemia, la realización de la compra para los mayores está estrechamente relacionada con la realización de comidas familiares, el uso general de las videollamadas y las videollamadas con los mayores, también tienen relación con las llamadas telefónicas y el ocio, tanto con mayores como sin ellos, pero son independientes de las visitas médicas.

La participación en eventos antes de la pandemia está estrechamente relacionada con las visitas posteriores a los mayores y los cumpleaños tras la pandemia y con las videollamadas con los mayores previas a esta. Las visitas a los mayores previas a la pandemia también tienen una gran relación con el uso de las videollamadas y las comidas familiares posteriores a la misma.

Estudiaremos a continuación con más detalle el caso de las mujeres que se recoge en la figura 3.6, seguirá el mismo código de colores que el caso anterior.

Antes de la pandemia, en el caso de las mujeres, dentro de sus hábitos personales destaca que el uso de las videollamadas tiene una relación directa con las comidas en los

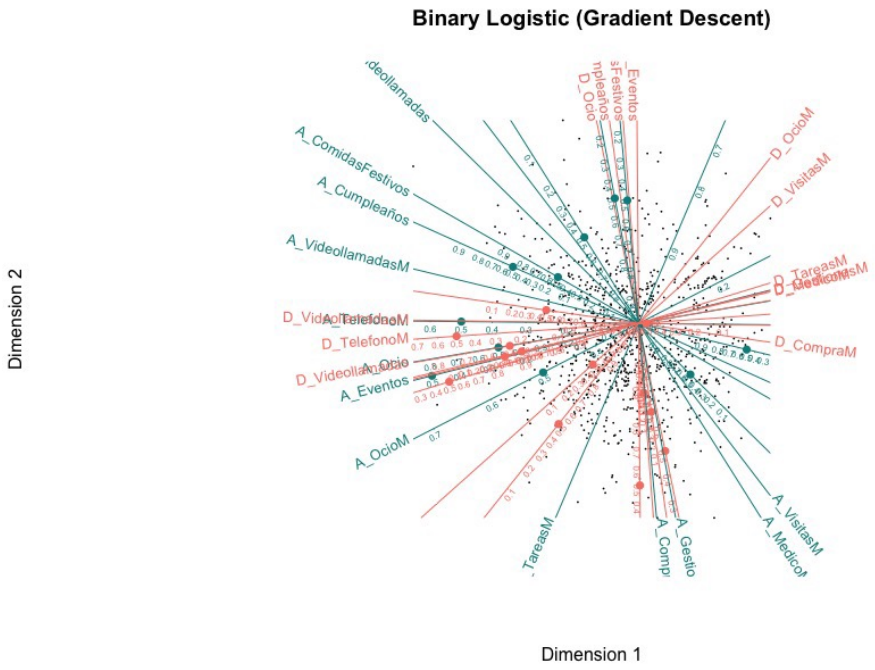

Figura 3.6: Biplot Logístico para los hábitos antes y después de la pandemia en el caso de las mujeres

días festivos y los cumpleaños. El ocio y los eventos también se encuentran altamente relacionados.

Después de la pandemia, todos los hábitos personales están muy relacionados, asistencia a eventos, cumpleaños, comidas familiares y actividades de ocio, excepto el uso de videollamadas que son independientes a todas ellas.

En el caso de las mujeres, aunque la relación no es muy fuerte, la asistencia a cumpleaños y comidas familiares y el uso de las videollamadas están directamente relacionadas antes y después de la pandemia. Sin embargo, el participar en eventos familiares y en actividades de ocio son independientes en los dos momentos que se han incluido en la encuesta.

En el biplot logístico para el caso de las mujeres, observamos que, en los hábitos previos a la pademia, con las personas mayores que no residen con las entrevistadas las

visitas a estas personas y el acompañarlas al médico están estrechamente relacionadas, así como el realizarles la compra y ayudarles con las gestiones, también se encuentran relacionadas entre ellas, aunque la relación es mas débil. Es independiente de todas ellas el realizar actividades de ocio con los mayores. A diferencia del caso general, la realización de las compras y las tareas domésticas sí que están relacionadas.

Tras las pandemia, las tareas, las gestiones y el acompañar a las personas mayores al médico están íntimamente relacionados de forma directa, también se establece una relación directa con el ocio con mayores, las visitas a estos y el realizarles la compra. La relación del uso de videollamadas con los mayores será inversa, y con gran importancia, con el resto de las variables mencionadas.

Las variables de ocio antes y después de la pandemia están inversamente relacionadas, el realizar las tareas domesticas a los mayores tienen una pequeña relación inversa y la realización de la compra a los mayores tiene una pequeña relación directa en los dos momentos. Las visitas al médico, la visita a los mayores y la realización de las gestiones son independientes antes y después de la pandemia.

Antes de la pandemia, las variables de uso de videollamadas, de comidas en días festivos y de cumpleaños están inversamente relacionadas con las visitas a los mayores y acompañarles a las consultas medicas, y son independientes de la realización de las tareas domesticas. La participación en actividades de ocio o eventos familiares es independiente de las visitas realizadas y de acompañarles al médico.

En el momento de realización de la encuesta, el uso general de las videollamadas está inversamente relacionado con la realización de las tareas domesticas a los mayores, el acompañarles a las visitas médicas y el ayudarles en la realización de gestiones, sin embargo, está estrecha y directamente relacionado con las llamadas telefónicas y videollamadas con los mayores. La participación en eventos familiares, cumpleaños, actividades de ocio y comidas en días festivos está relacionada de forma directa con la participación

en actividades de ocio con las personas mayores y con realizarles visitas, pero son independientes de las llamadas y videollamadas con los mayores, la realización de las tareas domésticas de estos, acompañarles al médico, ayudarles con las gestiones y hacerles la compra.

En la relación entre antes de la pandemia y después de esta se puede observar que, tanto la participación en actividades de ocio como la asistencia eventos familiares, son independientes. La relación entre la participación en cumpleaños, comidas en días festivos y el uso de las videollamadas antes y después de la pandemia es directa.

En los hábitos con las personas mayores podemos destacar que las videollamadas y las llamadas telefónicas antes de la pandemia, en el caso de las mujeres, están directamente relacionadas con las posteriores a esta. El prestar ayuda para la realización de las gestiones, acompañar a los mayores al médico, realizarles visitas y hacerles la compra antes y después de la pandemia son independientes. Además, la realización de las tareas domésticas en los dos momentos que se han preguntado están inversamente relacionadas.

El uso habitual de videollamadas antes de la pandemia está muy relacionado con la participación en cumpleaños, comidas en días festivos, eventos familiares y actividades de ocio en el momento en el que se realizó la encuesta, pero son independientes del ocio con los mayores y las visitas a estas personas. Sin embargo, las videollamadas a los mayores antes de la pandemia son independientes de la participación en cumpleaños, comidas en días festivos, eventos familiares y actividades de ocio en el momento en el que se realizó la encuesta. Además, podemos destacar que aquellas mujeres que antes de la pandemia realizaban la compra o las gestiones con los mayores están inversamente relacionadas con las situaciones planteadas anteriormente (cumpleaños, ocio, comidas en festivos y eventos), y que las que participaban en eventos familiares y actividades de ocio antes de la pandemia se relacionan de forma inversa con aquellas que en el momento en el que se realizó la encuesta realizan las tareas domesticas, las gestiones y acompañan a

las personas mayores a las consultas médicas.

Para finalizar, en el gráfico 3.7, analizaremos el caso de los hombres del estudio manteniendo los mismos colores que en los casos anteriores.

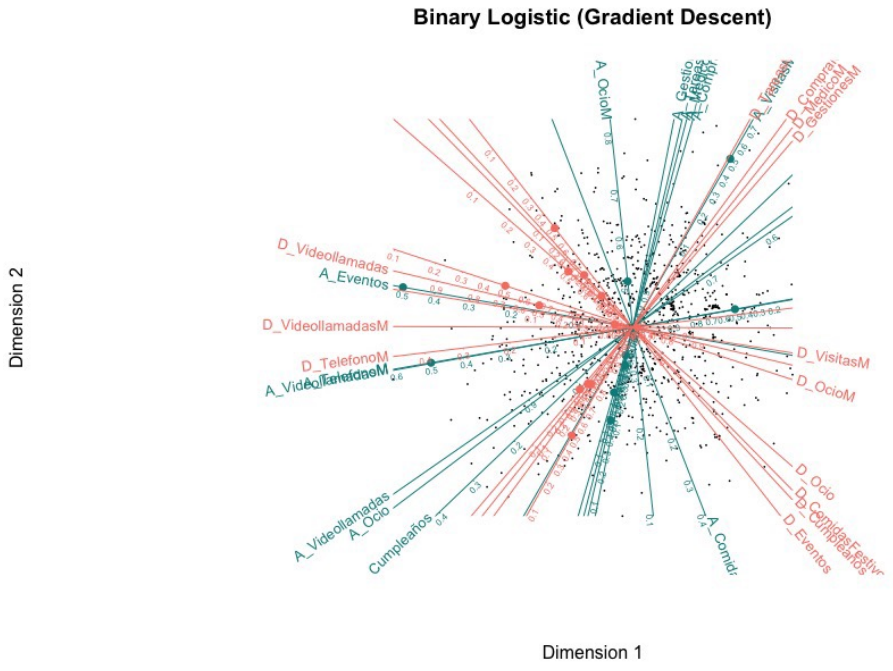

Figura 3.7: Biplot Logístico para los hábitos antes y después de la pandemia en el caso de los hombres

Para los hombres, entre sus hábitos personales destaca una relación directa y bastante fuerte entre el uso habitual de videollamadas, la participación en actividades de ocio y en cumpleaños, y, aunque la relación es menor, con la asistencia a eventos familiares. La participación en comidas en los días festivos es independiente del uso habitual de videollamadas, tiene una relación directa, aunque débil, con la participación en cumpleaños y una relación inversa con la asistencia a eventos familiares.

En cuanto a los hábitos personales de los hombres en el momento en el que se realizó la encuesta, se puede observar que la participación en cumpleaños, comidas en días festivos, eventos familiares y actividades de ocio están muy relacionadas entre ellas de

forma directa, e inversamente relacionadas con la realización habitual de videollamadas.

La asistencia de comidas en días festivos antes de la pandemia se encuentra muy relacionado con su participación en cumpleaños, comidas en días festivos, eventos familiares y actividades de ocio después de la pandemia. Sin embargo, estas situaciones son independientes de la realización de videollamadas habitualmente, la asistencia a cumpleaños y actividades de ocio antes de la pandemia. Otra de las relaciones que destacan en este gráfico es que, en los hombres, la asistencia a eventos familiares antes de la pandemia está muy relacionada, de forma directa, con el uso de las videollamadas regularmente después de la pandemia.

En cuanto a los hábitos de los hombres con los mayores que no viven en su mismo domicilio, antes de la pandemia, destaca que las llamadas y las videollamadas a los mayores están directa y estrechamente relacionadas y son independientes de las actividades de ocio realizadas con los mayores. La variable de ocio con los mayores está muy relacionada con las visitas y la asistencia a las comidas familiares, las consultas médicas, acompañarles en la realización de gestiones y la ayuda en las tareas domesticas de forma directa. Además, estas cuatro últimas variables mencionadas tiene una relación directa y muy fuerte entre ellas, y están relacionadas de forma inversa y débil con las llamadas telefónicas y las videollamadas.

Las visitas a los mayores y el ocio con ellos después de la pandemia pueden formar un grupo de variables por la relación directa existente entre ellas, y las videollamadas y las llamadas telefónicas otro por el mismo motivo, entre estos dos grupos se establece una relación fuerte e inversa. Con el resto de los hábitos con los mayores, que tienen una fuerte relación directa entre ellos, los dos grupos creados antes tiene relaciones que no son muy fuertes, con el primero se relacionan de forma directa y, por lo tanto, con el segundo de forma inversa.

La relación de las llamadas y las videollamadas antes y después de la pandemia están

directa e íntimamente relacionadas, también se mantienen relaciones fuertes y directas entre las gestiones, las tareas domésticas, las consultas médicas y el realizar la compra para los mayores antes y después de la pandemia, sin embargo, en el caso de los hombres, el ocio con los mayores tiene una relación no muy fuerte e inversa en los dos momentos de estudio y en las visitas a estas personas una relación directa que tampoco es demasiado fuerte. Cabe destacar que la realización de las gestiones, tareas domésticas, la compra y la asistencia a consultas médicas antes de la pandemia es independiente del ocio y las visitas a los mayores después de la pandemia.

Si estudiamos la relación entre los hábitos personales y los hábitos con las personas mayores, para el caso de los hombres encontramos que la asistencia a eventos familiares (bodas, bautizos y comuniones) es independiente de la realización de las gestiones, las tareas domésticas, la compra y la asistencia a consultas médicas con los mayores antes de la pandemia, y su relación con el ocio con las personas mayores y las visitas a estas tienen una relación muy débil, directa en el primer caso e inversa en el segundo. Además observamos que antes de la pandemia las videollamadas habituales, las actividades de ocio personal y los cumpleaños están inversa y estrechamente relacionados con las visitas a los mayores, y que la realización de actividades de ocio con los mayores y las comidas en días festivos se relacionan de esta misma forma.

En el momento en el que se realizó la encuesta, el uso habitual de videollamadas está muy relacionado, en el caso de los hombres, con la realización de videollamadas y llamadas telefónicas a los mayores, y por lo tanto, igual que ocurría en los hábitos con las personas mayores, inversamente relacionadas con las visitas y el ocio con mayores. Se observa también que, después de la pandemia, el ocio y las visitas a los mayores están relacionadas de forma directa con el ocio personal, las comidas en días festivos, los cumpleaños y los eventos familiares. Estos hábitos personales, además, se observa que no están relacionados con las tareas domésticas realizadas para los mayores, así como la realización de la compra, las consultas médicas, o la ayuda a los mayores para la realización de gestiones.

Por último, podemos destacar que el uso habitual de videollamadas y la participación en actividades de ocio y en cumpleños están fuertemente relacionadas de forma inversa con la realización de la compra a los mayores, acompañarles a consultas médicas, ayudarles en la realización de gestiones o hacerles las tareas domésticas. También es posible destacar que existe una relación inversa entre la asistencia a eventos antes de la pandemia y la realización de visitas o actividades de ocio con los mayores en el momento de realización de la encuesta.

Conclusiones del ejemplo

- Los hábitos personales antes de la pandemia tenían un comportamiento similar, mientras que se dispersan después de la pandemia.
- En los hábitos con las personas mayores ocurre al contrario que con los hábitos personales, estaban dispersos mientras que tras la pandemia se agrupan de dos en dos, las llamadas telefónicas con el ocio, la realización de la compra con las videollamadas y las visitas con las consultas médicas.
- Se observa que, a nivel general, las variables de acompañar a las consultas médicas a los mayores antes y después de la pandemia están íntimamente relacionadas de forma directa. Y la variable asociada a las llamadas telefónicas antes de la pandemia presenta un comportamiento opuesto al del resto de las variables en casi todos los casos.
- Los hábitos personales de las mujeres antes y después de la pandemía tienen una relación directa, aunque cabe destacar que la relación es más estrecha en los hábitos posteriores a la pandemia que los anteriores.
- En los hábitos con los mayores que no residen en el mismo domicilio que las mujeres del estudio, destaca que es independiente si los realizaban antes de la pandemia para realizarlos en el momento en el que se realizó la encuesta. Las llamadas tele-

fónicas y las videollamadas a los mayores tienen un comportamiento diferente al resto de hábitos, su relación con ellos es fuerte, pero inversa, y están muy relacionadas antes y después de la pandemia.

- A nivel general destaca que, antes de la pandemia, los hábitos personales de las mujeres estaban inversamente relacionados de los hábitos con sus mayores. Sin embargo, en el momento de la realización de la encuesta son independientes los hábitos personales y la mayor parte de los hábitos con los mayores. Existe una relación inversa entre las variables que miden la participación en eventos familiares y actividades de ocio antes de la pandemia con la mayor parte de los hábitos con los mayores tras la pandemia.

- En el caso de los hábitos personales de los hombres antes de la pandemia se observa que están muy relacionadas de forma directa las videollamadas, las actividades de ocio y la asistencia a cumpleaños; la relación con el resto de las actividades es menor, pero también directa. También cabe destacar que están inversamente relacionadas la asistencia a eventos familiares y a comidas en días festivos. Tras la pandemia los hábitos están muy relacionados de forma directa entre sí y con las comidas en días festivos antes de la pandemia, excepto las videollamadas que se comporta de forma opuesta, y está muy relacionada con los eventos familiares antes de la pandemia. Todos los hábitos después de la pandemia son independientes de las videollamadas, las actividades de ocio y los cumpleaños antes de la pandemia.

- Las videollamadas y llamadas telefónicas antes y después de la pandemia, igual que en el caso de las mujeres, están muy relacionadas de forma directa, tanto antes como después de la pandemia y entre ellas, excepto las videollamadas y llamadas telefónicas que en ambos casos son independientes. El comportamiento de los hombres en cuanto a las gestiones, las compras y las consultas médicas es igual antes y después de la pandemia, sin embargo con las visitas y el ocio con los mayores es independiente antes y después de esta.

- La realización de las tareas, las compras, las gestiones y las consultas médicas con

los mayores después de la pandemia están inversamente relacionadas con el uso de las videollamadas, las actividades de ocio y los cumpleaños de los hombres antes de la pandemia.

- En los hábitos personales vamos a destacar que en el biplot de las mujeres están muy relacionadas las variables de la realización de actividades de ocio antes de la pandemia y las videollamadas durante la misma, mientras que en los hombres la relación es débil.

- La relación de las variables de tareas con los mayores antes de la pandemia muestran que los hombres que hacen una las hacen todas, mientras que en el caso de las mujeres, aunque también tienen una relación directa, no es tan fuerte. Las visitas después de la pandemia son independientes de las tareas que se hicieran antes de la pandemia tanto en hombres como en mujeres, pero sí que es importante destacar que todas las demás variables de tareas realizadas para los mayores en el caso de las mujeres también son independientes o no tienen una relación muy fuerte, mientras que en el caso de los hombres están muy relacionadas. Además, es llamativo, también en el caso de los hombres, que las personas que antes hacían las visitas ahora realizan las tareas del hogar. Cabe destacar que las llamadas telefónicas antes de la pandemia están muy estrechamente relacionadas con las videollamadas con los mayores desde el inicio de la pandemia en el caso de las mujeres, mientras que en el caso de los hombres, aunque están relacionadas de forma directa, la relación es más débil. Podríamos concluir que las atenciones a los mayores tanto en hombres como en mujeres han aumentado desde la pandemia, pero no de la misma forma.

Capítulo 4

MANOVA basado en distancias

Notación

I : Número de individuos.

J : Número de variables predictoras.

Q : Número de variables respuesta.

K : Número de grupos en los que se dividen los individuos.

$\mathbf{X}_{(I\times J)}$: Matriz de variables explicativas o matriz de diseño con I individuos y J variables.

$\mathbf{Y}_{(I\times Q)}$: Matriz de variables respuesta con I individuos y Q variables.

$\mathbf{B}_{(J\times Q)}$: Matriz que contiene los parámetros de regresión desconocidos.

$\mathbf{0}$: Vector de ceros.

$\mathbf{\Sigma}$: Matriz de covarianzas común.

$\hat{\mathbf{B}}_{(J\times Q)}$: Matriz de parámetros de la regresión estimados.

$\mathbf{C}_{(V\times J)}$: Matriz de rango V de combinaciones lineales de las columnas de X con V filas y J columnas.

$\mathbf{M}_{(Q\times W)}$: Matriz de rango W de combinaciones lineales de las columnas de Y con Q filas y W columnas.

$\mathbf{\Omega}$: Estimador de varianza mínima.

$\mathbf{R}_{(V\times V)}$: Matriz relacionada con la inversa de la matriz de covarianzas entre predictores.

$\mathbf{E}_{(W\times W)}$: Matriz de covarianzas y productos cruzados del error.

$\mathbf{H}_{(W\times W)}$: Matriz de covarianzas y productos cruzados asociados a la hipótesis nula.

$\mathbf{a}$: Vector no nulo para los contrastes de hipótesis.

α : Nivel de significación.

λ : Valores propios.

$\boldsymbol{\eta}$: Predictores lineales relacionados con la matriz $\mathbf{X}$.

L : Función de enlace.

$\delta_{ii'}$: Distancia entre el individuo i y el individuo i'.

$s_{ii'}$: Similaridad entre el individuo i y el individuo i'.

w_j : Ponderación para la variable j.

$\mathbf{D}_K$: Matriz diagonal que contiene el tamaño muestral de cada uno de los grupos.

$\bar{\mathbf{Z}}$: Coordenadas canónicas de las medias.

$\mathbf{Z}$: Coordenadas de los individuos en el espacio canónico.

$\mathbf{S}$: Matriz de covarianzas dentro de los grupos.

$\boldsymbol{\Delta}$: Matriz de distancias entre individuos.

$\mathbf{P}_{(I\times I)}$: Matriz de pertenencia a los grupos.

$\bar{\boldsymbol{\Delta}}$: Matriz de distancias dentro de los grupos.

$\mathbf{G}$: Matriz de productos escalares.

B : Número de replicas bootstrap.

4.1. Introducción

En la primera de las técnicas que se van a estudiar en este trabajo, vamos a disponer de dos matrices de datos, una de ellas contiene un conjunto de variables respuesta y la otra un conjunto de predictores, normalmente variables ficticias que describen la pertenencia de los individuos a diferentes grupos, es decir, tenemos una matriz de predictores binarios y una de respuestas continuas, aunque, como veremos luego, esto puede extenderse a otros tipos. En este contexto es importante, no solamente determinar la forma de la relación entre ambos conjuntos, sino también la posible significación de las diferencias entre grupos basándose en la partición de la variabilidad total de los datos en dos partes,

una debida a las diferencias entre los grupos y otra debida a las diferencias dentro de los grupos.

Cuando se trata de realizar la comparación de más de dos grupos, está ampliamente extendido el uso del Análisis de la Varianza (ANOVA) o sus equivalentes no paramétricos para cada una de las variables por separado. Sin embargo, se trata de una técnica univariante que no tiene en cuenta las relaciones entre las variables, pudiendo provocar inferencias erróneas; solo si todas las comparaciones son independientes es posible controlar el riesgo Tipo I y si ninguna de las variables es significativa, es posible que existan combinaciones de ellas que sí lo sean. En la actualidad, como ya se ha comentado en la introducción, existen un gran número de bases de datos que contienen muchas variables. Para evitar estos problemas se emplean las técnicas multivariantes, la técnica más extendida es el Análisis Multivariante de la Varianza (MANOVA). Ambas técnicas se basan en el Modelo Lineal General (Univariante o Multivariante).

Para poder realizar una correcta aplicación del MANOVA es necesario que los datos sigan una distribución normal multivariante, las matrices de covarianzas deben ser iguales y el número de variables debe ser menor que el número de individuos. Debido a la naturaleza de los datos, es poco probable que se cumplan los tres principios de aplicación Xu y Cui (2008). Por ello, se han desarrollado diferentes alternativas no paramétricas que no han sido muy extendidas dentro de la bibliografía especializada. En este capítulo se presentan dos alternativas de MANOVA basado en distancias, que pueden emplearse cuando las restricciones del MANOVA no permiten su aplicación.

El Análisis Permutacional Multivariante de la Varianza (PERMANOVA) fue propuesto por Anderson (2001) como un test no paramétrico que, basándose en distancias, aproxima la distribución del estadístico de contraste a través de permutaciones. McArdle y Anderson (2001) amplió los principales resultados permitiendo el uso de cualquier Modelo Lineal Multivariante. El uso de esta técnica ha sido muy amplio en diversos campos, uno de los principales es la Ecología, disciplina en la que surgió este tipo de análisis.

Existen diversas técnicas, además del PERMANOVA, que permite la comparación de los centros de varios grupos. El trabajo de Clarke (1993) donde propone el Análisis de Similaridades (ANOSIM), precursor de la técnicas mencionada en el párrafo anterior, este autor con algunos colaboradores desarrollan el software PRIMER-e (Clarke *et al.*, 2017) que contiene ANOSIM y, recientemente, han incluido PERMANOVA. Clarke y Gorley fueron los autores principales, aunque ahora el número de colaboradores se ha incrementado.

En 1999 proponen como un análisis de distancias para datos estructurales (Gower y Krzanowski, 1999) que es, fundamentalmente, lo mismo que el PERMANOVA, sin embargo esta técnica está mucho menos extendida debido a que no existe un software específico para realizar su calculo y, probablemente, porque fue publicado en una revista de Estadística, que tienen menos difusión en los investigadores aplicados. Este test y su representación gráfica, como una proyección de los individuos sobre el Análisis de Coordenadas Principales de las medias de los grupos, han sido desarrollados posteriormente en el paquete de R MultBiplotR por Vicente-Villardon (2021).

Recientemente ha sido propuesto el MANOVA Bootstrap basado en distancias (BOOTMANOVA) por Vicente-Gonzalez y Vicente-Villardon (2019). Es una técnica similar a las anteriores, pero en su procedimiento se emplean técnicas Bootstrap. A diferencia del Análisis de Permutaciones que emplea muestreo sin reposición Splawa-Neyman *et al.* (1990), el Bootstrap Efron (1979) se basa en el remuestreo con reposición. El test de Permutaciones y el test Bootstrap han sido comparados o combinadas en multitud de ocasiones (ter Braak, 1992; Afanador *et al.*, 2013; Figueiredo, 2017; Janssen y Pauls, 2005). El uso de una u otra depende de la finalidad y los supuestos establecidos en el diseño de la investigación.

Para la correcta comprensión de este capítulo es necesario haber revisado los Modelos Lineales del capítulo 4.2, con base en ellos se establece la teoría del Análisis Mul-

tivariante de la Varianza en la sección 4.3. En la sección 4.4 se resumen algunos de los cálculos de las distancias que pueden ser utilizados en las técnicas que nos competen en este capítulo, contiene distancias para variables continuas (sección 4.4.1), distancias para variables binarias (sección 4.4.2), distancias para variables nominales (sección 4.4.3) y cómo realizar los cálculos de las distancias cuando las variables son de diversos tipos (sección 4.4.4). El desarrollo del PERMANOVA se encuentra en la sección 4.5 y el del BOOTMANOVA en 4.6. Ambas técnicas pueden llevar asociados diferentes gráficos que están recogidos en la sección 4.7, como se pueden observar en el ejemplo de la sección 4.9. Todos los cálculos es posible realizarlos con diferentes software también explicados dentro de este capítulo (sección 4.8).

4.2. Modelos Lineales

Los Modelos Lineales están ampliamente extendidos y son muy usados en la literatura con aplicaciones en un gran número de campos. La manera más frecuente en la que se encuentran los Modelos Lineales es la Regresión Lineal Simple, sin embargo, el conjunto de técnicas que están relacionadas con este tipo de modelo es muy amplio.

Cuando un conjunto de datos contiene las variables predictoras y otro las variables respuesta, es decir, las dos matrices sometidas a estudio no tienen papeles simétricos, las relaciones entre los dos conjuntos de datos se fundamentan en los Modelos Lineales.

Un gran número de técnicas están englobadas dentro de los Modelos Lineales Generales. En este capitulo se desarrollará brevemente el Modelo Lineal General Multivariante 4.2.1 que fundamentan parte de los resultados obtenidos en parte de este trabajo.

Hay un gran número de situaciones en las que no se cumple las condiciones de normalidad que deben existir para que se pueda aplicar el Modelo Lineal General Multivariante. Por ejemplo, si la respuesta es dicotómica, caso de especial interés en este documento.

Si el conjunto de variables respuesta no sigue una distribución normal multivariante el Modelo Lineal General no es óptimo. Nelder y Wedderburn (1972) proponen alternativas cuando las variables respuestas siguen distribuciones diferentes a la normal. En Myers y Montgomery (2018) encontramos un resumen de algunas de las alternativas propuestas. Se realizará una pequeña descripción del Modelo Lineal Generalizado en el capítulo 4.2.2.

4.2.1. Modelo Lineal General Multivariante

Partimos de una matriz de variables predictoras $\mathbf{X}_{(I\times J)}$ y una matriz de variables respuesta $\mathbf{Y}_{(I\times Q)}$ con J y Q variables, respectivamente, medidas en I individuos. El objetivo es estudiar la relación entre las respuestas $\mathbf{Y}$ y los predictores $\mathbf{X}$. La matriz $\mathbf{X}$ puede contener una primera columna de unos que permite incluir el término independiente en la ecuación.

Consideremos el modelo

$$\mathbf{Y} = \mathbf{XB} + \mathbf{U}, \tag{4.1}$$

donde $\mathbf{B}_{(J\times Q)}$ es la matriz de parámetros de regresión desconocidos y $\mathbf{U}$ es una matriz de residuales, cada una con media $\mathbf{0}$ y matriz de covarianzas común $\mathbf{\Sigma}$. Esta es la extensión del Modelo Lineal General Univariante utilizado, normalmente, en muchas aplicaciones.

Es de sobra conocido que, los estimadores de $\mathbf{B}$ existen si $\mathbf{X}$ es de rango completo. Los estimadores pueden ser calculados como

$$\hat{\mathbf{B}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}. \tag{4.2}$$

Se observa que, los estimadores del problema univariante para cada variable por sepa-

rado son los mismos que los que se obtienen en el modelo multivariante. En consecuencia, la obtención de las estimaciones no supone una ventaja frente al modelo univariante, es el contraste del modelo global el que proporciona una mejora en la aproximación multivariante.

En este tipo de modelos consideraremos la hipótesis de la forma

$$\mathbf{\Omega} = \mathbf{CBM} = \mathbf{0}, \tag{4.3}$$

donde $\mathbf{C}_{V\times J}$ y $\mathbf{M}_{Q\times W}$ son matrices dadas con rangos V y W respectivamente. Se recogen las combinaciones lineales de las columnas de $\mathbf{X}$ que se quieren contrastar en la matriz $\mathbf{C}$ y las combinaciones lineales especificas de las variables respuestas en la matriz $\mathbf{M}$. En la forma más clásica del modelo se omite la matriz de contrastes y se toma $\mathbf{M} = \mathbf{I}$.

$$\hat{\mathbf{\Omega}} = \mathbf{C}\hat{\mathbf{B}}\mathbf{M} = \mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{YM}, \tag{4.4}$$

es el estimador de varianza mínima de $\mathbf{\Omega}$.

Se pueden definir

$$\mathbf{R} = \mathbf{C}\ (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T, \tag{4.5}$$

$$\mathbf{E} = \mathbf{M}^T\mathbf{Y}^T[\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T]\mathbf{YM}, \tag{4.6}$$

y

$$\mathbf{H} = \mathbf{M}^T\mathbf{Y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T[\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T]^{-1}\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{YM} = \hat{\mathbf{\Omega}}^T\mathbf{R}^{-1}\hat{\mathbf{\Omega}}. \tag{4.7}$$

Donde $\mathbf{R}$ está relacionada con la inversa de la matriz de covarianzas entre los predictores, $\mathbf{E}$ es la matriz de covarianzas y productos cruzados del error y $\mathbf{H}$ es la matriz de covarianzas y productos cruzados asociada a la hipótesis.

Es posible contrastar la hipótesis (4.3) con el principio de unión-intersección de Roy (1953). Si y solo si las hipótesis univariantes $\mathbf{\Omega a} = \mathbf{CBMa} = \mathbf{0}$ se verifican para todos los

vectores no nulos $\mathbf{a}$, la hipótesis multivariante es verdadera.

El estadístico de contraste para cada uno de los contrastes univariantes es

$$F(\mathbf{a}) = \frac{(I-J)}{V}\frac{\mathbf{a}^T\mathbf{H}\mathbf{a}}{\mathbf{a}^T\mathbf{E}\mathbf{a}}.$$

Aceptamos la hipótesis multivariante (4.3) para un nivel α si

$$\bigcap_{\mathbf{a}} [F(\mathbf{a}) \leq F_{\alpha;V,I-J}],$$

para todo $\mathbf{a}$ no nulo.

Esta región de aceptación es equivalente a la definida por

$$\max_{\mathbf{a}} F(\mathbf{a}) \leq F_{\alpha;V,I-J}.$$

Introduciendo $\mathbf{a}^T\mathbf{E}\mathbf{a} = 1$ como restricción (denominador unidad), mediante multiplicadores de Lagrange, se puede demostrar que el máximo es proporcional a la mayor raíz de

$$|\mathbf{H} - \lambda\mathbf{E}| = 0, \tag{4.8}$$

donde $\mathbf{H}$ es la matriz asociada a la hipótesis definida en (4.7) y $\mathbf{E}$ es la matriz asociada al error que se definía en (4.6).

Las raíces características de

$$\mathbf{H}\mathbf{E}^{-1}, \tag{4.9}$$

son las mismas que las raíces no nulas de (4.8).

Mardia *et al.* (2009); Morrison (2005) o Seber (2009) recogen los estadísticos de contraste que emplean estas raíces características o una función de ellos.

El primer vector propio de (4.9) es el vector de coeficientes $\mathbf{a}$ que maximiza el cociente *F-ratio*. En el sentido de la regresión múltiple, este vector contiene la combinación lineal de las variables respuesta mejor explicada por las variables predictoras. De forma análoga, el segundo vector propio contiene la combinación lineal de la parte no explicada por el primer vector propio que mejor se relaciona con las variables predictoras. Los sucesivos vectores propios se construyen de forma similar. Utilizando los Modelos Lineales Generales Multivariantes buscamos las combinaciones lineales de las variables respuesta mejor explicadas por los predictores, a diferencia del Análisis de Componentes Principales (PCA) que busca combinaciones lineales de las respuestas que explique la mayor parte de la variabilidad posible, aunque en ambos casos pueden ser representados los individuos sobre el plano de forma similar.

Empleando la ecuación

$$\left(\mathbf{H}\mathbf{E}^{-1} - \lambda\right)\mathbf{a} = \mathbf{0}, \tag{4.10}$$

podemos obtener los vectores de la ecuación, donde, en general, la matriz (4.9) es no simétrica. Esta misma descomposición es posible obtenerla a partir de la matriz simétrica

$$\mathbf{E}^{-1/2}\mathbf{H}\,\mathbf{E}^{-1/2}. \tag{4.11}$$

Podemos escribir la ecuación (4.10) de la forma

$$[\mathbf{E}^{-1/2}\mathbf{H}\mathbf{E}^{-1/2} - \lambda\mathbf{I}]\mathbf{E}^{1/2}\mathbf{a} = \mathbf{0},$$

o

$$[\mathbf{E}^{-1/2}\mathbf{H}\mathbf{E}^{-1/2} - \lambda\mathbf{I}]\mathbf{g} = \mathbf{0}.$$

Si $\mathbf{g}$ es un vector propio de (4.11), entonces

$$\mathbf{a} = \mathbf{E}^{-1/2}\mathbf{v} \tag{4.12}$$

es un vector propio (4.10) con el mismo valor propio.

Como recogen Arnold (1981) y Seber (2009), para contrastar la hipótesis pueden definirse varios test estadísticos basados en estos valores propios y matrices.

Traza de Lawley-Hotteling :

$$T = traza(\mathbf{H}\mathbf{E}^{-1}) = \sum_i \lambda_i,$$

siendo λ_i los valores propios no negativos de (4.9).

Lambda de Wilks :

$$\Lambda = \frac{|\mathbf{E}|}{|\mathbf{H}+\mathbf{E}|} = \left|\mathbf{E}(\mathbf{H}+\mathbf{E})^{-1}\right| = \prod_i \lambda_i,$$

Estadístico de Pillai :

$$V = traza(\mathbf{E}(\mathbf{H}+\mathbf{E})^{-1}) = \sum_i \frac{\lambda_i}{1+\lambda_i},$$

que son los estadísticos utilizados por la mayor parte de los paquetes de software, junto con la mayor raíz característica mostrada anteriormente.

Todos los estadísticos pueden aproximarse asintóticamente mediante una F de Snedecor.

4.2.2. Modelo Lineal Generalizado

Como ya hemos mencionado en la introducción de este capítulo, existen muchas situaciones en las que los datos con los que buscamos trabajar no siguen una distribución normal y, por lo tanto, no es posible utilizar el Modelo Lineal General. Cuando las respuestas siguen otro tipo de distribuciones, como alternativa Finney (1952); Nelder (1966); Nelder y Wedderburn (1972) presentan el Modelo Lineal Generalizado expuesto en esta sección.

De forma breve y simplificada podemos decir que los Modelos Lineales Generalizados utilizan una función, a la que denominamos función de enlace, sobre la variable

respuesta que permita establecer una relación lineal con los predictores.

Para explicar este tipo de modelos partiremos de una matriz $\mathbf{X}$ que contiene los J predictores y una matriz de repuestas $\mathbf{Y}$ con Q variables dependientes. Además, fundamentamos el modelo en tres elementos:

- La función de distribución f de la matriz de respuestas $\mathbf{Y}$, que pertenecerá a la familia exponencial (normal, binomial, Poisson, etc.)
- Los predictores lineales, relacionados con la matriz $\mathbf{X}$, $\boldsymbol{\eta} = \mathbf{XB}$
- La función de enlace L, de forma que $E(\mathbf{Y}) = L^{-1}(\boldsymbol{\eta})$

Dentro de este tipo de modelos podemos incluir, por ejemplo, la Regresión Logística en el que la distribución utilizada es la binomial y la función de enlace es la *logit*. El modelo logit será utilizado posteriormente. No lo describimos aquí completamente porque no es necesario dentro de este capítulo.

4.3. Análisis Multivariante de la Varianza

En la mayor parte de los experimentos realizados en la actualidad, se obtienen respuestas de carácter multivariante con un número elevado de mediciones.

El Análisis Univariante de la Varianza de cada una de las variables por separado. Es una práctica habitual de los investigadores que, para evitar los falsos positivos, emplean métodos de corrección por comparaciones múltiples.

Como se mencionaba en la introducción, el ANOVA no es la técnica óptima debido a que el uso de cada una de las variables por separado, en lugar de todas de forma simultanea, puede tener varios inconvenientes:

- En un gran número de casos, el uso de modelos multivariantes aporta información sobre la relación entre las variables que de forma individual no podrían ser observadas, evitando así inferencias erróneas.

- La realización de un número menor de contrastes implica un mayor control del riesgo Tipo I. El uso de cada una de las variables por separado se realiza fundamentado en la independencia de estas, pero pueden no serlo.

- Puede existir información redundante que, utilizando técnicas multivariantes, sería eliminada.

- Otro de los inconvenientes se plantea cuando alguna de las variables no presenta diferencias significativas, es posible que la combinación de esa variable con otras del conjunto de datos sometido a estudio, sí que sea significativa y pueda pasar desapercibida.

En esta sección se presenta el Análisis Multivariante de la Varianza como la técnica más indicada para evitar este tipo de errores. Este método busca hacer máxima la F de Snedecor univariante empleando combianciones lineales de las variables medidas. El MANOVA también se puede entender como un Modelo Lineal General Multivariante (MGLM) explicado en la sección 4.2.1, esta será la forma empleada para el desarrollo de este trabajo.

En el MLGM partimos de una matriz de predictores $\mathbf{X}$ y una matriz de respuestas $\mathbf{Y}$ descritas en la sección 4.2.1. En el MANOVA, la matriz de predictores $\mathbf{X}$ está formada por variables categóricas. En la sección 4.3.1 explicaremos los modelos con un factor de variación y será en la sección 4.3.3 donde desarrollaremos los modelos más complejos con más de un factor de variación.

4.3.1. MANOVA con un factor de variación

La variable predictora categórica puede indicar el grupo al que pertenece la observación o el tratamiento que se ha aplicado en el diseño experimental. Esta técnica tiene como objetivo identificar si existen diferencias significativas entre dichos tratamientos o grupos.

Dado K niveles, grupos o categorías, las I filas de la matriz de datos $\mathbf{Y}$ estarán divididas en K grupos, con I_k individuos en cada categoría, siendo $k = 1, \dots, K$ y cumpliendo $I = I_1 + \cdots + I_k + \cdots + I_K$.

Es de sobra conocido que, en un modelo de regresión, pueden introducirse variables categóricas a través de variables dummy o indicadores que pueden tomar los valores 0 y 1 en función de si el individuo está dentro de la categoría de referencia o no lo está. Es posible definir una variable de este tipo para cada una de las categorías, pero no podrían ser introducidas de forma simultánea en un modelo de regresión con constante como variables independientes, ya que existiría una dependencia lineal y el modelo no podría se estimado debido a que la matriz $\mathbf{X}^T\mathbf{X}$ sería singular.

Matriz de diseño y estimación de los parámetros

La matriz de variables predictoras $\mathbf{X}$ será definida como la matriz de diseño de tamaño $I \times J$, que contiene por columnas los indicadores de todas las categorías menos una, junto a una columna, correspondiente a la constante, conformada por unos

$$\mathbf{X} = \begin{pmatrix} \mathbf{1}_{I_1} & \mathbf{1}_{I_1} & \mathbf{0}_{I_1} & \dots & \mathbf{0}_{I_1} \\ \mathbf{1}_{I_2} & \mathbf{0}_{I_2} & \mathbf{1}_{I_2} & \dots & \mathbf{0}_{I_2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{1}_{I_k} & \mathbf{0}_{I_k} & \mathbf{0}_{I_k} & \dots & \mathbf{0}_{I_k} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{1}_{I_{K-1}} & \mathbf{0}_{I_{K-1}} & \mathbf{0}_{I_{K-1}} & \dots & \mathbf{1}_{I_{K-1}} \\ \mathbf{1}_{I_K} & \mathbf{0}_{I_K} & \mathbf{0}_{I_K} & \dots & \mathbf{0}_{I_K} \end{pmatrix}, \tag{4.13}$$

donde $\mathbf{1}_{I_k}$ y $\mathbf{0}_{I_k}$ son los vectores de unos y ceros con I_k elementos del grupo k.

También es posible escribir la matriz de diseño tomando los indicadores de todas las categorías más la columna de unos, es decir, se diferencia de la matriz anterior en que no suprime ninguna de las categorías

$$\mathbf{X} = \begin{pmatrix} \mathbf{1}_{I_1} & \mathbf{1}_{I_1} & \mathbf{0}_{I_1} & \dots & \mathbf{0}_{I_1} & \mathbf{0}_{I_1} \\ \mathbf{1}_{I_2} & \mathbf{0}_{I_2} & \mathbf{1}_{I_2} & \dots & \mathbf{0}_{I_2} & \mathbf{0}_{I_2} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{1}_{I_{K-1}} & \mathbf{0}_{I_{K-1}} & \mathbf{0}_{I_{K-1}} & \dots & \mathbf{1}_{I_{K-1}} & \mathbf{0}_{I_{K-1}} \\ \mathbf{1}_{I_K} & \mathbf{0}_{I_K} & \mathbf{0}_{I_K} & \dots & \mathbf{0}_{I_K} & \mathbf{1}_{I_K} \end{pmatrix}. \tag{4.14}$$

Sin embargo, esta matriz de diseño (ecuación (4.14)) presenta algunas dificultades en la práctica, ya que no es posible calcular directamente la matriz de estimadores **B** debido a que del producto por su transpuesta ($\mathbf{X}^T\mathbf{X}$) se obtiene una matriz singular. Es posible solucionar este problema utilizando la inversa generalizada ($\left(\mathbf{X}^T\mathbf{X}\right)^-$) en lugar de la inversa regular ($\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$).

Construir la matriz de diseño **X** de diferentes formas da lugar a diversas parametrizaciones del modelo. Es de interés en este trabajo utilizar la opción más sencilla, será construida únicamente con los indicadores de cada una de las categorías

$$\mathbf{X} = \begin{pmatrix} \mathbf{1}_{I_1} & \mathbf{0}_{I_1} & \dots & \mathbf{0}_{I_1} & \mathbf{0}_{I_1} \\ \mathbf{0}_{I_2} & \mathbf{1}_{I_2} & \dots & \mathbf{0}_{I_2} & \mathbf{0}_{I_2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0}_{I_{K-1}} & \mathbf{0}_{I_{K-1}} & \dots & \mathbf{1}_{I_{K-1}} & \mathbf{0}_{I_{K-1}} \\ \mathbf{0}_{I_K} & \mathbf{0}_{I_K} & \dots & \mathbf{0}_{I_K} & \mathbf{1}_{I_K} \end{pmatrix}. \tag{4.15}$$

Utilizando esta matriz de diseño y considerando que la matriz de contrastes (**C**) y la matriz de combinaciones de las variables (**M**) son igual a la identidad (**I**), entonces

$$\mathbf{X}^T\mathbf{X} = diag(I_1, \dots, I_K) = \mathbf{D}_K, \tag{4.16}$$

es decir, la matriz diagonal que contiene el tamaño muestral de cada uno de los grupos.

De la misma forma, es posible afirmar que

$$\mathbf{R} = \mathbf{D}_K^{-1} = diag\left(\frac{1}{I_1}, \ldots, \frac{1}{I_K}\right),$$

es la matriz inversa de $\mathbf{D}_K$;

$$\hat{\Omega} = \hat{\mathbf{B}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \bar{\mathbf{Y}}$$

es la matriz de tamaño $K \times J$ que contiene los vectores de medias de cada grupo;

$$\mathbf{E} = \mathbf{Y}^T\mathbf{Y} - \bar{\mathbf{Y}}^T\mathbf{D}_K\bar{\mathbf{Y}}$$

y

$$\mathbf{H} = \bar{\mathbf{Y}}^T\mathbf{D}_K\bar{\mathbf{Y}}$$

son las matrices de sumas de cuadrados y productos dentro de los grupos y entre grupos.

Calculando los vectores propios de la matriz $\mathbf{E}^{-1}\mathbf{H}$ es posible obtener la combinación lineal de las variables observadas con mayor poder discriminante, que es de sobra conocido que corresponden con las coordenadas discriminantes o variables canónicas.

De esta forma, podemos interpretar la primera variable canónica como aquella que recoge la variabilidad máxima entre los grupos respecto a la variabilidad dentro de ellos, esta coordenada discriminante genera la F univariante más grande. Con las demás variables canónicas ocurre de forma análoga con la variabilidad no explicada en las coordenadas anteriores. Generalmente el número de variables canónicas es el mínimo entre el número de individuos menos el número de grupos y el número de grupos menos uno (mín $(I - K, K - 1)$), en la mayor parte de los casos corresponde con el número de grupos menos uno.

En definitiva, es posible asemejarlo a un Análisis de Componentes Principales (ACP) de la matriz de medias, teniendo en cuenta la variabilidad existente dentro de los grupos.

De forma análoga al ACP, es posible representar mediante un diagrama de dispersión las medias de los grupos sobre el espacio de coordenadas discriminantes. Las coordenadas canónicas de las medias, $\bar{\mathbf{Z}}$, pueden calcularse empleando los coeficientes de la ecuación (4.12)

$$\bar{\mathbf{Z}} = \bar{\mathbf{Y}}\mathbf{a} = \bar{\mathbf{Y}}\mathbf{E}^{-1/2}\mathbf{v}. \tag{4.17}$$

Esta representación tiene una propiedad importante, la distancia euclídea entre las medias en el espacio de la representación, es la distancia de Mahalanobis en el espacio inicial, y una aproximación al proyectar en las primeras variables canónicas.

De la misma forma que se proyectan las medias es posible proyectar los individuos originales, así se puede obtener un gráfico de visualización que permita observar la separación entre grupos. Las coordenadas de los individuos pueden ser representados en el espacio canónico de esta forma:

$$\mathbf{Z} = \mathbf{Y}\mathbf{a} = \mathbf{Y}\mathbf{E}^{-1/2}\mathbf{v}. \tag{4.18}$$

Utilizando la matriz de covarianzas dentro de los grupos,

$$\mathbf{S} = \frac{1}{I-K}\mathbf{E}, \tag{4.19}$$

es posible realizar esta misma representación, ya que se diferencian únicamente en una constante.

Para interpretar las variables canónicas es posible realizarlo de la misma forma que se realiza en el Análisis Factorial (AF), calculando las correlaciones entre las variables canónicas y las variables observadas. Al contrario que en el AF, las correlaciones pueden no estar optimizadas, por lo que en algunas ocasiones pueden tomar valores pequeños.

4.3.2. Matrices para las combinaciones lineales

Al estudiar los Modelos Lineales Generales Multivariantes en la sección 4.2.1, concretamente en la ecuación (4.3), se han definidos dos matrices de combinacionales lineales, una para cualquier conjunto de combinaciones de las columnas de $\mathbf{X}_{I\times J}$, $\mathbf{C}_{v\times J}$, y otra para cualquier combinación lineal específica de las variables respuesta de la matriz $\mathbf{Y}_{I\times Q}$, $\mathbf{M}_{Q\times w}$.

La matriz de combinaciones lineales de las medias puede ser definida como

$$\mathbf{C} = \begin{pmatrix} \mathbf{c}_1^T \\ \vdots \\ \mathbf{c}_V^T \end{pmatrix}, \tag{4.20}$$

donde $\mathbf{c}_v^T = (c_{v1}, \dots, c_{vJ})$ contiene los coeficientes para contrastar las combinaciones lineales de las columnas de $\mathbf{X}$. Si la matriz es la identidad ($\mathbf{C} = \mathbf{I}$), cada fila corresponde con una de las variables.

Para el caso del Análisis Multivariante de la Varianza con un único factor de variación, las combinaciones lineales pueden ser contrastes para las medias de los grupos. Si $\sum_{j=1}^{J} c_j = 0$ una combinación lineal de las medias de coeficientes, $\mathbf{c} = (c_1, \dots, c_J)^T$ es un contraste. Un ejemplo de la utilidad de este matriz puede ser las comparaciones de todos los grupos por parejas, todas ellas pueden realizarse mediante una matriz de contrastes

$$\mathbf{C} = \begin{pmatrix} 1 & -1 & 0 & 0 & \dots & 0 & 0 \\ 1 & 0 & -1 & 0 & \dots & 0 & 0 \\ 1 & 0 & 0 & -1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & 0 & 0 & 0 & \dots & -1 & 0 \\ 1 & 0 & 0 & 0 & \dots & 0 & -1 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 1 & 0 & 0 & \dots & 0 & -1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 & -1 \end{pmatrix}. \tag{4.21}$$

De la misma forma que en los análisis univariantes se pueden utilizar las correcciones por comparaciones múltiples, es posible realiza el contraste simultáneo de todas las comparaciones o hacer cada una de ellas por separado.

La matriz de combinaciones lineales de las variables respuestas se puede definir como

$$\mathbf{M} = \begin{pmatrix} \mathbf{m}_1 \\ \vdots \\ \mathbf{m}_W \end{pmatrix} \tag{4.22}$$

para la que $\mathbf{m}_w^T = \left(m_{1w}, \dots, m_{Qw}\right)^T$.

La forma más habitual de realizar los contrastes de todas las variables respuestas de forma simultánea, de esta forma, la matriz $\mathbf{M}$ corresponde con la matriz identidad $\mathbf{M} = \mathbf{I}$. En algunos casos, por ejemplo cuando se quiere realizar un Análisis de Perfiles, puede ser de interés hacer la comparación entre variables respuesta, en este tipo de análisis se calculan las diferencias entre diversas mediciones de una misma variable en diferentes momentos de tiempo, en lugar de las variables originales.

En este caso la matriz **M** sería de la siguiente forma:

$$\mathbf{M} = \begin{pmatrix} 1 & 0 & 0 & 0 & \dots & 0 & 0 \\ -1 & 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & -1 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & -1 & 1 & \dots & 0 & 0 \\ 0 & 0 & 0 & -1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & -1 & 1 \\ 0 & 0 & 0 & 0 & \dots & 0 & -1 \end{pmatrix}. \tag{4.23}$$

Es posible emplear estas dos matrices para personalizar la forma en la que se ajusta el modelo lo más posible al interés del investigador.

Es de sobra conocido que es posible separar las sumas de cuadrados en tantas partes como grados de libertad asociados a la hipótesis correspondiente tengamos. Se puede realizar esta separación utilizando contrastes ortogonales.

Dos contrastes $\mathbf{c} = (c_1, \dots, c_V)^T$ y $\mathbf{d} = (d_1, \dots, d_V)^T$ son ortogonales si la suma de productos de sus cocientes es 0,

$$\sum_{v=1}^{V} \mathbf{c}_v \mathbf{d}_v = 0.$$

Los contrastes de Helmert son una posible forma conocida para obtener un conjunto de contrastes ortogonales. Esta forma consiste en comparar, de forma recursiva, cada grupo con cada uno de demás contrastes. Por ejemplo para un caso en el que existan cuatro grupos, los contrastes podrían ser

$$\mathbf{C} = \begin{pmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \\ \mathbf{c}_3 \end{pmatrix} = \begin{pmatrix} 3 & -1 & -1 & -1 \\ 0 & 2 & -1 & -1 \\ 0 & 0 & 1 & -1 \end{pmatrix}. \tag{4.24}$$

4.3.3. Diseños más complejos

En un gran número de casos en las aplicaciones practicas reales, el número de factores de variación es mayor a dos, y además, puede ser de interés incluir intersecciones entre ellos. Todas estas opciones se recogen en la matriz de contrates **C**, permitirá aislar tantos los efectos principales como sus interacciones.

Si tenemos un ejemplo con dos factores de variación que tengan K_1 y K_2 niveles respectivamente. Es posible calcular los grados de libertad para cada uno de los efectos como $(K_1 - 1)$ y $(K_2 - 1)$. Los grados de libertad de la interacción también pueden ser calculados como $(K_1 - 1)(K_2 - 1)$.

La matriz de contrastes **C** será construida partiendo de contrastes ortogonales para cada uno de los efectos, $\mathbf{C}_1$ y $\mathbf{C}_2$, de la misma forma que se ha construido en la ecuación (4.24) y aislaremos la interacción entre ambos factores a través de la matriz de contrastes $\mathbf{C}_{12}$. Estas matrices tendrán los tamaños $(K_1 - 1) \times K$, $(K_2 - 1) \times K$ y $(K_1 - 1)(K_2 - 1) \times K$ respectivamente.

La matriz de contrastes

$$\mathbf{C} = \begin{pmatrix} \mathbf{C}_1 \\ \mathbf{C}_2 \\ \mathbf{C}_{12} \end{pmatrix}, \tag{4.25}$$

permite separar la variabilidad global en $(K - 1)$ contrates, uno para cada uno de los grados de libertad disponibles.

La tercera matriz será construida como el producto de cada una de las filas de la matriz $\mathbf{C}_1$ por cada una de las filas de la matriz $\mathbf{C}_2$.

Volvamos al ejemplo en el que tenemos tres grados de libertad, uno para cada uno de los efectos principales y otro para la intersección, es decir cada una de las matrices de (4.25) tiene únicamente una fila. Las dos primeras filas permitirán aislar cada uno de los efectos principales, la tercera la interacción entre ambos

$$\mathbf{C} = \begin{pmatrix} \mathbf{C}_1 \\ \mathbf{C}_2 \\ \mathbf{C}_{12} \end{pmatrix} = \begin{pmatrix} 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}. \tag{4.26}$$

En caso necesario,es posible hacer una proyección sobre la representación canónica de las medias para ayudar en la interpretación de los resultados.

4.4. Cálculo de distancias entre las filas de una matriz de datos

Existen diversas formas de calcular las distancias y similaridades entre datos para formar la matriz de distancias que se pueden emplear en la aplicación de estas técnicas. La elección de cualquiera de ellas está condicionada al tipo de datos, por ello en este documento se especifican distancias para variables continuas, binarias, nominales y diferentes tipos de variables. Dentro de cada uno de los tipos de variables, por regla general, la medida se elige en función del ámbito al que pertenecen los datos sometidos a estudio.

En esta sección se resumen algunas de las distancias más utilizadas. Gower (1971); Gray y Markel (1976); Cuadras Avellanas (1988); Zhang y Srihari (2003); Boriah *et al.* (2008); Choi *et al.* (2010); Han *et al.* (2011); Goshtasby (2012) recogen un resumen de todos los índices desarrollados en esta sección.

4.4.1. Distancias para variables continuas

Partiendo de una matriz de datos definida en el capítulo 4.2, $\mathbf{Y}_{(I\times Q)}$, con Q variables continuas, que puede ser entendida como la concatenación de vectores de medidas de los I individuos en las Q variables

$$\mathbf{Y} = \begin{pmatrix} \mathbf{y}_1^T \\ \vdots \\ \mathbf{y}_i^T \\ \vdots \\ \mathbf{y}_I^T \end{pmatrix}, \tag{4.27}$$

con $\mathbf{y}_i^T = \left(y_{i1}, \ldots, y_{iQ}\right)^T$.

En la siguiente lista se encuentran algunas de las medidas que podemos utilizar para datos continuos:

Pitagórica:

$$\delta_{ii'}^P = \sqrt{\frac{\sum_{j=1}^J \left(y_{ij} - y_{i'j}\right)^2}{J}}$$

Taxonómica:

$$\delta_{ii'}^T = \sqrt{\frac{\sum_{j=1}^J \frac{\left(y_{ij}-y_{i'j}\right)^2}{d^2}}{J}}$$

Ciudad:

$$\delta_{ii'}^C = \frac{\sum \frac{|y[i,]-y[j,]|}{d}}{J}$$

Euclídea Ordinaria:

$$\delta_{ii'} = \sqrt{\sum_{j=1}^J \left(y_{ij} - y_{Ij}\right)^2}$$

Minkowsky:

$$\delta_{ii'}^{d} = \left(\sum_{j=1}^{d} \left| y_{ij} - y_{Ij} \right|^{d} \right)^{1/d}$$

Divergencia:

$$\delta_{ii'}^{D} = \left(\frac{1}{J} \sum_{j=1}^{J} \left(\frac{\left| y_{ij} - y_{i'j} \right|}{y_{ij} + y_{i'j}} \right)^{2} \right)^{1/2}$$

Suma de las diferencias absolutas:

$$\delta_{ii'}^{ADS} = \sum \frac{\left| y_{ij} - y_{i'j} \right|}{\left| y_{ij} + y_{i'j} \right|}$$

Camberra:

$$\delta_{ii'}^{CA} = \sum_{j=1}^{J} \frac{\left| y_{ij} - y_{i'j} \right|}{y_{ij} + y_{i'j}}$$

Bray-Curtis:

$$\delta_{ii'}^{BC} = \frac{\sum_{j=1}^{J} \left| y_{ij} - y_{i'j} \right|}{\sum_{j=1}^{J} \left(y_{ij} + y_{i'j} \right)}$$

Soergel:

$$\delta_{ii'}^{S} = \frac{\sum_{j=1}^{J} \left| y_{ij} - y_{i'j} \right|}{\sum_{j=1}^{J} \text{máx} \left(y_{ij}, y_{i'j} \right)}$$

Ware Hedges:

$$\delta_{ii'}^{WH} = \frac{\sum \left| y_{ij} - y_{i'j} \right|}{\sum \text{máx} \left(y_{ij}, y_{i'j} \right)}$$

Todas las distancias presentadas se encuentran recogidas dentro del paquete PERMANOVA (Vicente-Gonzalez y Vicente-Villardon, 2021) en la función *"DistContinuous"*.

4.4.2. Distancias para variables binarias

Para el caso de los datos binarios, en primer lugar se debe calcular una medida de similaridad entre cada par de individuos ($s_{ii'}$) que se convertirá posteriormente en una medida de distancia a través de una de las siguientes fórmulas:

1 $\delta_{ii'} = s_{ii'}$

2 $\delta_{ii'} = 1 - s_{ii'}$

3 $\delta_{ii'} = \sqrt{1 - s_{ii'}}$

4 $\delta_{ii'} = -\log\left(s_{ii'}\right)$

5 $\delta_{ii'} = \frac{1}{s_{ii'}-1}$

6 $\delta_{ii'} = \sqrt{2\,(1 - s_{ii'})}$

7 $\delta_{ii'} = \frac{1-(s_{ii'}+1)}{2}$

8 $\delta_{ii'} = 1 - |s_{ii'}|$

9 $\delta_{ii'} = \frac{1}{s_{ii'}+1}$

Para mostrar algunas de las medidas de similaridad, definiremos dos vectores de la matriz $\mathbf{Y}$, que ahora estará compuesta por Q variables binarias, como $\mathbf{y}_i$ y $\mathbf{y}_{i'}$. En ambos vectores la codificación será 1 si está presente la variable y 0 si no lo está. A continuación se creará una tabla como la de la tabla 4.1.

i/i'	Presente	Ausente	Total
Presente	$a = \mathbf{y}_i^T \mathbf{y}_{i'}$	$b = \mathbf{y}_i^T (1 - \mathbf{y}_{i'})$	$a + b$
Ausente	$c = (1 - \mathbf{y}_i)^T \mathbf{y}_{i'}$	$d = (1 - \mathbf{y}_i)^T (1 - \mathbf{y}_{i'})$	$c + d$
Total	a+c	b+d	a+b+c+d

Tabla 4.1: Tabla de contingencia para el cruce de dos individuos

Por tanto, quedan definidas a, b, c y d, tal que a sea el número de caracteres presentes en las dos variables que están siendo comparadas, b y c sean el número de caracteres presentes unicamente en una de los variables sometidas a estudio y d sea el número de individuos que no está presente en ninguna de las dos variables. A partir de estas cuatro frecuencias será posible realizar los cálculos de las medidas de similitud y asociación entre variables.

Igual que en el caso de las variables continuas, en la siguiente lista se recogen algunos de los índices de similaridad.

Kulezynski1:

$$s_{ii'}^K = \frac{a}{b + c}$$

Russell y Rao

$$s_{ii'}^{RR} = \frac{a}{a+b+c+d}$$

Jaccard

$$s_{ii'}^{J} = \frac{a}{a+b+c}$$

Concordancia Simple

$$s_{ii'}^{SM} = \frac{a+d}{a+b+c+d}$$

Anderberg

$$s_{ii'}^{A} = \frac{a}{a+2(b+c)}$$

Rogers y Tanimoto

$$s_{ii'}^{RT} = \frac{a+d}{a+2(b+c)+d}$$

Sorensen, Dice y Czekanowski

$$s_{ii'}^{SDC} = \frac{a}{a+0.5(b+c)}$$

Sneath y Sokal

$$s_{ii'}^{SS} = \frac{a+d}{a+0.5(b+c)+d}$$

Hamman

$$s_{ii'}^{H} = \frac{a-(b+c)+d}{a+b+c+d}$$

Kulezynski2

$$0.5\left(\frac{a}{a+b}+\frac{a}{a+c}\right)$$

Anderberg2

$$s_{ii'}^{A2} = 0.25\left(\frac{a}{a+b}+\frac{a}{a+c}+\frac{d}{c+d}+\frac{d}{b+d}\right)$$

Ochiai

$$s_{ii'}^{O} = \frac{a}{\sqrt{(a+b)(a+c)}}$$

S13

$$s_{ii'}^{S13} = \frac{(ad)}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}/$$

ϕ de Pearson

$$s_{ii'}^{\phi} = \frac{(ad - bc)}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$$

Yule

$$s_{ii'}^{Y} = \frac{ad - bc}{ad + bc}$$

Sorensen-Dice

$$s_{ii'}^{SD} = \frac{2a}{2a + b + c}$$

Estos índices se pueden calcular utilizando el paquete PERMANOVA (Vicente-Gonzalez y Vicente-Villardon, 2021) con la función *"DistBinary"*. Usando esta función también se podrá elegir la transformación que utilizar para calcular la distancia a partir de la similaridad, por defecto emplearemos $\delta_{ii'} = \sqrt{1 - s_{ii'}}$.

4.4.3. Distancias para variables categóricas

Para realizar el cálculo de las distancias cuando la matriz $\mathbf{Y}_{(I\times Q)}$ está compuesta por Q variables nominales será necesario calcular una medida de similaridad ($s_{ii'j}$) y una ponderación (w_j) para cada variable, que serán combinadas para calcular la similaridad global

$$s_{ii'} = \sum_{j=1}^{J} w_j s_{ii'j}.$$

Las medidas de similitud y las ponderaciones pueden estar influenciadas por el número de categorías de la variable (K), la frecuencia de cada categoría de la variable j ($f_j(y)$) y la probabilidad de que en la variable j tome la categoría k dado por $\hat{p}_j(y) = \frac{f_j(y)}{I}$, o por $p_j^2(y) = \frac{f_j(y)(f_j(y)-1)}{I(I-1)}$.

Igual que en la sección anterior, la similaridad global será transformada en una medida de distancia a través de las fórmulas recogidas en la sección 4.4.2.

En la siguiente lista se detallan algunas de las posibles opciones para realizar el cálculo de las similaridades y ponderaciones para las variables nominales.

Método de superposición:

$$s_{ii'j} = \begin{cases} 1 & \text{si } y_{ij} = y_{i'j} \\ 0 & \text{en otro caso} \end{cases}; \qquad w_j = \frac{1}{J}$$

Eskin:

$$s_{ii'j} = \begin{cases} 1 & \text{si } y_{ij} = y_{i'j} \\ \frac{K^2}{K^2+2} & \text{en otro caso} \end{cases}; \qquad w_j = \frac{1}{J}$$

IOF

$$s_{ii'j} = \begin{cases} 1 & \text{si } y_{ij} = y_{i'j} \\ \frac{1}{1+\log f_j(y_{ij}) \times \log f_j(y_{i'j})} & \text{en otro caso} \end{cases}; \qquad w_j = \frac{1}{J}$$

OF

$$s_{ii'j} = \begin{cases} 1 & \text{si } y_{ij} = y_{i'j} \\ \frac{1}{1+\log \frac{I}{f_j(y_{ij})} \times \log \frac{I}{f_j(y_{i'j})}} & \text{en otro caso} \end{cases}; \qquad w_j = \frac{1}{J}$$

Goodall3

$$s_{ii'j} = \begin{cases} 1 - p_j^2(y_{ij}) & \text{si } y_{ij} = y_{i'j} \\ 0 & \text{en otro caso} \end{cases}; \qquad w_j = \frac{1}{J}$$

Lin

$$s_{ii'j} = \begin{cases} 2\log \hat{p}_j(y_{ij}) & \text{si } y_{ij} = y_{i'j} \\ 2\log\left(\hat{p}_j(y_{ij}) + \hat{p}_j(y_{i'j})\right) & \text{en otro caso} \end{cases}; \qquad w_j = \frac{1}{\sum_{j=1}^{J} \log \hat{p}_j(y_{ij}) + \log \hat{p}_j(y_{i'j})}$$

Estas distancias pueden encontrarse en el paquete MultBiplotR (Vicente-Villardon, 2021) en la función *"NominalDistances"*.

4.4.4. Distancias para variables de diferentes tipos

Es posible que las Q variables de la matriz $\mathbf{Y}_{(I\times Q)}$ no sean todas de un único tipo. En este caso, para calcular las distancias, primero se calculará para cada variable una similaridad ($0 \leq s_{ii'j} \leq 1$) y una ponderación ($0 \leq w_{ii'j} \leq 1$), que para obtener la similaridad

global, serán combinadas empleando una media ponderada

$$s_{ii'} = \frac{\sum_{j=1}^{J} s_{ii'j} w_{ii'j}}{\sum_{j=1}^{J} w_{ii'j}}. \tag{4.28}$$

El cálculo de la similaridad y la ponderación estará condicionado al tipo de variable. A continuación se detallan los cálculos para cada uno de ellos.

Variables Binarias :

Similaridad:

- $s_{ii'j} = 1$ cuando coinciden ambas variables.
- $s_{ii'j} = 0$ cuando no hay coincidencia entre las variables.

Ponderación:

- $w_{ii'j} = 0$ cuando existe ausencia de ambas variables, correspondería a la d de la tabla 4.1.
- $w_{ii'j} = 1$ para el resto de casos.

Variables Nominales :

Similaridad:

- $s_{ii'j} = 1$ para coincidencias entre variables.
- $s_{ii'j} = 0$ para divergencias sin tener en cuenta el número de categorías.

Ponderación:

- $w_{ii'j} = 0$ para datos perdidos.
- $w_{ii'j} = 1$ para el resto de casos.

Variables Cuantitativas :

Similaridad:

- $s_{ii'j} = 1 - \frac{|y_{ij} - y_{i'j}|}{R_j}$ donde R_j es el rango de la j-ésima variable.

Ponderación:

- $w_{ii'j} = 0$ para datos perdidos.
- $w_{ii'j} = 1$ para el resto de casos.

Una vez calculada la similaridad global con la ecuación (4.28), debemos transformarla en una medida de distancia con una de las formulas recogidas en la sección 4.4.2, en nuestro caso utilizaremos por defecto $\delta_{ii'} = \sqrt{1 - s_{ii'}}$.

4.5. Análisis de la Varianza basado en distancias y permutaciones (PERMANOVA)

Como ya se ha mencionado en la introducción del capítulo, Anderson (2001); McArdle y Anderson (2001) describieron el PERMANOVA. Esta técnica consiste en realizar un análisis de permutaciones del MANOVA basado en distancias.

Los pasos a seguir para realizar un PERMANOVA son los siguientes:

- Realizar el cálculo de la matriz de distancias o disimilitudes entre individuos.
- Calcular las sumas de cuadrados totales y las sumas de cuadrados dentro de los grupos, que por diferencia, permiten calcular las sumas de cuadrados entre los grupos.
- Empleando las sumas de cuadrados del punto anterior, se calculará la pseudo-F inicial de la misma forma que será calculada en el análisis univariante.
  Estos tres puntos permitirán obtener un MANOVA basado en distancias entre los datos originales.
- A continuación, se realiza la estimación de la distribución muestral, para ello se supone que la hipótesis nula es cierta. Serán empleadas al azar un número elevado de permutaciones, ya que no es posible utilizarlas todas. Para cada una de las permutaciones se calculará la F que nos permita crear la distribución buscada.

- Finalmente, se obtendrá el p-valor asociado calculando la proporción de los valores F que se han obtenido de las permutaciones que sean mayores que la pseudo-F inicial, calculada de las distancias de los datos originales.

La justificación teórica de esta técnica la recogen McArdle y Anderson (2001).

En la introducción también se ha comentado que existe una técnica análoga a la que ha sido descrita en esta sección desarrollada por Gower y Krzanowski (1999). Su propuesta se fundamenta en realizar un análisis de distancias para datos estructurados multivariantes, utilizando como representación gráfica un Análisis de Coordenadas Principales de las medias de los grupos sobre las que realizan una proyección del conjunto completo de individuos. La técnica propuesta por Gower y Krzanowski (1999) ha tenido una menor repercusión que el PERMANOVA debido a que no existe un software asociado para su realización. Otros de los motivos para su menor difusión ha sido la publicación en una revista propiamente de Estadística, que tiene una menor divulgación entre los investigadores. Recientemente se ha publicado en un paquete de R que contiene también esta técnica (Vicente-Villardon *et al.*, 2006), se puede realizar tanto el test como la representación de los centroides mediante Coordenadas Principales de las medias de los grupos sobre las que se proyectan el conjunto completo de individuos.

Los software con los que se pueden realizar el PERMANOVA serán descritos más adelante en la sección 4.8.

4.6. Análisis de la Varianza basado en distancias y bootstrap (BOOTMANOVA)

Goodnight y Schwartz (1997); Krishnamoorthy y Lu (2010); Aelst y Willems (2011); Konietschke *et al.* (2015); Xu (2015); Lin *et al.* (2021) han utilizado anteriormente técnicas bootstrap asociadas a MANOVA, sin embargo en la mayor parte de los casos no se emplean distancias.

Haciendo una generalización del PERMANOVA resumido en el apartado anterior (sección 4.5), se empleará, en lugar de un análisis de permutaciones, el método bootstrap para realizar la estimación de la distribución muestral.

De forma análoga a la sección 4.5, se empleará el cálculo de distancias de la sección 4.4, se realiza el cálculo de la pseudo-F inicial como la suma de cuadrados de las distancias entre grupos y dentro de los grupos. La distribución del estadístico será estimado a través de bootstrap bajo la hipótesis nula calculando el correspondiente p-valor.

4.6.1. Diseños con un factor de variación

Esta técnica tiene como objetivo buscar si existen diferencias significativas entre los K grupos o tratamientos definidos. Por ello, la hipótesis estadística asociada a este objetivo en notación matemática sería la siguiente:

$$\begin{cases} \mathrm{H}_0 \ : \ \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_K \\ \mathrm{H}_a \ : \ \exists i, j, \in (1, 2, \ldots K), \boldsymbol{\mu}_i \neq \boldsymbol{\mu}_j \end{cases}, \tag{4.29}$$

donde $\boldsymbol{\mu}_k$ es el vector de medias del grupo k, donde $k = 1, 2..., K$. Esta hipótesis puede ser definida también como:

$$\begin{cases} \mathrm{H}_0 \ : \ \text{Todos los grupos son iguales} \\ \mathrm{H}_a \ : \ \text{Existen diferencias entre algunos de los grupos} \end{cases}. \tag{4.30}$$

Igual que en el PERMANOVA partimos de una matriz de datos con I filas divididas en K grupos con I_k individuos en cada uno de ellos, donde $k = 1, \ldots, K$ y $I = I_1 + I_2 + \cdots + I_K$, y J variables. Tomamos además, la matriz de diseño $\mathbf{X}_{(I \times K)}$ con I filas y K columnas que se ha construido de la forma (4.15).

Como ya se ha mencionado en las introducciones del capítulo y de esta sección, en el BOOTMANOVA es necesario calcular la matriz de distancias a través de las similitudes

o disimilitudes de la sección anterior (4.4).

A partir de la matriz de distancias calculadas que denominaremos $\mathbf{\Delta}_{(I\times I)}$ con I filas y I columnas. Esta matriz tendrá ceros en diagonal principal.

$$\mathbf{\Delta} = \begin{pmatrix} 0 & \delta_{12} & \dots & \delta_{1(I-1)} & \delta_{1I} \\ \delta_{21} & 0 & \dots & \delta_{2(I-1)} & \delta_{2I} \\ \vdots & \vdots & 0 & \vdots & \vdots \\ \delta_{(I-1)1} & \delta_{(I-1)2} & \dots & 0 & \delta_{(I-1)I} \\ \delta_{I1} & \delta_{I2} & \dots & \delta_{I(I-1)} & 0 \end{pmatrix}. \tag{4.31}$$

De la misma forma, la matriz que contiene los cuadrados de las distancias entre individuos será denominada $\mathbf{\Delta}^2_{(I\times I)}$:

$$\mathbf{\Delta}^2 = \begin{pmatrix} 0 & \delta^2_{12} & \dots & \delta^2_{1(I-1)} & \delta^2_{1I} \\ \delta^2_{21} & 0 & \dots & \delta^2_{2(I-1)} & \delta^2_{2I} \\ \vdots & \vdots & 0 & \vdots & \vdots \\ \delta^2_{(I-1)1} & \delta^2_{(I-1)2} & \dots & 0 & \delta^2_{(I-1)I} \\ \delta^2_{I1} & \delta^2_{I2} & \dots & \delta^2_{I(I-1)} & 0 \end{pmatrix}. \tag{4.32}$$

Es posible realizar los cálculos para las sumas de cuadrados total empleando las distancias

$$\mathrm{SC_T} = \frac{1}{I}\sum_{i=1}^{I-1}\sum_{i'=i+1}^{I}\delta^2_{ii'}. \tag{4.33}$$

Teniendo en cuenta que la matriz es simétrica y completa, es de sobra conocido que la suma de cuadrados de las distancias de cada uno de los puntos al centroides es igual a la suma de los cuadrados de las interdistancias entre todos los puntos divididos entre el número de ellos

$$\sum_{i=1}^{I} \delta^2(\mathbf{x}_i, \bar{\mathbf{x}}) = \frac{1}{I} \sum_{i<I} \delta^2(\mathbf{x}_i, \mathbf{x}_{i'}). \tag{4.34}$$

La ecuación (4.33) de las sumas de cuadrados totales puede ser escrita de forma matricial con la suma de las distancias de la diagonal inferior de la matriz completa entre el número de observaciones

$$SC_T = \frac{1}{I}\mathbf{1}^T \frac{1}{2}\Delta^2 \mathbf{1}, \tag{4.35}$$

donde $\mathbf{1}_I$ es un vector columna con I unos.

Las sumas de cuadrados dentro de los grupos es un poco más compleja, las distancias que deben ser calculadas serán hasta el centroide de cada uno de los grupos. Empleando la propiedad (4.34), ilustrada en la figura 4.1, es posible realizar los cálculos de las sumas de cuadrados dentro de los grupos.

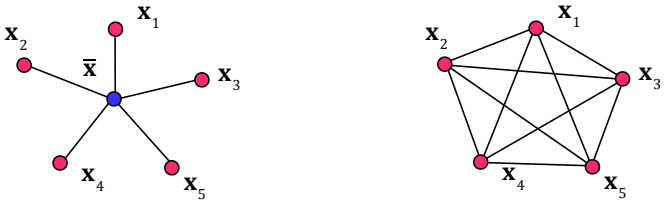

Figura 4.1: Sumas de cuadrados de las distancias

Incluiremos la matriz de pertenencia $\mathbf{P}_{(I\times I)}$ donde tenía I filas e I columnas compuesta por ceros y unos, donde $p_{ii'}$ será igual a 1 si los individuos i e i' pertenecer a un mismo grupo y 0 si pertenecer a grupos distintos. Así, podemos definir la matriz $\tilde{\Delta} = \mathbf{P} * \Delta$, donde $*$ es el producto elemento a elemento de la matriz $\mathbf{P}$ y la matriz Δ.

De esta forma, en un diseño en el que haya dos grupos, la matriz de distancias dentro de los grupos tendría la siguiente forma:

$$\tilde{\Delta}_2 = \begin{pmatrix} 0 & \delta_{12} & \dots & \delta_{1(I_1-1)} & \delta_{1I_1} & 0 & 0 & 0 & 0 & 0 \\ \delta_{21} & 0 & \dots & \delta_{2(I_1-1)} & \delta_{2I_1} & 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & 0 & \vdots & \vdots & 0 & 0 & 0 & 0 & 0 \\ \delta_{(I_1-1)1} & \delta_{(I_1-1)2} & \dots & 0 & \delta_{(I_1-1)I_1} & 0 & 0 & 0 & 0 & 0 \\ \delta_{I_1 1} & \delta_{I_1 2} & \dots & \delta_{I_1(I_1-1)} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \delta_{(I_1+1)(I_1+2)} & \dots & \delta_{(I_1+1)(I_2-1)} & \delta_{(I_1+1)I_2} \\ 0 & 0 & 0 & 0 & 0 & \delta_{(I_1+2)(I_1+1)} & 0 & \dots & \delta_{(I_1+2)(I_2-1)} & \delta_{(I_1+2)I_2} \\ 0 & 0 & 0 & 0 & 0 & \vdots & \vdots & 0 & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \delta_{(I_2-1)(I_1+1)} & \delta_{(I_2-1)(I_1+2)} & \dots & 0 & \delta_{(I_2-1)I_2} \\ 0 & 0 & 0 & 0 & 0 & \delta_{I_2(I_1+1)} & \delta_{I_2(I_1+2)} & \dots & \delta_{I_2(I_2-1)} & 0 \end{pmatrix}. \quad (4.36)$$

Como se puede observar en la ecuación (4.36), si dos individuos i e i' pertenecen a un mismo grupo, $\tilde{\delta}_{ii'} = \delta_{ii'}$ y si pertenecen a grupos distintos $\tilde{\delta}_{ii'} = 0$.

Utilizando las matrices definidas anteriormente y $\mathbf{D}_K$ definida en la ecuación (4.16), es posible calcular las sumas de cuadrados dentro de los grupos a partir de la siguiente formula. Ha sido definida en forma matricial

$$SC_D = \mathbf{1}_K^T \mathbf{D}_K^{-1/2} \mathbf{X}^T (\frac{1}{2}\tilde{\Delta}^2) \mathbf{X} \mathbf{D}_K^{-1/2} \mathbf{1}_K. \quad (4.37)$$

La suma de cuadrados entre grupos puede definirse como la diferencia entre la suma de cuadrados total y la suma de cuadrados dentro de los grupos

$$SC_E = SC_T - SC_D. \quad (4.38)$$

Esta primera parte del proceso ha sido ilustrado a través de un esquema que se encuentra en la figura 4.2, en ella se recoge desde el cálculo de las distancias hasta las sumas de cuadrados total y dentro de los grupos de nuestro modelo.

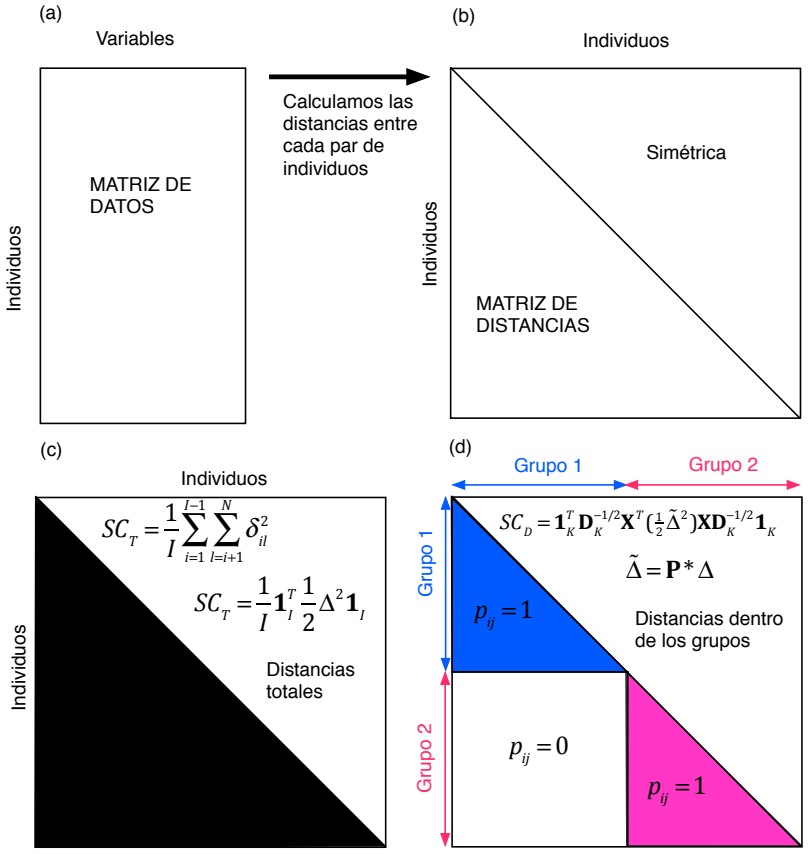

Figura 4.2: Cálculo de las distancias. (a) Matriz de datos brutos. (b) Matriz simétrica que contiene las distancias. (c) Sumas de cuadrados totales. (d) Sumas de cuadrados dentro de los grupos.

Para realizar el cálculo del pseudo estadísitco F inicial, por analogía al ANOVA, será realizado a partir de las sumas de cuadrados

$$F = \frac{SC_E \,/(K-1)}{SC_D \,/(I-K)}. \tag{4.39}$$

La F calculada coincidirá con la F univariante del procedimiento tradicional si, tenemos una única variable dependiente, la distancia utilizada es la distancia euclídea y se cumple el principio de normalidad. En este caso además seguirá una distribución F de Snedecor, sino la distribución es desconocida ya que las variables no son normales y es posible emplear una medida de distancia distinta.

En este caso, para calcular la distribución muestral de nuestro estadístico F emplearemos el método bootstrap. Esta técnica consiste en realizar un remuestreo con remplazamiento de los I elementos de la muestra original B veces. Para cada una de las repeticiones se estimará F como se ha calculado en la (4.39), esto permitirá obtener la distribución del estadístico. El número de remuestreos B suele estar entre 1000 y 2000 muestras. El desarrollo de estas técnicas se encuentra en Efron (1979); Efron y Tibshirani (1986, 1994).

Si los vectores de medias de todos los grupos son iguales, es decir se cumple la hipótesis nula planteada en la ecuación (4.29), las observaciones podrían intercambiarse y el remuestreo realizado sobre los individuos no influiría, ya que las etiquetas de cada uno de los grupos se pueden asignadas al azar. Calcularemos entonces un $p - valor$ asociado comparando cada una de las F obtenidas en los B remuestreos realizados, que formaban la distribución muestral y serán denominadas F^{π} en el resto del documento, con el valor original.

El p-valor será definido como la probabilidad de que los resultados de la muestra realizada empleando métodos bootstrap, si la hipótesis nula (ecuación (4.29)) es cierta, sean más extremos que el resultado obtenido en la muestra original.

De la misma forma que se ha realizado en el PERMANOVA descrito por Anderson (2001), es posible definir obtener el p-valor de dos formas diferentes

$$p - val = \frac{\text{Número de } F^{\pi} \geq F}{\text{Número total de } F^{\pi}}, \tag{4.40}$$

o bien

$$p - val = \frac{(\text{Número de } F^{\pi} \geq F) + 1}{(\text{Número total de } F^{\pi}) + 1}. \tag{4.41}$$

Una justificación teórica del modelo que hemos propuesto es posible realizarla describiendo, a partir del Modelo Lineal General Multivariante, la comparación entre grupos en su notación matricial.

Partimos de una matriz de datos centrados a la que denominaremos $\mathbf{Y}$. La información más relevante para diferentes técnicas de Análisis Multivariante se encuentra en su matriz de covarianzas ($\mathbf{Y}^T\mathbf{Y}$) y en su matriz de productos escalares ($\mathbf{Y}\mathbf{Y}^T$). La matriz distancias, independientemente del tipo que sea, también puede ser obtenida a partir de la matriz de productos escalares como define Gower (1966). Emplearemos también la matriz de diseño $\mathbf{X}$ definida en la ecuación (4.27).

Como hemos mencionado anteriormente, un gran número de cálculos de los Análisis Multivariantes emplean las matrices de productos cruzados $\mathbf{Y}^T\mathbf{Y}$, una de ellas es el cálculo de las sumas de cuadrados, que como también se ha mencionado a lo largo del capítulo, permite el cálculo tradicional de este tipo de técnicas. Si calculamos la traza de la matriz de productos cruzados, es decir, sumamos los valores de su diagonal principal, es posible demostrar que se obtiene la Suma de Cuadrados Total

$$SC_T = tr\left(\mathbf{Y}^T\mathbf{Y}\right). \tag{4.42}$$

Partiendo de la hipótesis particular

$$\hat{\Omega} = \mathbf{B} = \mathbf{0}, \tag{4.43}$$

se empleará el Modelo Lineal General Multivariante que ha sido descrito en la sección 4.2.1 para contrastarla.

Es posible obtener los vectores estimados utilizando el estimador mínimo cuadrático de la ecuación (4.2):

$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{B} = \hat{\mathbf{H}}\mathbf{Y},$$

donde $\hat{\mathbf{H}} = \hat{\mathbf{X}}\left(\hat{\mathbf{X}}^T\hat{\mathbf{X}}\right)^{-1}\hat{\mathbf{X}}^T$. La diferencia entre los datos reales y los estimados permite

calcular la matriz de residuales $\hat{\mathbf{U}} = \mathbf{Y} - \hat{\mathbf{Y}} = \left(\mathbf{I} - \hat{\mathbf{H}}\right)\mathbf{Y}$.

Tanto la matriz de covarianzas como la matriz de productos escalares puede descomponerse en una parte que estará explicada por el modelo y una parte residual. De forma matricial puede escribirse como:

$$\mathbf{Y}^T\mathbf{Y} = \hat{\mathbf{Y}}^T\hat{\mathbf{Y}} + \hat{\mathbf{U}}^T\hat{\mathbf{U}}, \tag{4.44}$$

o

$$\mathbf{Y}\mathbf{Y}^T = \hat{\mathbf{Y}}\hat{\mathbf{Y}}^T + \hat{\mathbf{U}}\hat{\mathbf{U}}^T. \tag{4.45}$$

Como definimos en la (4.38), es posible descomponer las sumas de cuadrados total en dos partes, una parte explicada por el modelo, que corresponderá a las sumas de cuadrados entre grupos, y otra parte residual, que será asociada a la variabilidad existente dentro de los grupos de estudio.

$$SC_T = SC_E + SC_D \tag{4.46}$$

Utilizando la descomposición de la matriz de covarianzas definida en la (4.44) y de forma análoga al caso de las sumas de cuadrados total (ecuación (4.42)), es posible calcular las sumas de cuadrados entre grupos y dentro de los grupos utilizando trazas:

$$SC_E = tr\left(\hat{\mathbf{Y}}^T\hat{\mathbf{Y}}\right), \tag{4.47}$$

$$SC_D = tr\left(\hat{\mathbf{U}}^T\hat{\mathbf{U}}\right). \tag{4.48}$$

De la misma forma que la descomposición de la ecuación (4.46), y utilizando las definiciones de las sumas de cuadrados de las ecuaciones (4.42), (4.47) y (4.48), se puede

afirmar que

$$tr(\mathbf{Y}^T\mathbf{Y}) = tr(\hat{\mathbf{Y}}^T\hat{\mathbf{Y}}) + tr(\hat{\mathbf{U}}^T\hat{\mathbf{U}}), \tag{4.49}$$

de esta forma, es posible definir un pseudo-estadístico de contraste calculando, a partir de las trazas y los grados de libertad del modelo, una F:

$$F = \frac{tr\left(\hat{\mathbf{Y}}^T\hat{\mathbf{Y}}\right)/K-1}{tr\left(\hat{\mathbf{U}}^T\hat{\mathbf{U}}\right)/I-K}. \tag{4.50}$$

En esta $pseudo-F$ no se emplean las covarianzas y correlaciones entre las variables o individuos, $\mathbf{C} = \mathbf{I}$ y $\mathbf{M} = \mathbf{I}$, a diferencia de la hipótesis general del Modelo Lineal General Multivariante planteado en la ecuación (4.3). Si solo existe una variable se habla del estadístico empleado en el ANOVA, de la misma forma que en el estadístico F de la ecuación (4.39).

Utilizaremos el bootstrap no paramétrico sobre los individuos para estimar la distribución muestral.

Partiendo la matriz de distancias denominada $\mathbf{\Delta} = (\delta_1, \delta_2, \dots, \delta_{I-1}, \delta_I)$, para todo δ_i vector de distancias del elemento $i = (1, 2, \dots, I-1, I)$, tomaremos una muestra con I elementos elegida al azar que denominaremos $\mathbf{\Delta}^* = (\delta_1^*, \delta_2^*, \dots, \delta_{I-1}^*, \delta_I^*)$. Se recalculará el valor del estadístico con la nueva muestra $\mathbf{\Delta}^*$, cada valor de F será denominada F^π, igual que en la explicación anterior. Este proceso será repetido un número de veces B fijado con anterioridad antes de comenzar el análisis.

Utilizando las formulas (4.40) o (4.41) se calcularán los p-valores asociados al estadístico.

Si en lugar de la matriz de covarianzas empleamos la matriz de productos escalares entre los individuos, es posible obtener la misma descomposición ya que, partiendo de dos matrices $\mathbf{A}$ y $\mathbf{B}$, se cumple que $tr(\mathbf{AB}) = tr(\mathbf{BA})$. De esta forma, si el $tr\left(\mathbf{Y}^T\mathbf{Y}\right) =$

$tr\left(\mathbf{Y}\mathbf{Y}^T\right)$, es posible realizar la misma descomposición del modelo que en la ecuación (4.49), pero empleando la matriz de productos escalares como en (4.45)

$$tr\left(\mathbf{Y}\mathbf{Y}^T\right) = tr\left(\hat{\mathbf{Y}}\hat{\mathbf{Y}}^T\right) + tr\left(\hat{\mathbf{U}}\hat{\mathbf{U}}^T\right). \tag{4.51}$$

Aunque la matriz de datos $\mathbf{Y}$ sea desconocida, es posible realizar esta partición partiendo de la matriz de productos escalares, ya que es posible realizar las estimaciones de la forma

$$\hat{\mathbf{Y}}\hat{\mathbf{Y}}^T = \hat{\mathbf{H}}\left(\mathbf{Y}\mathbf{Y}^T\right)\hat{\mathbf{H}},$$

$$\hat{\mathbf{U}}\hat{\mathbf{U}}^T = \left(\mathbf{I} - \hat{\mathbf{H}}\right)\left(\mathbf{Y}\mathbf{Y}^T\right)\left(\mathbf{I} - \hat{\mathbf{H}}\right).$$

Podemos obtener la matriz de productos escalares, a la que denominaremos $\mathbf{G}$, como ya habíamos mencionado anteriormente, a través de la fórmula propuesta por Gower (1966), partiendo de cualquier matriz de distancias observadas $\mathbf{\Delta} = \left(\delta_{ij}\right)$

$$\mathbf{G} = (\mathbf{I} - \tfrac{1}{I}\mathbf{1}_I\mathbf{1}_I^T)(\tfrac{1}{2}\mathbf{\Delta}^2)(\mathbf{I} - \tfrac{1}{I}\mathbf{1}_I\mathbf{1}_I^T). \tag{4.52}$$

Esta matriz $\mathbf{G}$ sigue manteniendo las propiedades mencionadas anteriormente y puede descomponerse como hemos mencionado en la ecuación (4.51). De esta forma es posible calcular la suma de cuadrados total a través de su traza ($SC_T = tr(\mathbf{G})$). Empleando los desarrollos anteriores, la suma de cuadrados entre grupos puede ser calculada como $SC_E = tr\left(\hat{\mathbf{H}}\mathbf{G}\hat{\mathbf{H}}\right)$ y la suma de cuadrados dentro de los grupos como $SC_D = tr\left[\left(\mathbf{I} - \hat{\mathbf{H}}\right)\mathbf{G}\left(\mathbf{I} - \hat{\mathbf{H}}\right)\right]$ y por lo tanto, se puede obtener el estadístico de contraste de la ecuación (4.53).

$$F = \frac{tr\left(\hat{\mathbf{H}}\mathbf{G}\hat{\mathbf{H}}\right)/K - 1}{tr\left[\left(\mathbf{I} - \hat{\mathbf{H}}\right)\mathbf{G}\left(\mathbf{I} - \hat{\mathbf{H}}\right)\right]/I - K}. \tag{4.53}$$

De nuevo, empleando remuestreos bootstrap para calcular la distribución muestral y los procedimientos descritos anteriormente, es posible obtener el p-valor asociado al estadístico calculado a través de una de las dos ecuaciones ya mencionadas a lo largo del documento (ecuación (4.40) o (4.41)). De acuerdo con McArdle y Anderson (2001), este es un estadístico de tipo III y es adecuado para regresión, MANOVA, MANCOVA e incluso para diseños no balanceados.

4.6.2. Diseños generales

Para generalizar los modelos, será necesario utilizar una hipótesis que incluya la matriz de contrastes **C** que nos permita aislar los modelos. La hipótesis contrastada debe ser

$$\mathbf{\Omega} = \mathbf{CB} = \mathbf{0} \tag{4.54}$$

Se ha suprimido la **M** de la hipóteisis del MLGM planteada en la ecuación (4.3), ya que en este contexto no es de interés realizar el contraste entre las variables.

Para contrastar la hipótesis (4.54), utilizando las ecuaciones (4.5), (4.6) y (4.7), con estadísticos relacionados con las raíces características de $\mathbf{HE}^{-1}$.

De esta forma el estimador de la hipótesis es

$$\hat{\mathbf{\Omega}} = \mathbf{C}\hat{\mathbf{B}} = \mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \tag{4.55}$$

Es posible utilizar una pseudo F que tenga en cuenta el número de grados de libertad de $\mathbf{C}\hat{\mathbf{B}}$ para contrastar la hipótesis.

Para contrastar esta hipótesis es posible utilizar un pseudo estadístico F que tenga en cuenta los grados de libertad de $\mathbf{C}\hat{\mathbf{B}}$, $(V - 1)$:

$$F = \frac{tr\left(\hat{\Omega}\mathbf{R}^{-1}\hat{\Omega}\right)/(V-1)}{tr\left(\hat{\mathbf{U}}^T\hat{\mathbf{U}}\right)/(I-K)}. \tag{4.56}$$

El denominador de este estadístico no depende del planteamiento de la hipótesis, por lo tanto se mantendrá igual que en los casos anteriores (ecuación (4.50)).

Si se trata un único contraste, con una variable y se ha empleado la distancia euclídea, este estadístico corresponde con una t de *Student* al cuadrado con $(I-K)$ grados de libertad, o lo que es lo mismo, una distribución F de *Snedecor* con 1 y $(I-K)$ grados de libertad.

Este tipo de modelos, que incluyen la matriz $\mathbf{C}$, permiten estudiar diseños más complejos, realizar las comparaciones por parejas o aislar efectos.

Para realizar el contraste de la hipótesis (4.55), es posible utilizar la matriz de productos escalares ($\mathbf{G}$), de forma análoga a lo realizado en las hipótesis más sencillas (ecuación (4.43)). Para ello es necesario tener en cuenta que

$$tr\left(\hat{\Omega}^T\mathbf{R}^{-1}\hat{\Omega}\right) = tr\left(\hat{\Omega}^T\mathbf{R}^{-1/2}\mathbf{R}^{-1/2}\hat{\Omega}\right) = tr\left(\mathbf{R}^{-1/2}\hat{\Omega}^T\hat{\Omega}\mathbf{R}^{-1/2}\right) =$$
$$= tr\left(\mathbf{R}^{-1/2}\mathbf{C}\hat{\mathbf{B}}\hat{\mathbf{B}}^T\mathbf{C}^T\mathbf{R}^{-1/2}\right) = tr\left(\mathbf{R}^{-1/2}\mathbf{C}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{C}^T\mathbf{R}^{-1/2}\right) =$$
$$= tr\left(\mathbf{R}^{-1/2}\mathbf{C}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{G}\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{C}^T\mathbf{R}^{-1/2}\right).$$

Por ello, la F puede ser calculada de forma similar a la representada en la ecuación (4.53):

$$F = \frac{tr\left[\mathbf{R}^{-1/2}\mathbf{C}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{G}\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{C}^T\mathbf{R}^{-1/2}\right]/K-1}{tr\left[\left(\mathbf{I}-\hat{\mathbf{H}}\right)\mathbf{G}\left(\mathbf{I}-\hat{\mathbf{H}}\right)\right]/I-K}. \tag{4.57}$$

Igual que hemos realizado en los apartados anteriores, la distribución muestral será calculada a través de boostrap, obteniendo un p-valor asociado de una de las formas descritas en las ecuaciones (4.40) y (4.41).

4.7. Representaciones gráficas

Las representaciones gráficas de los Modelos Lineales Multivariantes son conocidas y de gran interés, en concreto los gráficos asociados al MANOVA. Este tipo de representaciones busca estudiar la dimensionalidad de la hipótesis alternativa. El Análisis Canónico de Poblaciones o Coordenadas Discriminantes permite realizar este tipo de representaciones gráficas. En esta sección se buscará una representación gráfica que pueda ser asociada a los MANOVAs basados en distancias, que son objeto de estudio en este capítulo, permitiendo que se realice un mejor estudio de la hipótesis alternativa.

La matriz de productos escalares, $\mathbf{G}$, se emplea como base en las Coordenadas Discriminantes, esto la convierte en una matriz muy útil para realizar las representaciones gráficas de cualquiera de los MANOVAS basados en distancias que se han recogido en este capítulo.

4.7.1. Análisis de Coordenadas Principales

Partiendo de una matriz de distancias observadas entre individuos $\mathbf{\Delta}$, se calculará la matriz de productos escalares $\mathbf{G}$, utilizando la ecuación definida en (4.52), que permita obtener la configuración de puntos $\mathbf{Z}$ que reproduce los productos escalares, y por lo tanto las distancias entre los individuos de estudio, de la forma más fidedigna posible. El Análisis de Coordenadas Principales, propuesto por Gower (1966), realiza este proceso.

En esta sección 4.7.1 se recogerán las ideas principales de la técnica propuesta por Gower.

Es posible escribir la descomposición en valores y vectores propios de la matriz $\mathbf{G}$ de la forma

$$\mathbf{G} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T, \tag{4.58}$$

donde $\mathbf{V}$ contiene los vectores propios de la matriz $\mathbf{G}$ y $\mathbf{\Lambda}$ es la matriz diagonal que contiene, ordenados de mayor a menor, los valores propios de esta misma matriz $\mathbf{\Lambda} = diag\,(\lambda_1, \lambda_2, \dots, \lambda_I)$.

Existirá una configuración euclídea que permita reproducir las distancias observadas en una dimensión $(I-1)$, si la matriz de vectores propios $\mathbf{V}$ es definida positiva, y por lo tanto, todos los valores propios son no negativos.

Es posible calcular una matriz $\mathbf{Z}$ que contenga las coordenadas buscadas

$$\mathbf{Z} = \mathbf{V}\mathbf{\Lambda}^{1/2}. \tag{4.59}$$

Las variabilidad del modelo se encuentra en la matriz $\mathbf{Z}$, que se ordenará de forma decreciente, de la misma forma que lo hacían los valores propios. Esto permite elegir solo las primeras columnas para realizar una representación en dimensión reducida. Utilizando los valores propios, es posible realizar el calculo de la cantidad de variabilidad recogida por las A primeras coordenadas utilizando la fórmula

$$\frac{\sum_{j=1}^{A} \lambda_j}{\sum_{j=1}^{I-1} \lambda_j}.$$

Es posible realizar este proceso con cualquier distancia de las definidas en la sección 4.4, pero si la distancia no es euclídea, es posible que existan valores negativos y/o de pequeña magnitud. Además, si la distancia utilizada es la distancia euclídea, estas coor-

denadas coincidirán con las de las Coordenadas Principales.

Para solventar los valores negativos o de pequeña magnitud en la matriz de productos escalares calculada a partir de las distancias existen en la literatura multitud de alternativas. En este trabajo emplearemos la opción más sencilla. Se realizará la aproximación a la matriz semidefinida positiva más próxima sustituyendo los valores negativos por ceros para reconstruir la matriz **G**.

Usar el Análisis de Coordenadas Principales (ACoA) después de realizar cualquiera de los dos MANOVAs basados en distancias descritos a lo largo de este capítulo, equivaldría a realizar un Análisis de Componentes Principales (ACP)sobre el MANOVA convencional. El ACP busca encontrar las direcciones que maximizan la variabilidad total, independientemente de si la variabilidad entre grupos lo es o no, si tras realizar un MANOVA tradicional, nos interesa buscar las direcciones que hacen que la variabilidad entre los grupos sea máxima en relación a la variabilidad dentro de ellos, debe ser sustituido el ACP por el Análisis Canónico. Para los MANOVAs basados en distancias Gower y Krzanowski (1999) propone, para recoger mejor las diferencias entre grupos, realizar un ACoP sobre los centroides, aunque no las tenga en cuenta; este fue el mismo objetivo que tuvieron Anderson y Willis (2003) al proponer la realización de una Análisis Canónico sobre las Coordenadas Principales.

4.7.2. Análisis de Coordenadas Principales de la matriz de medias

Para realizar el Análisis de Coordenadas Principales de la matriz de medias debemos partir de la matriz de distancias entre los centroides a la que denominaremos $\bar{\mathbf{\Delta}}$. Gower y Krzanowski (1999) describen como calcular esta matriz a partir de $\mathbf{\Delta}$, $\mathbf{D}_K$ y $\mathbf{X}$

$$\bar{\mathbf{\Delta}} = \mathbf{D}_K^{-1}\mathbf{X}^T\mathbf{\Delta}\mathbf{X}\mathbf{D}_K^{-1}. \tag{4.60}$$

A continuación, realizaremos un ACoP, descrito en la sección 4.7.1, de la matriz de distancias a los centroides calculada en el paso anterior.

En primer lugar, realizaremos los cálculos necesarios para obtener la matriz de productos escalares partiendo de la matriz de distancias a los centroides

$$\bar{\mathbf{G}} = \left(\mathbf{I} - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^T\right)\left(\frac{1}{2}\bar{\mathbf{\Delta}}^2\right)\left(\mathbf{I} - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^T\right). \tag{4.61}$$

En segundo lugar, se debe obtener la descomposición en valores y vectores propios

$$\bar{\mathbf{G}} = \bar{\mathbf{V}}\bar{\mathbf{\Lambda}}\bar{\mathbf{V}}^T, \tag{4.62}$$

donde $\bar{\mathbf{V}}$ contiene la matriz de vectores propios y $\bar{\mathbf{\Lambda}}^{1/2}$ la matriz diagonal que contiene los valores propios ordenados de forma descendiente, igual que en el caso anterior.

El último paso será realizar los cálculos para obtener las Coordenadas Principales

$$\bar{\mathbf{Z}} = \bar{\mathbf{V}}\bar{\mathbf{\Lambda}}^{1/2}. \tag{4.63}$$

Es posible emplear una versión ponderada que tenga en cuenta los tamaños muestrales de los grupos que, aunque no se encuentran explicados en este documento, pero sí fue descrita por los mismos autores.

Gower (1968) también propuso una fórmula para realizar la representación de los individuos iniciales sobre el gráfico calculado con las medias.

Suponemos que es posible calcular la matriz de distancias de cada punto de la matriz original $\mathbf{Y}$ a las medias de cada grupo, dicha matriz será denominada $\mathbf{\Delta}_{\bar{\mathbf{Y}}}$ de dimensión

$I \times K$ ya que tiene las distancias de los I individuos a los centroides de los K grupos.

A partir de las matrices descritas en esta sección 4.7.2 y los elementos de la diagonal principal de la matriz $\bar{\mathbf{G}}$, contenidos en el vector $\bar{\mathbf{g}}$, es posible obtener las coordenadas de los individuos de la matriz original en la representación de las medias, $\mathbf{Z}_{\bar{Z}}$, empleando la siguiente fórmula

$$\mathbf{Z}_{\bar{Z}} = \frac{1}{2}\left(\mathbf{1}_I \bar{\mathbf{g}} - \Delta_{\bar{Y}}\right)\bar{\mathbf{Z}}\bar{\Lambda}^{-2}. \tag{4.64}$$

Si la matriz de datos original tiene datos continuos, es posible calcular los centroides y, por lo tanto las distancias de los individuos a dichos puntos. Sin embargo, cuando los datos son binarios, los centroides pueden no ser un vector de datos binarios, y por lo tanto, esto imposibilitaría el cálculo de las distancias a los centroides que permiten la realización de representación gráfica sobre el gráfico creado a partir de la matriz de las medias. Una posible solución para esta problemática sería calcular las Coordenadas Principales en la dimensión completa, obtener los centros y, partiendo de ellos, las distancias.

4.7.3. Regiones de confianza bootstrap para los centroides

Basándonos en el Análisis Canónico, crearemos regiones de confianza para los centroides, pero ahora basadas en bootstrap. Para mostrar la región de confianza usaremos los datos perturbados que muestren la variabilidad de los centroides. Estas regiones se utilizan para hacer contrastes aproximados de comparación de parejas de medias, utilizándose para comprobar la estructura de la hipótesis alternativa. Para la versión clásica del Análisis Canónico existe un procedimiento basado en bootstrap que puede encontrase en Duarte *et al.* (1998). Las regiones obtenidas con este procedimiento serán similares a las obtenidas en el artículo de Amaro *et al.* (2008).

A diferencia de los procedimientos bootstrap anteriores, haremos el remuestreo bajo

el supuesto de que hay diferencias entre los grupos, hasta ahora los remuestreos los hemos realizado bajo la hipótesis nula de que todos los grupos son iguales. Los remuestreos bootstrap serán realizados dentro de cada uno de los grupos.

Dividiremos entonces los individuos de la matriz de distancias $\mathbf{\Delta}$, que hemos utilizado hasta el momento y que contiene las distancias de los I individuos, en K grupos. De esta forma, la matriz de distancias es ahora

$$\mathbf{\Delta} = \left(\delta_{1(1)}, \dots, \delta_{I_1(1)}, \dots, \delta_{1(k)}, \dots, \delta_{I_k(k)}, \dots, \delta_{1(K)}, \dots, \delta_{I_K(K)}\right).$$

Ahora, tomaremos muestras al azar con remplazamiento dentro de los grupos para obtener

$$\mathbf{\Delta}_b^* = \left(\delta^*_{1(1)}, \dots, \delta^*_{I_1(1)}, \dots, \delta^*_{1(t)}, \dots, \delta^*_{I_k(k)}, \dots, \delta^*_{1(K)}, \dots, \delta^*_{I_K(K)}\right)$$

siendo $b = 1, \dots, B$ y B el número de muestras bootstrap tomadas.

Las distancias entre las medias igual que habíamos realizado en (4.60),

$$\bar{\mathbf{\Delta}}_b^* = \mathbf{D}_K^{-1}\mathbf{X}^T\mathbf{\Delta}_b^*\mathbf{X}\mathbf{D}_K^{-1}, \tag{4.65}$$

y, de la misma forma que en (4.61), la matriz de productos escalares será

$$\bar{\mathbf{G}}_b = \left(\mathbf{I} - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^T\right)\left(\frac{1}{2}\left(\bar{\mathbf{\Delta}}_b^*\right)^2\right)\left(\mathbf{I} - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^T\right). \tag{4.66}$$

Igual que realizamos en (4.62) es posible realizar la descomposición en valores y vectores propios de la matriz de productos escalares calculada

$$\bar{\mathbf{G}}_b = \bar{\mathbf{V}}_b\bar{\mathbf{\Lambda}}_b\bar{\mathbf{V}}_b^T, \tag{4.67}$$

y como en (4.63) calcular las coordenadas principales para los centroides

$$\bar{\mathbf{Z}}_b = \bar{\mathbf{V}}_b\bar{\mathbf{\Lambda}}_b^{1/2}. \tag{4.68}$$

Cada una de las replicas bootstrap realizadas sobre las coordenadas de las medias de los grupos representados se encuentran en las matrices $\bar{\mathbf{Z}}_b$.

Las coordenadas descritas anteriormente pueden presentar algunos problemas, ya que es posible encontrar variabilidad añadida por el método empleado para la obtención de los valores y vectores propios, que debe ser tenida en cuenta antes de realizar la representación gráfica. Es posible que la variabilidad de las replicas de las coordenadas de los puntos sean importantes incluso cuando los valores ajustados son parecidos. Esta variabilidad se puede deber a diversos factores:

- Debido a que los vectores propios son únicos salvo el signo, es posible que exista una reflexión de los ejes.
- Puede ocurrir una rotación de los vectores propios, aunque se mantenga el espacio bidimensional que ha sido generado.
- Cuando los valores propios tienen magnitudes similares, es probable que se produzca una inversión del orden de dichos valores.
- La variabilidad de las remuestras puede ser más pequeña que en la muestra original al producirse las repeticiones, esto provoca que pueda existir una compresión de los valores.

Es posible encontrar un desarrollo ampliado de los problemas que se pueden plantear en este tipo de representaciones en Milan y Whittaker (1995).

De los errores planteados en el documento, el primero puede ser solucionado calculando el producto escalar de cada uno de los vectores propios de la matriz inicial con su correspondiente de la replica bootstrap, si el producto escalar es negativo realizar un cambio de signo. El Análisis de Procrustes permiten corregir el resto de las situaciones

planteadas, aunque son problemas un poco más complejos.

Procrustes

Partimos de dos configuraciones de puntos, en este caso tomaremos para los datos originales la configuración de puntos inicial $\bar{\mathbf{Z}}$ obtenida de las coordenadas principales de las medias (ecuación (4.63)) y para las medias en una replica bootstrap la configuración $\bar{\mathbf{Z}}_b$. Suponemos que ambas están centradas evitando el problema de la traslación en el Análisis Procrustes.

Mantendremos la primera como fija y realizaremos las transformaciones sobre la segunda buscando conseguir la configuración que más se aproxime a la fijada, es decir a nuestra configuración inicial. Para ello será necesario obtener una nueva matriz $\bar{\mathbf{J}}_b$, definida como

$$\bar{\mathbf{J}}_b = \mathbf{t}_b \bar{\mathbf{Z}}_b \mathbf{T}_b,$$

donde $\mathbf{t}_b$ será una constante y $\mathbf{T}_b$ una matriz ortogonal, que hagan mínima la discrepancia entre $\bar{\mathbf{Z}}$ y $\bar{\mathbf{Z}}_b$. La matriz $\bar{\mathbf{Z}}_b$ será rotada y re-escalada haciendo que su coincidencia con $\bar{\mathbf{Z}}$ sea máxima.

Para calcular $\mathbf{t}_b$ y $\mathbf{T}_b$ será necesario describir la matriz $\mathbf{K}_b$ como

$$\mathbf{K}_b = \bar{\mathbf{Z}}_b^T \bar{\mathbf{Z}},$$

y su correspondiente descomposición en valores singulares

$$\mathbf{K}_b = \mathbf{P}_b \mathbf{\Lambda}_b \mathbf{Q}_b^T.$$

Así $\mathbf{T}_b$ puede definirse como

$$\mathbf{T}_b = \mathbf{P}_b \mathbf{Q}_b^T,$$

y $\mathbf{t}_b$ de la forma

$$\mathbf{t}_b = \frac{tr\left(\mathbf{T}_b^T \bar{\mathbf{Z}}_b \left(\mathbf{I} - I^{-1}\mathbf{1}\mathbf{1}^T\right) \bar{\mathbf{Z}}\right)}{tr\left(\bar{\mathbf{Z}}_b^T \left(\mathbf{I} - I^{-1}\mathbf{1}\mathbf{1}^T\right) \bar{\mathbf{Z}}\right)}.$$

Tras esta modificación, podemos suponer que ambas configuraciones son comparables y pueden ser representadas sobre el mismo espacio. Para obtener la réplica de la configuración sustituiremos entonces $\bar{\mathbf{Z}}_b \Leftarrow \bar{\mathbf{J}}_b$.

Para cada una de las coordenadas de los grupos, obtenemos entonces un conjunto de réplicas. Este conjunto de coordenadas puede ser representado como una elipse de concentración no paramétrica o una envolvente convexa de todos los puntos. de Leeuw y Meulman (1986) presenta un procedimiento para realizar el cálculo de la elipse correspondiente.

4.8. Software para los MANOVAs basados en distancias

Desde la creación del PERMANOVA se han desarrollado diversos software que permiten realizar los cálculos necesarios para aplicar la técnica a diversos tipos de datos.

Inicialmente se emplea el software desarrollado por Anderson (2005), este programa ha sido integrado dentro del software PRIMER-e explicado en la sección 4.8.1. Otro de los programas estadísticos que integra un módulo para el cálculo del PERMANOVA es PAST 4.8.2.

Dentro del software de código abierto R también existen varios paquetes que integran estos cálculos, desarrollados en la sección 4.8.3, los paquetes vegan y pairwiseAdonis contienen únicamente el PERMANOVA, sin embargo, el paquete PERMANOVA contiene toda la información desarrollada en este capítulo.

4.8.1. PRIMER-e (Clarke *et al.*, 2017)

Este software no contiene en su versión estándar el PERMANOVA, debe utilizarse un paquete adicional al programa para poder realizar sus cálculos.

El paquete PRIMER contiene un gran número de Análisis Multivariantes No Paramétricos que emplean métodos de permutación robustos para hacer sus cálculos. En la última versión se incluye también la posibilidad de realizar análisis multivía.

El paquete adicional PERMANOVA, derivado del software original desarrollado por Anderson, permite realizar la técnica que da nombre al paquete, el PERMANOVA, con modelos complejos, además de algunas técnicas complementarias que pueden ser de interés, como el Análisis Discriminante.

4.8.2. PAST (Hammer *et al.*, 2001)

PAST es un programa que solo ha sido desarrollado para equipos con sistema operativo Windows. Es un paquete gratuito con análisis univariantes y multivariante orientados, mayoritariamente, a datos del campo de la Ecología.

Este software, igual que el anterior, contiene una interfaz gráfica que facilita al usuario la realización de numerosos análisis multivariantes y sus gráficos asociados. Una de las características que se destacan de este software es la realización de gráficos en 3D.

El PERMANOVA tanto de una vía como de dos se podrá encontrar en el Menú que contiene el resto de Análisis Multivatiantes.

4.8.3. R (R Core Team, 2021)

R es un software libre en el que existen un gran número de paquetes que permiten la realización de multitud de técnicas estadísticas y representaciones gráficas.

Para la realización de las técnicas presentadas en este capítulo, también existen varios paquetes que permiten realizar los cálculos del PERMANOVA, sin embargo, debido a que la técnica BOOTMANOVA y las representaciones gráficas asociadas a ellas se han presentado en trabajos recientes, se ha construido un paquete que contenga estas técnicas.

vegan (Oksanen *et al.*, 2017)

Este paquete, que ya hemos mencionado en el capítulo de biplot (3.4), permitirá la realización de pruebas no paramétricas multivariantes, entre ellas el PERMANOVA.

La función que debemos utilizar será "`adonis2`". Esta función requerirá el modelo que queremos contrastar, las respuestas estarán en un data frame y los predictores en otro. Será posible elegir la distancia y el número de permutaciones que vamos a utilizar.

Con este paquete no será posible incluir un gran número de variables predictoras ni realizar los contrastes a posteriori.

pairwiseAdonis

Este paquete será complementario al anterior, se encuentra en GitHub, y permitirá realizar los contrastes a posteriores de los resultados de la función "`adonis2`".

PERMANOVA (Vicente-Gonzalez y Vicente-Villardon, 2021)

Durante este trabajo se ha desarrollado este paquete, aunque se están implementando mejoras que no han sido incluidas anteriormente.

Con él será posible realizar los MANOVAs basados en distancias presentadas en este capítulo. Podremos calcular las distancias recogidas en la sección 4.4, los MANOVAs recogidos en la sección 4.3, se podrán realizar los cálculos tanto del PERMANOVA, de la sección 4.5, como del BOOTMANOVA de la sección 4.6 con los modelos simples y los más complejos, y por lo tanto, será posible realizar los contrastes a posteriori.

Las representaciones gráficas descritas en la sección 4.7 también están incluidas dentro del paquete.

4.9. Ejemplo MANOVA basado en distancias

Este tipo de técnicas, como ha quedado resaltado a lo largo de todo el capítulo, son de gran utilidad en un gran número de contextos en los que los datos recogidos no siguen una distribución normal, como ocurre en muchos casos del campo de la Eclología, o el número de individuos es menor que el número de variables, como ocurre en un gran número de casos en los que se trabaja con datos genéticos, por ejemplo.

En esta sección vamos a presentar dos ejemplos que corresponden con estas situaciones, aunque puede ser aplicado en un gran número de contextos. Se han elegido estos dos casos ya que se trata de datos de respuesta binaria, presencia o ausencias de las variables de estudio. En ambos ejemplos se utilizarán datos genómicos, que en la actualidad tienen un gran interés para la ciencia, principalmente en el ámbito biológico, en concreto utilizaremos secuenciación de ARN.

En el primero de los casos, relacionado con la biología, vamos a trabajar con la pre-

sencia o ausencia del hongo *Colletotrichum graminicola* en plantas de maíz de diferentes países, extraídos de un experimento real realizado en un proyecto del Instituto de Investigación en Agrobiotecnología (CIALE) con el que estamos colaborando (sección 4.9.1).

El segundo ejemplo que vamos a presentar corresponde con la presencia o ausencia de SNP del cromosoma 10 para comparar 10 grupos raciales. Los datos están asociados a la fase III de un proyecto internacional denominado HapMap (sección 4.9.2).

En ambos casos realizaremos tanto el PERMANOVA, como el BOOTMANOVA, y si fuera necesario los contrastes a posteriori.

Para realizar estos análisis utilizaremos el software estadístico R (R Core Team, 2021), en concreto, el paquete PERMANOVA (Vicente-Gonzalez y Vicente-Villardon, 2021), desarrollado para ejecutar los cálculos de este tipo de análisis.

Existen trabajos en diferentes campos que aplican el PERMANOVA, sin embargo solo Vicente-Gonzalez y Vicente-Villardon (2019) contiene el BOOTMANOVA.

4.9.1. Ejemplo 1. *Colletotrichum graminicola*

Anderson *et al.* (2004) presenta que aproximadamente el 30 % de las enfermedades en plantas se deben a hongos fitopatógenos influyendo notablemente tanto en el ámbito de la Ecología como en el de la Agricultura. En este mismo artículo es posible encontrar algunos ejemplos en el ámbito ecológico que afectan a especies que han sido infectadas por hongos atacando la supervivencia de la especie. Desde el punto de vista de la agricultura, la infección por hongos puede provocar importantes perdidas económicas en los productores.

En este trabajo vamos a presentar un hongo concreto, el *Colletotrichum graminicola* (o Glomerella graminícola). Este tipo de hongo puede infectar a un gran número de

plantas, sin embargo uno de los hospedadores principales son las plantas del maíz. En esta planta provoca una enfermedad denominada antracnosis.

La antracnosis produce unas manchas, similares a las manchas de humedad, principalmente en las hojas y tallos de la planta que infecta. Si no se elimina el hongo y controla la enfermedad puede llegar a producir que las plantas o tejidos infectados se marchiten y mueran. Existen algunos estudios como el de O'Connell *et al.* (2012), que recoge el impacto económico que tuvo en EE.UU..

Su genoma se encuentra organizado en un total de 13 cromosomas, 3 de los cuales son minicromosonas con un tamaño menor de 1 Mb; el tamaño total del genoma es de 57 Mb.

Este primer ejemplo, como hemos mencionado en la introducción de la aplicación práctica, pertenece a un proyecto realizado por el CIALE, sobre 9 tipos de cepas de la planta del maíz infectadas con el hongo *Colletotrichum graminicola*.

Base de datos

La base de datos utilizada contendrá un total de 103 cepas diferentes, que están divididas en 9 tipos según su país de procedencia:

- Argentina (AR)
- Brasil (BR)
- Canadá (CA)
- Croacia (HR)
- Eslovenia (SI)
- Francia (FR)
- Portugal (PT)

- Suiza (CH)
- EE.UU. (US)

De la secuenciación del ARN del hongo *Colletotrichum graminicola* se obtienen 13183 variables, sin embargo el método de secuenciación genera que los datos recuperados tenga la información por duplicada, por ello eliminaremos aquellas variables cuya presencia era 0, de esta forma reducimos el número de variables a 6419.

Objetivos del ejemplo 1

Los objetivos de nuestros ejemplo serán:

Objetivo 1. Estudiar si existen diferencias significativas entre las cepas de maíz sometidas a estudio.

Objetivo 2. Analizar, si fuera pertinente, entre qué grupos de cepas existen diferencias significativas.

Objetivo 3. Crear agrupaciones entre los tipos de cepas sometidas a estudio.

Objetivo 4. Extraer las variables con mayor relevancia para cada uno de los grupos.

Metodología

Como mencionábamos en la introducción de esta sección, este ejemplo permitirá ilustrar las técnicas presentadas en el presente capítulo.

Por ello, emplearemos en primer lugar el Análisis PERMANOVA y BOOTMANOVA para realizar los contrastes generales y el análisis Post-Hoc en caso de que sea necesario, utilizaremos la corrección de Bonferroni. A continuación, realizaremos el Análisis de Coordenadas Principales y un Análisis de Coordenadas Principales sobre los centroides con sus regiones de confianza bootstrap asociadas. Por último, terminaremos realizando su representación biplot asociada.

Resultados

En primer lugar, realizaremos el análisis PERMANOVA propuesta por McArdle y Anderson (2001) para buscar diferencias entre los tipos de cepas del maíz. La matriz de distancias de partida se calculará empleando el índice de similaridad denominado "Concordancia Simple". Se utilizarán 10000 permutaciones y se extraerán 4 dimensiones para la posterior representación gráfica. Los valores para el contraste general es posible encontrarlas en la tabla 4.2.

Explicada	Residual	G.L. Num	G.L. Denom	F-exp	p-valor
5.586306	9.092764	8	94	7.218828	0.00049975

Tabla 4.2: Resultados PERMANOVA de las especies del tipo de cepa de maíz

A continuación, realizamos el análisis BOOTMANOVA propuesto recientemente por Vicente-Gonzalez y Vicente-Villardon (2019) con el mismo objetivo. Igual que en el PERMANOVA emplearemos 10000 réplicas bootstrap y extraeremos 4 dimensiones. Los resultados del contraste general se pueden encontrar en la tabla 4.3.

Explicada	Residual	G.L. Num	G.L. Denom	F-exp	p-valor
5.586306	9.092764	8	94	7.218828	9.999e-05

Tabla 4.3: Resultados BOOTMANOVA de las especies del tipo de cepa de maíz

Con ambas técnicas observamos que existen diferencias altamente significativas entre los tipos de cepas de maíz. La diferencia entre los resultados de ambas técnicas es pequeña, sin embargo el BOOTMANOVA puede mejorar los resultados del PERMANOVA en algunos casos concretos.

Puede ser de interés en este tipo de trabajos estudiar si alguna de las cepas se ha infectado menos con el hongo, para ello podemos utilizar los contrastes para cada uno

de las cepas de maíz. Para ello será necesario construir la matriz de contrastes C de la forma:

$$\mathbf{C} = \begin{pmatrix} AR & BR & CA & HR & SI & FR & PT & CH & US & \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & AR \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & BR \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & CA \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & HR \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & SI \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & FR \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & PT \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & CH \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & US \end{pmatrix}$$

Los resultados son iguales en ambas técnicas, así que en la tabla 4.4 se recogen los resultados obtenidos para el PERMANOVA.

	Explained	Residual	df Num	df Denom	F-exp	p-value	p-value adj.
C Argentina	0.2400687	8.939156	1	94	2.524450	0.0017998	0.016
C Brasil	1.7310059	8.939156	1	94	18.202452	0.0001000	0.001
C Canadá	0.8370673	8.939156	1	94	8.802210	0.0001000	0.001
C Croacia	0.3853276	8.939156	1	94	4.051925	0.0001000	0.001
C Francia	0.4423493	8.939156	1	94	4.651539	0.0001000	0.001
C Portugal	0.1734589	8.939156	1	94	1.824013	0.0180982	0.163
C Eslovenia	0.2213611	8.939156	1	94	2.327731	0.0033997	0.031
C Suiza	0.5305575	8.939156	1	94	5.579096	0.0001000	0.001
C EE.UU.	0.9226418	8.939156	1	94	9.702072	0.0001000	0.001

Tabla 4.4: Resultados de los contrastes para cada tipo de cepa de maíz

Podemos destacar la cepa de Portugal, es la única que no es significativa. El resto de

contrastes son altamente significativos excepto la cepa de Argentina y la de Eslovenia que solo son estadísticamente significativas.

Ya que existen diferencias estadísticamente significativas entre los grupos, es de interés saber entre qué tipos de cepas se presentan estas diferencias. Se ha recogido todas las combinaciones en la tabla 4.5.

En la mayor parte de las combinaciones se encuentran diferencias altamente significativas entre las cepas, aunque destaca, por ejemplo la cepa de Argentina que no presenta diferencias significativas con la mitad de las cepas, Croacia, Francia, Portugal y Eslovenia. Además de con la cepa de Argentina, la de Croacia no presenta diferencias significativas con la de Portugal y con la de Eslovia. La variedad francesa no presenta diferencias significativas con las cepas portuguesas y suizas. Por último, la variedad de Portugal, además de con los cepas de los países ya mencionados, no presenta diferencias con la de Eslovenia.

	Explained	Residual	df Num	df Denom	F-exp	p-value	p-value adj.
AR-BR	0.2455091	8.939156	1	94	2.5816596	0.0001000	0.004
AR-CA	0.1622990	8.939156	1	94	1.7066610	0.0001000	0.004
AR-HR	0.0378994	8.939156	1	94	0.3985330	0.0206979	0.745
AR-FR	0.0417259	8.939156	1	94	0.4387700	0.0119988	0.432
AR-PT	0.0328306	8.939156	1	94	0.3452314	0.0507949	1.000
AR-SI	0.0403269	8.939156	1	94	0.4240587	0.0126987	0.457
AR-CH	0.0644330	8.939156	1	94	0.6775476	0.0005000	0.018
AR-US	0.1821381	8.939156	1	94	1.9152796	0.0001000	0.004
BR-CA	0.4108316	8.939156	1	94	4.3201143	0.0001000	0.004
BR-HR	0.2920548	8.939156	1	94	3.0711124	0.0001000	0.004
BR-FR	0.2902573	8.939156	1	94	3.0522110	0.0001000	0.004
BR-PT	0.2455235	8.939156	1	94	2.5818104	0.0001000	0.004
BR-SI	0.2545319	8.939156	1	94	2.6765389	0.0001000	0.004
BR-CH	0.2981651	8.939156	1	94	3.1353649	0.0001000	0.004
BR-US	0.4225859	8.939156	1	94	4.4437161	0.0001000	0.004
CA-HR	0.1781262	8.939156	1	94	1.8730923	0.0001000	0.004
CA-FR	0.2103640	8.939156	1	94	2.2120895	0.0001000	0.004
CA-PT	0.1342767	8.939156	1	94	1.4119917	0.0001000	0.004
CA-SI	0.1410030	8.939156	1	94	1.4827217	0.0001000	0.004
CA-CH	0.2103417	8.939156	1	94	2.2118559	0.0001000	0.004
CA-US	0.0812981	8.939156	1	94	0.8548931	0.0001000	0.004
HR-FR	0.0610680	8.939156	1	94	0.6421630	0.0009999	0.036
HR-PT	0.0531519	8.939156	1	94	0.5589208	0.0014999	0.054
HR-SI	0.0458715	8.939156	1	94	0.4823630	0.0041996	0.151
HR-CH	0.0890184	8.939156	1	94	0.9360765	0.0001000	0.004
HR-US	0.1856924	8.939156	1	94	1.9526546	0.0001000	0.004
FR-PT	0.0400217	8.939156	1	94	0.4208498	0.0142986	0.515
FR-SI	0.0546213	8.939156	1	94	0.5743720	0.0006999	0.025
FR-CH	0.0516402	8.939156	1	94	0.5430241	0.0027997	0.101
FR-US	0.2242653	8.939156	1	94	2.3582694	0.0001000	0.004
PT-SI	0.0394110	8.939156	1	94	0.4144280	0.0144986	0.522
PT-CH	0.0582513	8.939156	1	94	0.6125437	0.0008999	0.032
PT-US	0.1510679	8.939156	1	94	1.5885592	0.0001000	0.004
SI-CH	0.0737736	8.939156	1	94	0.7757687	0.0002000	0.007
SI-US	0.1503182	8.939156	1	94	1.5806766	0.0001000	0.004
CH-US	0.2291435	8.939156	1	94	2.4095661	0.0001000	0.004

Tabla 4.5: Contrastes a Posteriori de los tipos de cepas del maíz

Por lo tanto, se pueden subdividir las cepas en dos conjuntos, por un lado Argentina, Croacia, Francia, Portugal y Eslovenia, y por otro Brasil, Canadá y Estados Unidos, Suiza se encuentra entre los dos conjuntos aunque más próximo al primero.

Realizando la representación gráfica del Análisis de Coordenadas Principales podemos observar que en el grupo de Brasil, Canadá y EE.UU. quedará dividido en dos, por un lado Brasil y por otro las cepas de Canadá y Estados Unidos. Se puede observar esta representación gráfica en la figura 4.3.

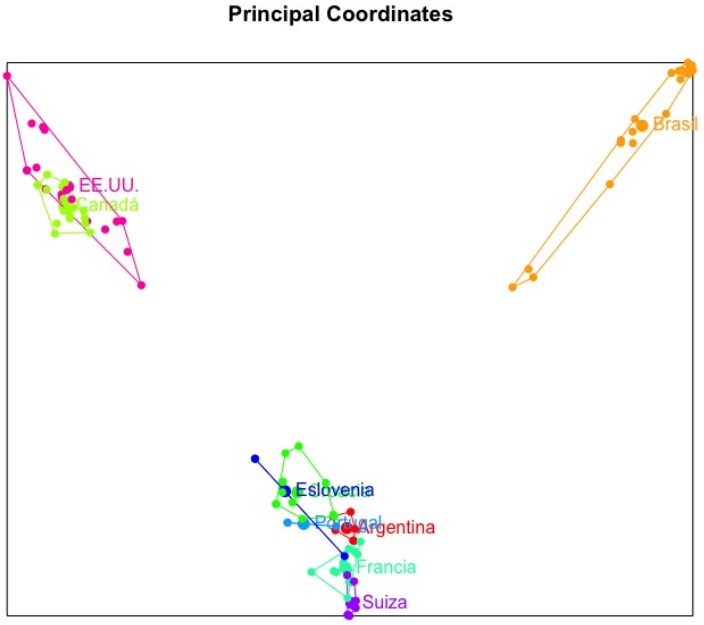

Figura 4.3: Análisis de Coordenadas Principales de los tipos de cepas de maíz

La representación generada como resultado del PERMANOVA y del BOOTMANOVA, si utilizamos las cuatro dimensiones calculadas, se encuentra en la figura 4.4. Esta representación permite observar que se pueden separar algunas cepas más.
El gráfico que se encuentra en la parte superior izquierda, el gráfico a), presenta los mismos grupos que ya se habían mencionado en la figura 4.3, en el que las cepas de Brasil se presentaba separadas del resto de los tipos de cepas.
En la siguiente representación, gráfico b), permite observar como la dimensión 3 diferencia a Eslovenia del resto de países del grupo.
El gráfico c), separará a Suiza y a Portugal, que hasta ahora en todos los gráficos se habían presentado superpuestos, es decir, la dimensión 4 permitirá separar estos dos tipos de cepas.

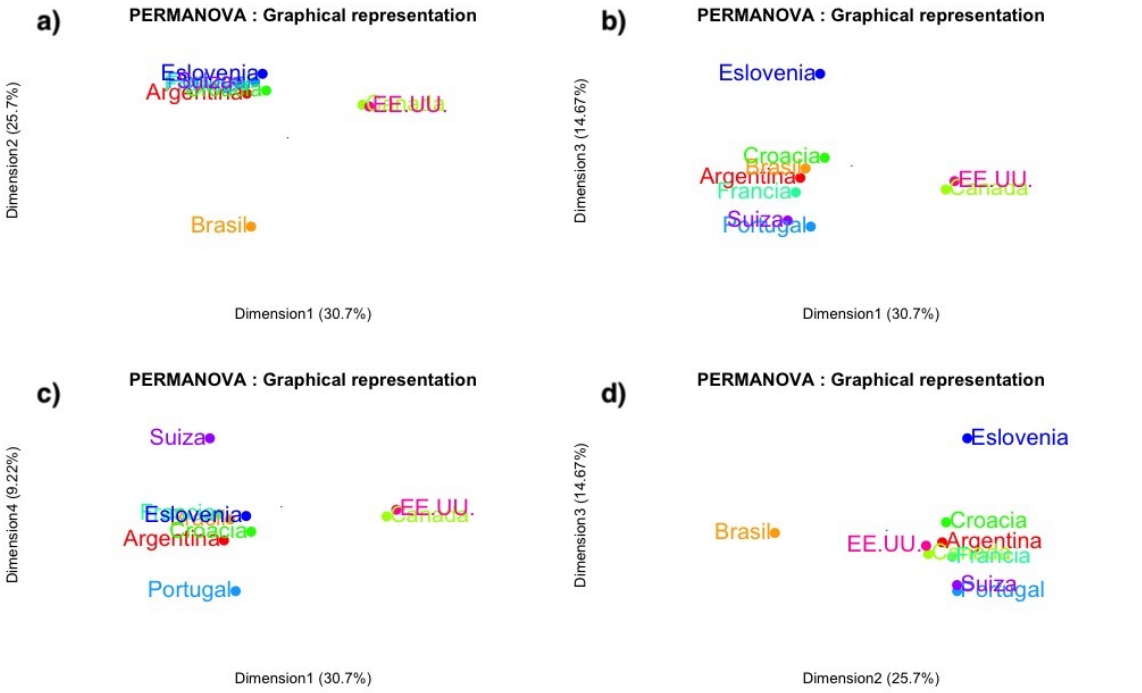

Figura 4.4: Representación gráfica del PERMANOVA en a) las dimensiones 1 y 2, b) las dimensiones 1 y 3, c) las dimensiones 1 y 4 y d) las dimensiones 2 y 3.

Las cepas de EE.UU. y Canadá se mantendrán juntas en todas las representaciones gráficas, igual que ocurre con las cepas de Argentina y Croacia.

Si incluimos las regiones bootstrap descritas en la sección 4.7.2 se pueden observar los resultados a los contrastes que habíamos descrito en la tabla 4.5. Igual que en el caso anterior, se pueden observar diferenciadas las regiones de confianza de los tres grupos especificados anteriormente.

Además este gráfico (figura 4.5) permite ilustrar los contrastes a posteriori. Se puede observar que en aquellos casos en los que habíamos señalado que no existen diferencias significativas las elipses de las regiones de confianza se superponen, por ello en el grupo más grande de circunferencias superpuestas es posible encontrar a Argentina, Croacia, Francia, Portugal y Eslovenia. Sin embargo, Brasil que presenta diferencias altamente significativas con el resto de países se encuentra separada del resto de grupos en tres de

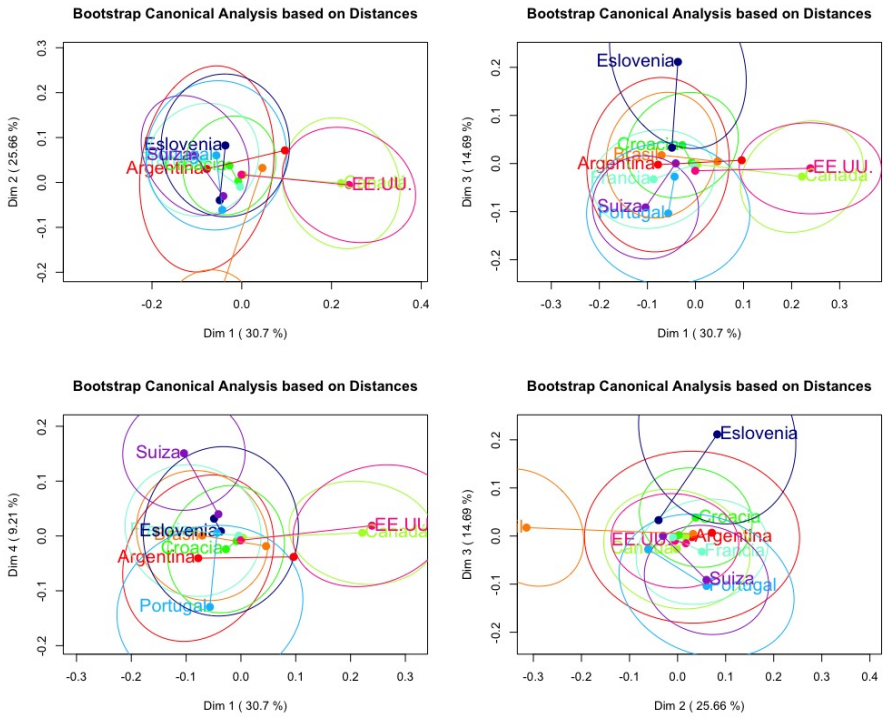

Figura 4.5: Regiones bootstrap para los tipos de cepas de maíz

los cuatro planos presentados.

A continuación, realizaremos las representaciones biplot asociadas. si realizamos un biplot logístico empleando el descenso del gradiente podemos caracterizar a un gran número de grupos. Utilizando una representación del biplot logístico calculado con el método del descenso del gradiente podemos observar algunas de las caracterizaciones de los grupos presentados en los gráficos anteriores. Estas representaciones gráficas serán calculadas únicamente con las variables que sean significativas al cruzarlas con los grupos, es decir 2379 variables, y se proyectarán únicamente aquellas cuya calidad de representación sea mayor del 50 %.

La representación de las dimensiones 1 y 2 muestran las variables que caracterizan los grupos presentados en el PCoA.

Los tipos de maíz de Argentina, Croacia, Eslovenia, Francia, Portugal y Suiza, que

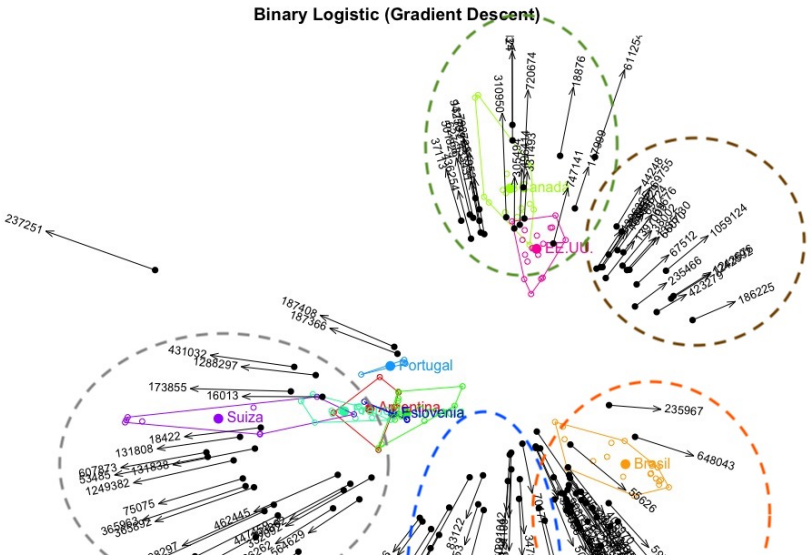

Figura 4.6: Dimensiones 1 y 2 del Biplot Logístico por el método del Descenso del Gradiente

conforman un primer grupo, se caracterizan por tener una probabilidad alta de poseer las variables que se encuentran en la esquina inferior izquierda, englobadas en una elipse gris, y bajas probabilidades de tener las variables de la esquina superior derecha, englobadas en una circunferencia marrón.

Las variables de la elipse verde caracterizarán al segundo grupo, formado por Canadá y Estados Unidos. El grupo 2 será altamente probable que tenga las variables de la elipse verde, mientras que será poco probable que se expresen los genes de la circunferencia azul.

El tercer grupo, que solo lo forma Brasil, estará caracterizado por las altas probabilidades de que se presenten las variables que están dentro de la circunferencia naranja.

Utilizando el plano formado por las dimensiones 2 y 3 y el construido a partir de las dimensiones 2 y 4, es posible caracterizar mejor el grupo 2. Se puede observar en le gráfico 4.7. En ambos gráficos, en el lado derecho se observan las variables que es altamente probable que se expresen en los tipos de cepas de maíz de los países de Canadá y Estados Unidos.

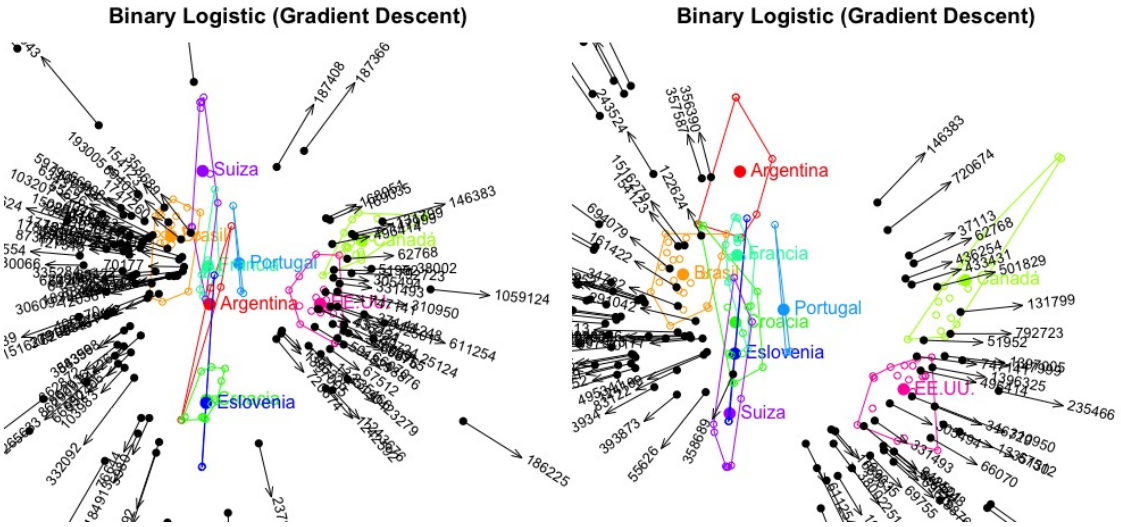

Figura 4.7: Dimensiones 2 y 3 y 2 y 4 del Biplot Logístico por el método del Descenso del Gradiente

Conclusiones

- Se ha comprobado que se presentan diferencias altamente significativas entre los tipos de cepas sometidas a estudio tanto con el análisis PERMANOVA, como con el BOOTMANOVA.

- Presentan diferencias significativas entre casi todos los tipos de cepas excepto entre Argentina y Croacia, Francia, Portugal y Eslovenia; entre Croacia y Portugal y Eslovenia; entre Francia y Portugal y Suiza; y entre Portugal y Eslovenia. Todas ellas pueden ser observadas en la representación gráfica asociada en la que se incluyen regiones de confianza bootstrap.

- Se pueden describir tres grandes grupos de tipos de cepas, el primero estará formado por Argentina, Croacia, Eslovenia, Francia, Portugal y Suiza, el segundo por Canadá y Estados Unidos, y el tercero y último, unicamente por Brasil.

- Se han agrupado las variables proyectadas sobre el gráfico con una calidad de representación mayor del 50 % en 5 grupos que caracterizan cada uno de los tres conjuntos de tipos de cepas presentados anteriormente.

4.9.2. Ejemplo 2. Proyecto HapMap

Como ya se mencionaba en la introducción de esta sección, uno de los ámbitos científicos que mayor cantidad de datos genera es aquel que estudia la genética de los seres vivos. En este caso vamos a estudiar directamente la propia cadena de ADN, utilizando las cadenas complementadas directamente en el microarray.

La secuencia de bases de la cadena de ADN entre dos personas se parece en un 99,9 %, el 0,1 % restante es el que permite la existencia de diferencias en el color del cabello, de los ojos o de la piel, pero no únicamente estas diferencias, también afecta al riesgo del individuo a padecer una enfermedad o la posibilidad de tener algunas de ellas, o al grupo sanguíneo al que pertenece.

Dentro de la secuencia de nucleótidos del ADN existen puntos donde la secuencia cambia en una única base, los nucleóticos donde esto ocurre se dice que sufren un polimorfismo mononucleotídico o SNP (siglas del nombre inglés "Single Nucleotide Polymorphism"). Un patrón de un gran número de SNPs se constituyen como un único bloque que recibe el nombre de haplotipo y, generalmente, la probabilidad de que exista una recombinación genética dentro de un haplotipo es muy baja por lo cual, de forma general, se hereda la secuencia de SNPs completa.

En este ejemplo emplearemos la base de datos de la Fase III del proyecto HapMap, realizando los análisis con el software estadístico R (R Core Team, 2021), en concreto con los paquetes PERMANOVA (Vicente-Gonzalez y Vicente-Villardon, 2021) y MultBiplotR (Vicente-Villardon, 2021).

Más información sobre este proyecto se puede obtener en los artículos asociados al HapMap en la bibliografía (International HapMap Consortium y others, 2005; McVean *et al.*, 2005).

También existen algunos artículos que presentan su disconformidad con este proyec-

to (Terwilliger y Hiekkalinna, 2006).

La utilización de los datos generados por este proyecto ha sido muy diversa (Manolio *et al.*, 2008; Bell *et al.*, 2011; McVean *et al.*, 2005; Deloukas y Bentley, 2004; Gitschier, 2009; Thorgeirsson *et al.*, 2008; Smyth *et al.*, 2006), algunas de las aplicaciones más habituales es la búsqueda de genotipos o SNP asociados a enfermedades.

Base de datos

El HapMap es un proyecto que comenzó en una reunión en el año 2002 con la intención de muestrear una serie de individuos y generar un mapa de haplotipos (Haplotypes Map) del genoma humano. El estudio fue realizado por centro de investigación de cinco países diferentes (Reino Unido, Canadá, Japón, China, Nigeria y Estados Unidos) y el proyecto fue dividido en tres etapas diferentes.

Fase I: Sus resultados son publicados en el año 2005. La muestra recogida en esta primera etapa fue de 269 personas divididas entre Nigeria, Japón, China y Estados Unidos (diferenciando el linaje de ascendencia) en representación de la población mundial. Se encuentran más de un millón de resultados.

Fase II: Los resultados de esta fase del proyecto fueron publicados en el año 2007. Se continua con la misma muestra, pero se pretende encontrar un mayor número de SNP (3,2 millones) (International HapMap Consortium, 2007; Skipper, 2007).

Fase III: En la última fase del proyecto publicada en 2009 se aumentan las poblaciones de las 5 de la fase I a 11. Los SNP añadidos como resultados al estudio fueron 1,6 millones.

Dentro de cada uno de los haplotipos existen una serie de SNPs que lo identifican y se denominan tag SNPs. Los tag SNPs se presentarán como los resultados del proyecto. En total el proyecto identifica en torno 10 millones de SNPs de los cuales 500000 son tag

SNPs.

Los datos que vamos a analizar en este apartado (4.9.2) corresponden a la Fase III y contienen las 11 poblaciones. La tabla 4.6 contiene todas las áreas poblacionales donde se han recogido las muestras acompañadas de las letras con las que se han codificado en el software para realizar los análisis:

Código	**Población**
CEU	Utah con ascendencia Europa del norte y occidental.
CHB	Chinos Han de Beijing, China.
JPT	Japoneses de Tokyo, Japón.
YRI	Yoruba de Ibadan, Nigeria.
ASW	Estadounidenses del suroeste con ascendencia africana.
CHD	Chinos en la metrópolis de Denver, Colorado, Estados Unidos.
GIH	Indios Gujarati residentes en Houston, Texas, Estados Unidos.
LWK	De étnia Luhya de Webuye, Kenia.
MKK	Massais de Kinyawa, Kenia.
MEX	De Los Ángeles, California, Estados Unidos con ascendencia mejicana.
TSI	Residentes en la Toscana de Italia.

Tabla 4.6: Poblaciones de la fase III del proyecto HapMap con la codificación realizada en la aplicación práctica.

Todos los datos recogidos llevan asociado un compromiso y consentimiento informado internacional del proyecto (Rotimi *et al.*, 2007) así como un estudio ético de la investigación (International HapMap Consortium y others, 2004).

Se han utilizado solamente los datos del cromosoma 10 y se han eliminado todos aquellos polimorfismos que tienen datos perdidos resultando en una matriz de 1397 individuos en los que se han obtenido 30684 alelos correspondientes a 15342 polimorfismos.

Objetivos del ejemplo

Para el ejemplo del proyecto HapMap los objetivos planteados serán los siguientes:

Objetivo 1. Analizar la presencia de diferencias significativas en los polimorfismos el cromosoma 10 en las diferentes poblaciones del estudio.

Objetivo 2. Buscar comportamientos diferenciados de estos polimorfismos en cada una de las poblaciones en relación al resto de ellas.

Objetivo 3. En el caso de que existan diferencias significativas, estudiar entre cuales de estas poblaciones se presentan esas diferencias.

Objetivo 4. Presentar diferentes agrupaciones creadas entre las poblaciones de estudio.

Metodología

Este ejemplo permitirá ilustrar las técnicas presentadas en este capítulo.

En primer lugar se emplearán el PERMANOVA y el BOOTMANOVA, que debido a sus dimensiones obtienen resultados muy similares, tanto para los contrastes generales como para el análisis de contrastes a posteriori en caso de que sea necesario, con la corrección de Bonferroni debido a su tamaño. Al realizar el Análisis de Coordenadas Principales y el Análisis de Coordenadas Principales sobre sus centroides con sus regiones de confianza bootstrap asociadas.

Finalmente se realizarán los biplot asociados a este análisis.

Resultados

Comenzamos aquí con el Análisis de Coordenadas Principales para todos los individuos a partir de la matriz de distancias calculada usando como índice de similaridad el coeficiente de concordancia simple $\frac{a+d}{a+b+c+d}$. La tabla 4.7 muestra la varianza explicada por las 10 primeras dimensiones que, en conjunto no supera el 14 %. Esto ocurre de forma

habitual en los casos en los que el número de variables e individuos es muy elevado.

	Eigenvalues	Variance Explained	Cummulative
Dim 1	0.011898	7.795	7.795
Dim 2	0.005504	3.606	11.400
Dim 3	0.000739	0.484	11.884
Dim 4	0.000718	0.470	12.355
Dim 5	0.000635	0.416	12.770
Dim 6	0.000359	0.235	13.006
Dim 7	0.000340	0.223	13.228
Dim 8	0.000336	0.220	13.448
Dim 9	0.000331	0.217	13.665
Dim 10	0.000327	0.214	13.879

Tabla 4.7: Varianza explicada por las coordenadas principales de los datos de Hapmap

Realizando la representación gráfica del primer plano principal del Análisis de Coordenadas Principales se muestra en la figura 4.8.

En este gráfico es posible observar una separación bastante clara entre las diferentes razas del estudio. En la parte izquierda del gráfico se encuentran los individuos de raza negra, es decir los Yoruba, los de étnia Luhya, los Estadounidenses del suroeste con ascendencia africana y los Massais.
En la parte inferior derecha aparecen aquellos que tienen raza asiática, los japoneses, los chinos en la metrópolis de Denver y los chinos Han de Beijing.
En la parte superior se encuentran aquellos individuos que tienen raza blanca, es decir los individuos residentes en la Toscana y los individuos de Utah con ascendencia de Europa del norte y occidental.
Los individuos restantes, que se encuentran en las posiciones intermedias recoge a los mejicanos y a los indios.

Será de interés tener como referencia los demás planos que conforman las 10 dimensiones, por ello se han recogido en la figura 4.9.

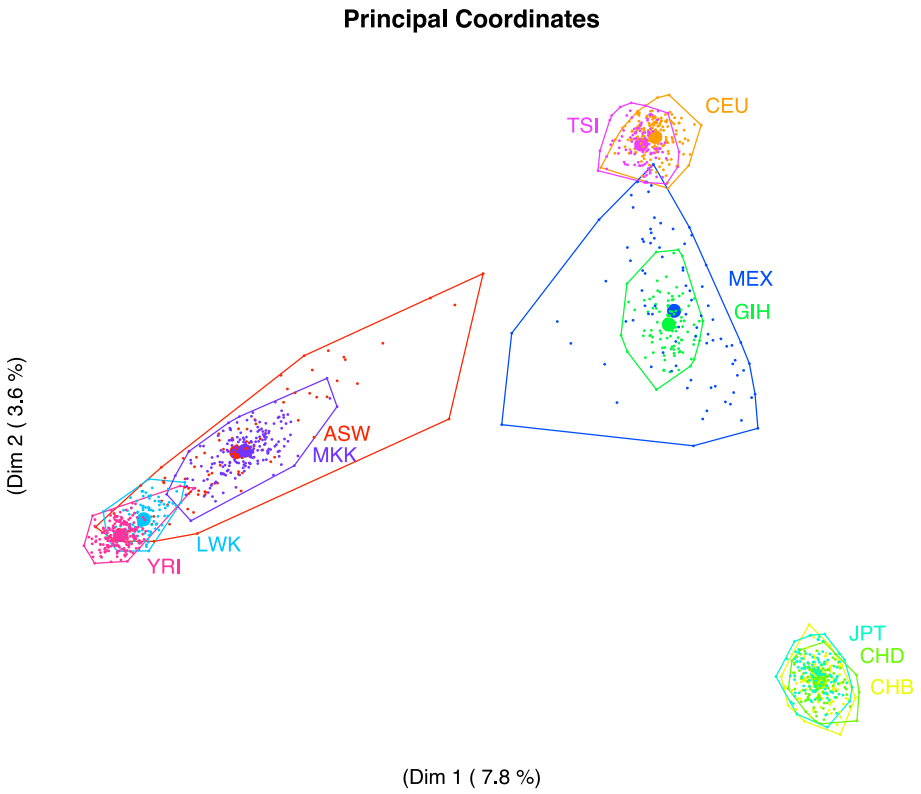

Figura 4.8: Primer plano principal del ACoP para los datos completos

El plano formado por las dimensiones 3 y 4 es capaz de discriminar los grupos MEX y GIH y separar la población MKK de las del resto de raza negra.
El resto de los planos no muestra ninguna separación clara entre los grupos de estudio. Los resultados son similares a los obtenidos por Demey *et al.* (2008) con los datos de la fase I (solo 4 grupos) y el cromosoma 22, que en aquel caso la tercera coordenada principal separaba los dos grupos de asiáticos. Igual que en ese caso, en este análisis, los grupos no tienen obligatoriamente que separarse, ya que maximizamos la variabilidad total y no la variabilidad entre grupos.

A continuación, realizaremos los contrastes generales para buscar diferencias entre las poblaciones de estudio utilizando MANOVAs basados en distancias. En la tabla 4.8 se encuentran los resultados obtenido tanto para el PERMANOVA, con 1000 permutaciones, y BOOTMANOVA con 1000 repeticiones bootstrap.

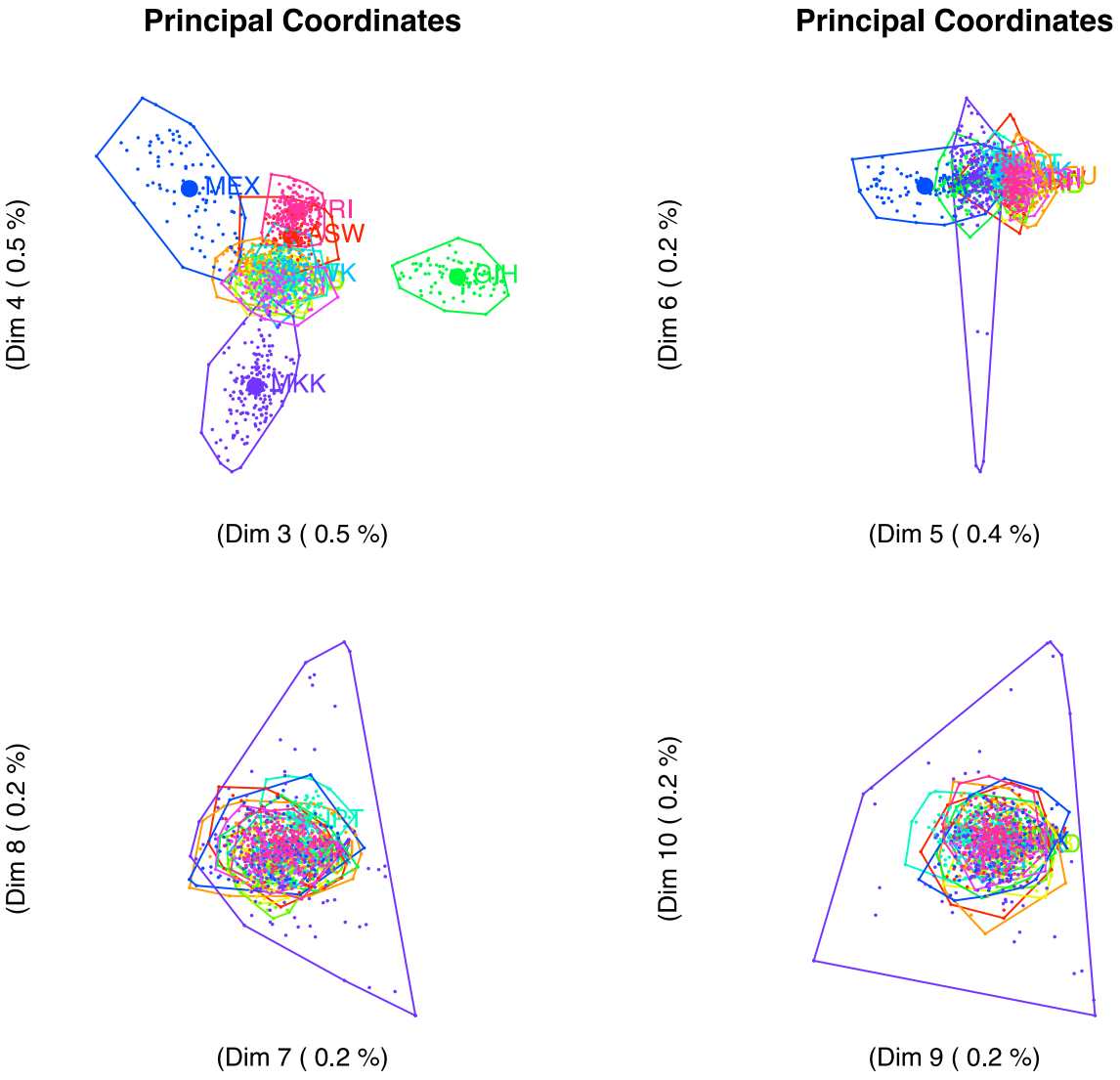

Figura 4.9: Dimensiones de la 3 a la 10 del ACoP para los datos completos

Se encuentran diferencias altamente significativas entre las poblaciones del estudio. Es de interés, igual que en el caso anterior, saber si alguno de los contrastes individuales es significativo. Estos contrastes individuales se encuentran en la tabla 4.9.

Todos los contrastes son altamente significativos, se puede afirmar que todas las poblaciones del estudio poseen características únicas.

Para estudiar entre qué grupos se presentan diferencias significativas realizaremos

	Explicada	Residual	G.L. Num	G.L. Denom	F-exp	p-value
Total	4.169	18.622	10	139	3.112	0.000100

Tabla 4.8: PERMANOVA y BOOTMANOVA global para los datos de HAPMAP

	Explained	Residual	G.L. Num	G.L. Denom	F-exp	p-value
C ASW	1.05	183.78	1	1386	7.92	0.00
C CEU	2.89	183.78	1	1386	21.78	0.00
C CHB	3.38	183.78	1	1386	25.52	0.00
C CHD	2.84	183.78	1	1386	21.45	0.00
C GIH	1.59	183.78	1	1386	11.96	0.00
C JPT	2.96	183.78	1	1386	22.30	0.00
C LWK	2.51	183.78	1	1386	18.94	0.00
C MEX	1.60	183.78	1	1386	12.10	0.00
C MKK	2.48	183.78	1	1386	18.71	0.00
C TSI	1.75	183.78	1	1386	13.22	0.00
C YRI	5.35	183.78	1	1386	40.36	0.00

Tabla 4.9: Contrastes del PERMANOVA y BOOTMANOVA para el proyecto HapMap

los contrastes a posteriori que se recogen en la tabla 4.10 que muestra las comparaciones entre todos los grupos de estudio.

Se encuentran diferencias altamente significativas entre la mayor parte de los grupos, sin embargo, existen tres parejas que no se puede afirmar que presenten diferencias entre sus individuos:

- Los individuos de Utah con ascendencia Europea del norte y occidental (CEU) y los Residentes en la Toscana de Italia (TSI).
- Los chinos Han de Beijing, China (CHB) y los chinos en la metrópolis de Denver, Colorado, Estados Unidos (CHD).
- Los chinos en la metrópolis de Denver, Colorado, Estados Unidos (CHD) y los ja-

poneses de Tokyo, Japón.

Se puede observar una relación mayor, o una menor diferenciación, entre los individuos con ascendencia europea por una lado, y de ascendencia asiática por otro, que con el resto de los individuos del estudio.

Emplearemos las representaciones gráficas sobre los centroides descritas en este capítulo para ilustrar los resultados obtenidos a partir de este análisis. Se incluirán los gráficos con regiones de confianza bootstrap para ayudar en la interpretación de los resultados.

Comenzaremos realizando la representación gráfica del PERAMANOVA. El primer plano factorial se encuentra en la figura 4.10.

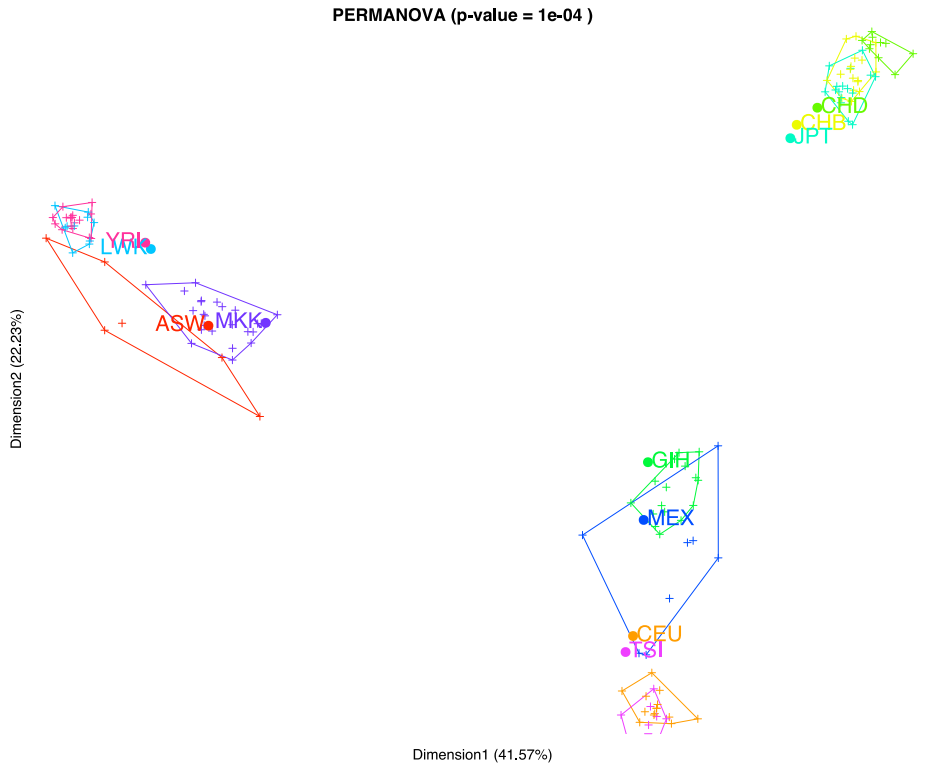

Figura 4.10: Dimensiones de la 3 a la 10 del ACoP sobre los centroides para los datos completos

El primer plano principal del ACoP, que aparece en la figura 4.8, es muy similar al que obtenemos como resultado de la representación gráfica del PERMANOVA, figura

4.10. Se observa que la distribución de los grupos es la misma que hemos mencionado anteriormente.

Igual que realizábamos en el caso del Análisis de Coordenadas Principales, utilizaremos las 10 primeras componentes para crear estos gráficos de Coordenadas Principales sobre los centroides. Las representaciones del resto de planos aparecen en la figura 4.11.

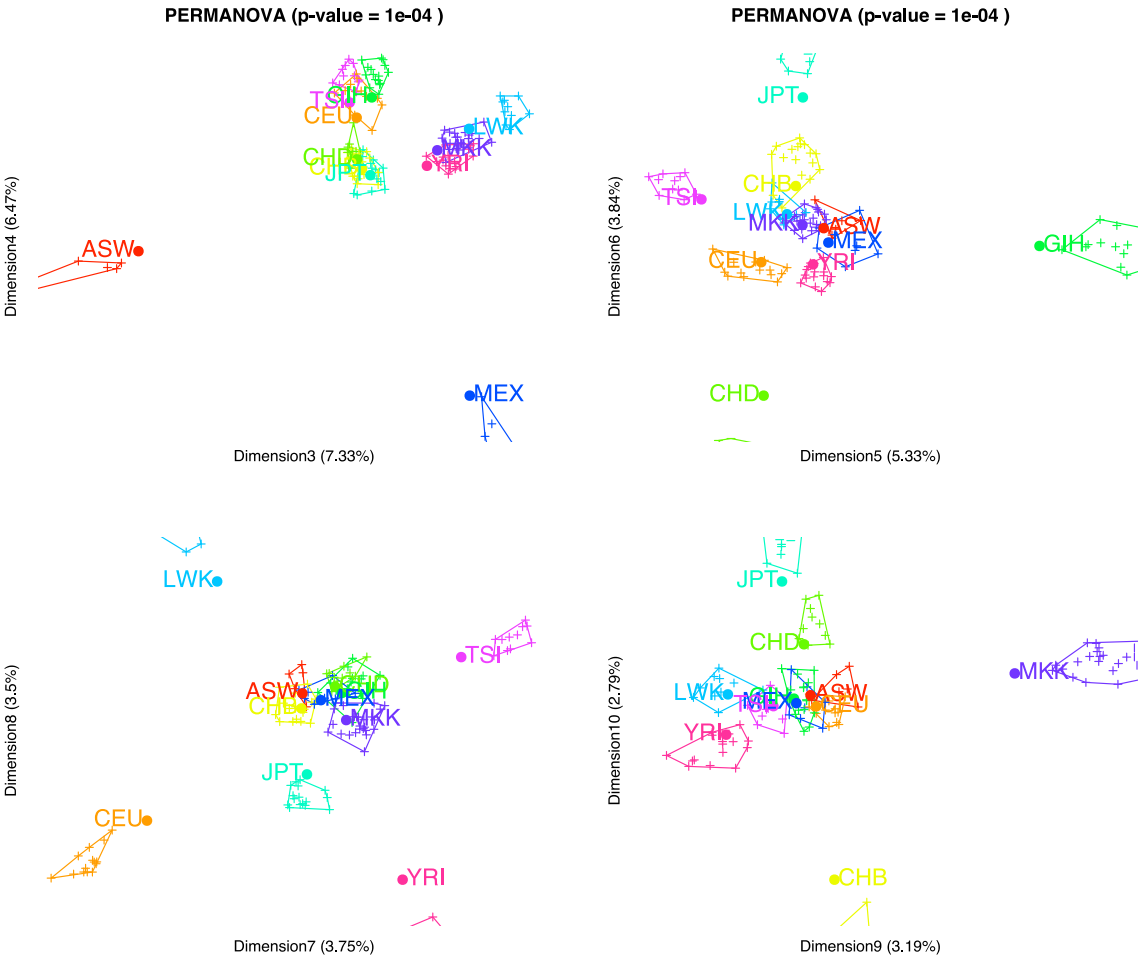

Figura 4.11: Dimensiones de la 3 a la 10 del ACoP sobre los centroides para los datos completos

Utilizando el plano 3-4 es posible separar los grupos de individuos del suroeste de Estados Unidos con ascendencia africana (ASW) y aquellos individuos con ascendencia mejicana que son de Los Ángeles, California, Estados Unidos, del resto de los grupos (MEX).

El plano que utiliza las dimensiones 5 y 6 permitirá diferenciar 6 poblaciones del resto de los individuos que se encuentran en la zona central del gráfico: los residentes en la Toscana de Italia (TSI); los japoneses de Tokyo, Japón (JPT); los indios Gujarati residentes en Houston, Texas, Estados Unidos (GIH); los yotuba de Ibadan, Nigeria (YRI); los chinos de la metrópolis de Denver, Colorado, Estados Unidos (CHD); y los individuos de Utah con ascendencia de Europa del norte y occidental (CEU).
El plano siguiente, creado por las dimensiones 7 y 8, permite separar, además de algunos grupos que ya se han separado en planos anteriores, como TSI, CEU, JPT, YRI, el grupo de individuos de étnia Luhya de Webuye, Kenia.
El último plano construido, plano 9-10, separará las dos poblaciones restantes que hasta ahora no habían aparecido, los chinos Han de Beijing, China y los massais de Kinyawa, Kenia.

Se puede observar que este tipo de gráficos nos permite separar con mayor precisión todas las poblaciones sometidas a estudio.

Empleando las regiones de confianza bootstrap se ha generado el gráfico 4.12 para el plano 1-2.

Se observan los grupos definidos anteriormente de forma más clara que en las dos representaciones anteriores, figuras 4.8, 4.10.

Utilizando los planos restantes, en el gráfico análogo al de la figura 4.11, es posible estudiar los contrastes a posteriori recogidos en la tabla 4.10. Estos gráficos es posible encontrarlos en la figura 4.13

En el plano 3-4 se pueden observar las diferencias de los grupos GIH y MEX del resto de poblaciones. En el plano 5-6 son los grupos MKK y JPT y en el plano 9-10 CHD y CHB del resto de las poblaciones. Cabe destacar la relación entre los individuos de Utah con ascendencia de Europa del norte y occidental y los residentes en la Toscana de Italia, que se encuentran superpuestos en todos los planos y en los contrastes a posteriori no

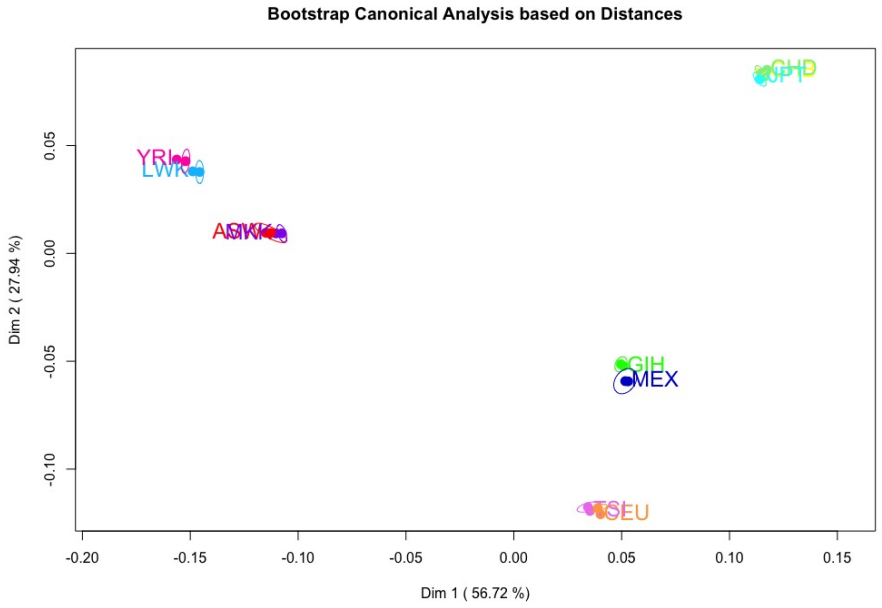

Figura 4.12: Regiones de confianza bootstrap del HapMap. Dimensiones 1 y 2.

se han podido detectar diferencias significativas entre ellos.

Conclusiones

- Se ha comprobado que existen diferencias altamente significativas en las poblaciones sometidas a estudio. Los resultados obtenidos con el análisis PERMANOVA y con el BOOTMANOVA, en este caso, son los mismos.
- Todas las poblaciones presentan diferencias en los polimorfismos del cromosoma 10.
- Un gran número de poblaciones presentan diferencias altamente significativas entre ellas, excepto entre la población de Utah con ascendencia de Europa del norte y occidental y los residentes en la Toscana de Italia; los chinos Han de Beijing, China y los chinos en la metrópolis de Denver, Colorado, Estados Unidos; los chinos Han de Beijing, China y los japoneses de Tokyo, Japón; y por último, los chinos en la metrópolis de Denver, Colorado, Estados Unidos y los japoneses de Tokyo, Japón.
- Se observan cuatro grupos de poblaciones diferenciadas:

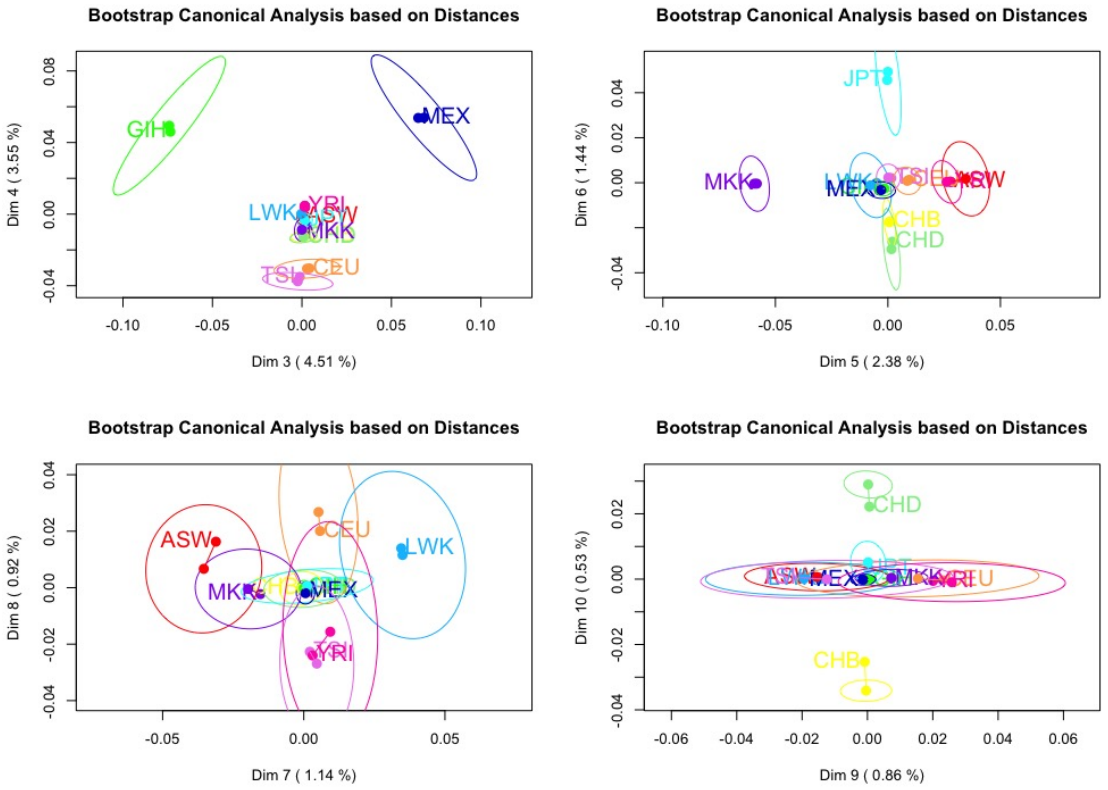

Figura 4.13: Regiones de confianza bootstrap del HapMap. Dimensiones de la 3 a la 10.

Grupo 1 Individuos de raza negra: Yoruba de Ibadan, Nigeria (YRI); de étnia Luhya de Webuye, Kenia (LWK), estadounidenses del suroeste con ascendencia africana (ASW) y los masais de Kinyawa (MKK).

Grupo 2 Individuos de raza asiática: japoneses de Tokyo, Japón (JPT); chinos en la metrópolis de Denver, Colorado, Estados Unidos (CHD) y chinos Han de Beijing, China (CHB).

Grupo 3 Individuos de raza blanca: residentes en la Toscana de Italia (TSI) y los individuos de Utah con ascendencia de Europa del norte y occidental (CEU).

Grupo 4 Individuos de otras razas: individuos de Los Ángeles, California, Estados Unidos con ascendencia mejicana e indios Gujarati residentes en Houston, Texas, Estados Unidos.

	Explicada	Residual	G.L. Num	G.L. Denom	F-exp	p-value
ASW-CEU	0.066	18.622	1	139	0.489	0.000100
ASW-CHB	0.089	18.622	1	139	0.666	0.000100
ASW-CHD	0.072	18.622	1	139	0.536	0.000100
ASW-GIH	0.063	18.622	1	139	0.470	0.000100
ASW-JPT	0.085	18.622	1	139	0.637	0.000100
ASW-LWK	0.031	18.622	1	139	0.233	0.032897
ASW-MEX	0.049	18.622	1	139	0.367	0.000200
ASW-MKK	0.043	18.622	1	139	0.319	0.000600
ASW-TSI	0.059	18.622	1	139	0.438	0.000100
ASW-YRI	0.042	18.622	1	139	0.316	0.000800
CEU-CHB	0.094	18.622	1	139	0.699	0.000100
CEU-CHD	0.077	18.622	1	139	0.573	0.000100
CEU-GIH	0.046	18.622	1	139	0.342	0.000200
CEU-JPT	0.089	18.622	1	139	0.664	0.000100
CEU-LWK	0.092	18.622	1	139	0.686	0.000100
CEU-MEX	0.038	18.622	1	139	0.284	0.003000
CEU-MKK	0.103	18.622	1	139	0.770	0.000100
CEU-TSI	0.027	18.622	1	139	0.203	0.136186
CEU-YRI	0.124	18.622	1	139	0.927	0.000100
CHB-CHD	0.026	18.622	1	139	0.192	0.223178
CHB-GIH	0.072	18.622	1	139	0.538	0.000100
CHB-JPT	0.028	18.622	1	139	0.209	0.101290
CHB-LWK	0.112	18.622	1	139	0.838	0.000100
CHB-MEX	0.066	18.622	1	139	0.490	0.000100
CHB-MKK	0.137	18.622	1	139	1.022	0.000100
CHB-TSI	0.088	18.622	1	139	0.660	0.000100
CHB-YRI	0.147	18.622	1	139	1.095	0.000100
CHD-GIH	0.060	18.622	1	139	0.445	0.000100
CHD-JPT	0.029	18.622	1	139	0.213	0.087091
CHD-LWK	0.093	18.622	1	139	0.693	0.000100
CHD-MEX	0.053	18.622	1	139	0.392	0.000100
CHD-MKK	0.115	18.622	1	139	0.858	0.000100
CHD-TSI	0.072	18.622	1	139	0.536	0.000100
CHD-YRI	0.123	18.622	1	139	0.920	0.000100
GIH-JPT	0.069	18.622	1	139	0.512	0.000100
GIH-LWK	0.087	18.622	1	139	0.646	0.000100
GIH-MEX	0.041	18.622	1	139	0.308	0.001000
GIH-MKK	0.101	18.622	1	139	0.752	0.000100
GIH-TSI	0.044	18.622	1	139	0.332	0.000600
GIH-YRI	0.116	18.622	1	139	0.868	0.000100
JPT-LWK	0.108	18.622	1	139	0.805	0.000100
JPT-MEX	0.061	18.622	1	139	0.458	0.000100
JPT-MKK	0.132	18.622	1	139	0.985	0.000100
JPT-TSI	0.084	18.622	1	139	0.625	0.000100
JPT-YRI	0.140	18.622	1	139	1.049	0.000100
LWK-MEX	0.069	18.622	1	139	0.516	0.000100
LWK-MKK	0.033	18.622	1	139	0.243	0.025097
LWK-TSI	0.084	18.622	1	139	0.627	0.000100
LWK-YRI	0.030	18.622	1	139	0.226	0.049095
MEX-MKK	0.081	18.622	1	139	0.603	0.000100
MEX-TSI	0.035	18.622	1	139	0.263	0.008799
MEX-YRI	0.096	18.622	1	139	0.714	0.000100
MKK-TSI	0.093	18.622	1	139	0.696	0.000100
MKK-YRI	0.042	18.622	1	139	0.314	0.001300
TSI-YRI	0.114	18.622	1	139	0.850	0.000100

Tabla 4.10: Comparaciones por parejas de los grupos de HapMap

Capítulo 5

Análisis de la Redundancia para datos binarios

Notación

I : Número de individuos de la matriz de estudio.

J : Número de variables predictoras del estudio.

K : Número de variables respuesta del estudio.

$\mathbf{X}_{(I\times J)}$: Matriz de variables explicativas o matriz de diseño con I individuos y J variables.

$\mathbf{Y}_{(I\times K)}$: Matriz de variables respuesta con I individuos y K variables.

R : Rango de la matriz $\mathbf{X}$.

S : Rango reducido de la matriz $\mathbf{X}$ ó $\mathbf{Y}$.

Σ_{XY} : Matriz de covarianzas entre $\mathbf{X}$ e $\mathbf{Y}$.

Σ_{YX} : Matriz de covarianzas entre $\mathbf{Y}$ e $\mathbf{X}$.

$\mathbf{B}_{(J\times K)}$: Matriz que contiene los parámetros de regresión desconocidos.

$\mathbf{U}_{(I\times R)}$: Matriz de vectores singulares por la derecha ($\mathbf{u}_r$) con I filas y R columnas.

$\mathbf{V}_{(J\times R)}$: Matriz de vectores singulares por la izquierda ($\mathbf{v}_r$) con J filas y R columnas.

$\boldsymbol{\lambda}_r$: Valores singuales no negativos decrecientes de $\mathbf{X}^T\mathbf{X}$ y $\mathbf{X}\mathbf{X}^T$.

π_{ij} : Probabilidad esperada del individuos i en la variable j.

$\mathbf{Z}$: Matriz de logits ajustados de la regresión.

L : Función de enlace.

$\mathbf{b}_0$: Intersección del modelo logístico.

$\mathbf{P}_{(I\times R)}$: Marcadores ajustados de los lugares.

$\mathbf{T}$: Proyección de las respuestas sobre las Componentes Principales.

5.1. Introducción

Entre los diversos métodos que permiten la extensión del Modelo Lineal General Multivariante, cuando se dispone de un conjunto de varias variables respuesta, se encuentra el denominado Análisis de la Redundancia (RDA) que fue propuesto por Rao (1964), y redescubierto como una alternativa al Análisis Canónico de Correlaciones (CCA) por van den Wollenberg (1977). El análisis trata de maximizar la varianza explicada por el modelo que, en este caso se denomina redundancia y da nombre a la técnica.

Desde un punto de vista, el RDA puede entenderse como una extensión del Modelo Lineal General Mutltivariante en la que se obtiene una reducción de la dimensión en las respuestas que facilita la interpretación permitiendo representaciones gráficas mediante biplots. Desde otro punto de vista, el RDA puede entenderse como un ACP de la matriz de respuestas en el que las componentes son combinaciones lineales de las variables predictoras. Takane y Shibayama (1991) hace esta misma propuesta con el nombre de Análisis de Componentes Principales Restringido o Análisis de Componentes Principales con información externa.

El Análisis de la Redundancia ha sido extendido para incluir variables cualitativas empleando cuantificaciones como ocurre en el sistema Gifi (Israels, 1984). En la práctica esta técnica no está muy extendida y no se encuentra en la mayor parte de la literatura reciente. Otra alternativa para las variables de tipo binario es realizar un Análisis de la Redundancia basado en distancias, igual que ocurría con el MANOVA; esta alternativa fue propuesta por Legendre y Anderson (1999); McArdle y Anderson (2001) aproximadamente a la vez que las técnicas descritas en el capítulo 4. En esta alternativa se emplea

el Análisis de Coordenadas Principales basado en las distancias calculadas a partir de las variables respuesta.

En este capítulo, proponemos el Análisis de la Redundancia para datos binarios basado en Modelos Lineales Generalizados (con enlace logístico) en lugar de en cuantificaciones o distancias.

Comenzaremos desarrollando la teoría del Análisis de la Redundancia para un conjunto de variables repuestas continuas en la sección 5.2. A continuación, se realiza la extensión de este RDA para variables respuesta de tipo binario, el desarrollo de esta técnica se encuentra en la sección 5.3.

La sección 5.4 contendrá las representaciones biplot asociadas al Análisis de la Redundancia, tanto para respuestas continuas como para respuestas binarias. El software existente y el desarrollado en este trabajo para el Análisis de la Redundancia ha sido recogido en la sección 5.5. Finalizaremos el capítulo con un ejemplo que permita explicar la utilidad de las técnicas así como su aplicabilidad (sección 5.6).

En resumen, el Análisis de la Redundancia es una técnica que extrae la parte de un conjunto de variables respuesta que se encuentra mejor explicada por el conjunto de predictores. Al asociar las técnicas biplot a este tipo de gráficos permitimos realizar una representación simultánea de las variables predictoras, las variables respuesta y los individuos en un plano de dimensión reducida.

5.2. Análisis de la Redundancia para datos continuos

Como en capítulos anteriores, partimos de una matriz predictores $\mathbf{X}_{I\times J}$ con J variables continuas y una matriz de respuestas $\mathbf{Y}_{I\times K}$ con K variables continuas, ambas con I individuos. Supondremos que ambas matrices están centradas y probablemente estan-

darizadas por columnas si las variables no son de escalas comparables.

En el Análisis de la Redundancia se busca una combinación lineal de las columnas de $\mathbf{X}$ que maximice la varianza explicada de todas las K variables respuesta simultáneamente.

El Análisis de la Redundancia se puede obtener de los valores y vectores propios de la matriz

$$\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}, \tag{5.1}$$

o de la matriz

$$\mathbf{Y}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}. \tag{5.2}$$

El procedimiento de cálculo del RDA, según Legendre y Legendre (2012), puede ser descrito en dos pasos, una Regresión Lineal Multivariante seguida de un Análisis de Componentes Principales. Esta visión se entiende como una forma intuitiva de entender la técnica y realizar los cálculos pertinentes que puede ser fácilmente generalizable cuando disponemos de respuestas binarias.

El procedimiento es el siguiente:

- En primer lugar realizamos una regresión lineal multivariante de la matriz $\mathbf{Y}$, es decir, realizaremos una regresión lineal de cada columna de $\mathbf{Y}$ en $\mathbf{X}$.
  El cálculo de la matriz de coeficientes de regresión puede realizarse a partir de la fórmula

  $$\mathbf{B} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}. \tag{5.3}$$

  No es necesario incluir las intersecciones del modelo debido a que habíamos definido $\mathbf{X}$ e $\mathbf{Y}$ como matrices centradas y estandarizadas.

- A partir de la matriz **B** y la matriz **X**, serán calculados los valores ajustados de la regresión

$$\hat{\mathbf{Y}} = \mathbf{XB}, \tag{5.4}$$

donde $\hat{\mathbf{Y}}$ contiene la parte de la matriz de respuestas **Y** que está explicada por los predictores de **X**.
Igual que en la ecuación (4.1), la matriz de residuales **E** puede ser calculada como $\mathbf{E} = \mathbf{Y} - \hat{\mathbf{Y}}$ y contendrá la parte no explicada por el modelo.

- A continuación, realizaremos un Análisis de Componentes Principales utilizando la Descomposición en Valores Singulares (SVD), igual que en la ecuación (3.2):

$$\hat{\mathbf{Y}} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T, \tag{5.5}$$

donde **U** contiene los vectores propios de $\hat{\mathbf{Y}}\hat{\mathbf{Y}}^T$, **V** los vectores propios de $\hat{\mathbf{Y}}^T\hat{\mathbf{Y}}$ y $\mathbf{\Lambda}$ la raíz cuadrada de los valores propios no negativos de ambas matrices, que son los mismos.

  - Existen solo mín $(J, K, I - 1)$ valores propios no nulos y, generalmente, solo es necesario emplear un número reducido de ellos para describir la variabilidad explicada por el modelo.
  - **V** también contiene los vectores propios de la matriz descrita en la ecuación (5.2).
  - Es posible realizar la interpretación como un conjunto de componentes principales de **Y** que han sido construidas a partir de combinaciones lineales de los predictores recogidos en **X**. De esta forma **V** contendrá un conjunto de vectores ortogonales que definen el subespacio de la misma forma que lo hace el PCA.

5.3. Análisis de la Redundancia para datos binarios

Cuando las variables repuesta son binarias, no es posible utilizar el Modelo Lineal General, por ello emplearemos el Modelo Lineal Generalizado que, para datos binarios, usará la distribución binomial con el enlace *logit*. La generalización es inmediata y el esquema para la realización de este análisis será muy similar al explicado anteriormente, como se puede observar en la figura 5.1.

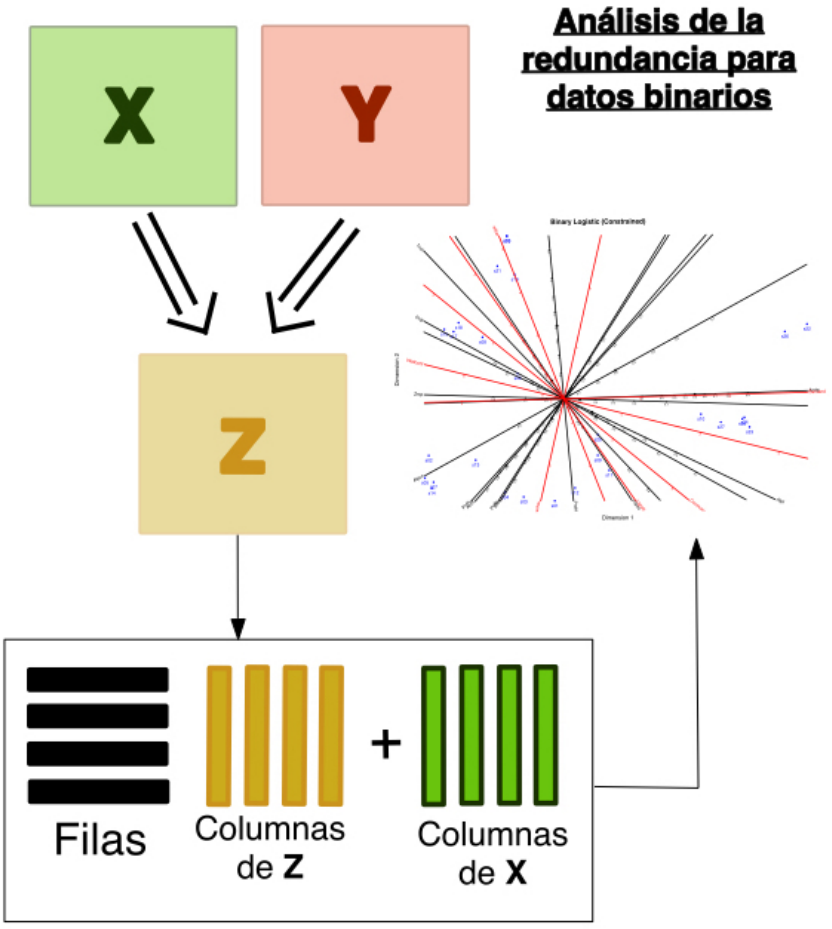

Figura 5.1: Esquema de la realización del Análisis de la Redundancia para datos binarios

Volvemos a partir de una matriz de predictores $\mathbf{X}_{(I\times J)}$ con J variables explicativas continuas y una matriz de respuestas $\mathbf{Y}_{(I\times K)}$ con K variables respuesta, pero a diferencia del caso anterior, estas variables serán de tipo binario, 1 si existe presencia de la variable medida y 0 en caso contrario, que se convertirán en las probabilidades observadas. Ambas matrices contienen las medidas para I individuos.

La matriz de predictores será centrada y estandarizada por columnas si las variables no tienen escalas comparables.

La matriz de variables respuesta no puede ser centrada de la forma habitual. Por ello, se calcula la matriz de probabilidades esperadas $\mathbf{\Pi}_{(I\times K)}$ utilizando la regresión logística para cada columna de $\mathbf{Y}$ en todo el conjunto de predictores $\mathbf{X}$. Los logits esperados serán recogidos en la matriz $\mathbf{Z}_{(I\times K)}$.

Las probabilidades esperadas, calculadas mediante la regresión logística, son:

$$\pi_{ik} = \frac{e^{z_{ik}}}{1 + e^{z_{ik}}} = \frac{e^{\left(\beta_{k0}+\beta_{k1}x_{i1}+\cdots+\beta_{kJ}x_{iJ}\right)}}{1 + e^{\left(\beta_{k}0+\beta_{k1}x_{i1}+\cdots+\beta_{kJ}x_{iJ}\right)}}. \tag{5.6}$$

En escala logit

$$z_{ik} = logit\,(\pi_{ik}) = \ln\left(\frac{\pi_{ik}}{1-\pi_{ik}}\right) = z_{ik} = \beta_{k0} + \beta_{k1}x_{i1} + \cdots + \beta_{kJ}x_{iJ}, \tag{5.7}$$

con $i = 1, \ldots, I$ y $k = 1, \ldots, K$.

Obsérvese que, tanto en el caso continuo como en el caso binario, la matriz $\mathbf{X}$ tiene que ser de rango completo para que sea posible calcular las regresiones.

Ahora, vamos a realizar la correspondiente generalización del algoritmo para datos cuantitativos realizando las adaptaciones necesarias para datos dicotómicos. Igual que para el caso continuo vamos a resumir el procedimiento varios pasos:

- En primer lugar, realizaremos una Regresión Logística estándar de cada columna de la matriz $\mathbf{Y}$ en $\mathbf{X}$. Debido a que las respuestas no pueden ser centradas, es necesario mantener la intersección en los modelos.
  La matriz de coeficientes de la regresión para los predictores es $\mathbf{B}_{J\times K} = \left(\beta_{jk}\right)$ y la intersección $\mathbf{b}_0 = (\beta_{k0})$. Pueden usarse penalizaciones, como Ridge, para evitar

el sobreajuste cuando el número de predictores es alto o existe un problema de separación.

- Calculamos los *logits* ajustados de la regresión

$$\mathbf{Z} = \mathbf{1}\mathbf{b}_0^T + \mathbf{XB}, \tag{5.8}$$

y

$$\mathbf{\Pi} = \frac{e^{\mathbf{Z}}}{1 + e^{\mathbf{Z}}}. \tag{5.9}$$

Las columnas de $\mathbf{Z}$ no están centradas y contiene las estimaciones del logits en la escala lineal.

- Realizaremos un Análisis de Componentes Principales de los logits ajustados de la descomposición en valores singulares del segundo término de la ecuación (5.8).

$$\hat{\mathbf{Z}} = \mathbf{1}\mathbf{b}_0^T + \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T. \tag{5.10}$$

De esta forma podemos hacer una representación, en dimensión reducida, de la matriz de respuestas en escala logit que convertiremos después en un biplot logístico.

5.4. Representación biplot asociada

Es posible asociar representaciones biplot, como las del capítulo 3, a este tipo de análisis, tanto para datos continuos como para datos binarios.

De la misma manera que en la ecuación (3.2), tanto en el Análisis de la Redundancia para datos continuos (ecuación (5.5)), como en el Análisis de la Redundancia para datos binarios (ecuación (5.10)), la SVD permitirá construir los biplot.

En la sección 5.4.1 desarrollamos la representación biplot asociada a datos continuos y en la sección 5.4.2 se desarrollará la representación biplot asociada a las respuestas binarias. Ambas representaciones pueden convertirse en un triplot representando las

variables predictoras, las respuestas y los individuos sobre el mismo plano de dimensión reducida.

5.4.1. Representaciones biplot asociadas a Análisis de la Redundancia para datos continuos

Partiendo de la SVD, obtenida en la ecuación (5.5), es posible obtener un biplot de la forma

$$\hat{\mathbf{Y}} = \mathbf{P}\mathbf{Q}^T, \tag{5.11}$$

donde tomaremos $\mathbf{P} = \mathbf{U}\mathbf{\Lambda}$ y $\mathbf{Q} = \mathbf{V}$ (o cualquiera de las alternativas para datos continuos presentadas en el capítulo 3).

En las aplicaciones del ámbito de la Ecología, se denomina a la matriz $\mathbf{P}$ marcadores ajustados de los puntos de muestreo y son también $\mathbf{P} = \hat{\mathbf{Y}}\mathbf{V} = \mathbf{X}\mathbf{B}\mathbf{V}$. Este biplot también puede denominarse biplot restringido, como hemos indicado en la introducción de este capítulo, el RDA puede denominarse también Análisis de Componentes Principales Restringido.

Como en el PCA, los valores de $\mathbf{Y}$ también puede ser proyectados sobre las componentes principales, $\mathbf{T} = \mathbf{Y}\mathbf{V}$, para aproximar los valores observados en lugar de los ajustados.

Esto define un biplot $\mathbf{Y} = \mathbf{T}\mathbf{V}^T$ para las respuestas. E igual que ocurría en el biplot definido anteriormente, los marcadores $\mathbf{T}$ serán denominados marcadores ajustados de los lugares en las aplicaciones ecológicas.

Es posible calcular los coeficientes de las variables predictoras de $\mathbf{X}$ a partir de la matriz $\mathbf{C} = \mathbf{B}\mathbf{V}$ para calcular de los marcadores ajustados de los puntos de muestreo, por lo tanto, es posible emplear estos vectores en un biplot de interpolación para realizar

proyecciones de nuevas observaciones.

Dado un nuevo grupo de individuos con $\mathbf{X}_h$ en los predictores, los marcadores ajustados son

$$\mathbf{P}_h = \mathbf{X}_h\mathbf{C} = \mathbf{X}_h\mathbf{B}\mathbf{V}, \tag{5.12}$$

y las predicciones de las respuestas en el biplot final son

$$\hat{\mathbf{Y}} = \mathbf{P}_h\mathbf{V}^T. \tag{5.13}$$

También es posible obtener un biplot de predicción para la matriz de variables explicativas $\mathbf{X}$ realizando una regresión de los predictores en los marcadores ajustados de los puntos de muestreo. Los marcadores en el biplot puede ser

$$\mathbf{H} = \left(\mathbf{P}^T\mathbf{P}\right)^{-1}\mathbf{P}^T\mathbf{X}, \tag{5.14}$$

y el biplot para las predicciones es

$$\mathbf{X} \simeq \mathbf{P}\mathbf{H}^T. \tag{5.15}$$

Es posible construir un triplot típico combinando las matrices $\mathbf{P}$ (o $\mathbf{T}$), $\mathbf{Q}$ y $\mathbf{H}$.

5.4.2. Representaciones biplot asociadas al Análisis de la Redundancia para datos binarios

Igual que en las representaciones para datos continuos, partimos de la SVD de la 5.10 que es un biplot en escala logística como el descrito por Vicente-Villardon *et al.* (2006) de la forma

$$\hat{\mathbf{Z}} = \mathbf{1}\mathbf{b}_0^T + \mathbf{P}\mathbf{Q}^T, \tag{5.16}$$

tomando $\mathbf{P} = \mathbf{U\Lambda}$ y $\mathbf{Q} = \mathbf{V}$.

Igual que anteriormente, es posible afirmar que $\mathbf{P}$ contiene los marcadores ajustados de los lugares.

Es posible realizar el logit de la aproximación de la forma

$$logit\left(\pi_{ij}\right) = b_{j0} + \sum_{s=1}^{S} p_{is}q_{js} = b_{j0} + \mathbf{p}_i^T\mathbf{q}_j. \tag{5.17}$$

También se pueden aproximar las probabilidades esperadas en la escala original utilizando

$$\pi_{ij} = \frac{e^{\left(b_{j0}+\sum_{s=1}^{S} p_{is}q_{js}\right)}}{1 + e^{\left(b_{j0}+\sum_{s=1}^{S} p_{is}q_{js}\right)}}. \tag{5.18}$$

Vicente-Villardon *et al.* (2006); Legendre y Anderson (1999) estudian la geometría para realizar la representación gráfica. La idea general es que las direcciones de representación definidas por los coeficientes de regresión son las direcciones que mejor predicen las probabilidades esperadas. Es posible completar con escalas graduadas las direcciones del biplot para poder predecir las probabilidades.

Se puede obtener un biplot de predicción para $\mathbf{X}$ realizando una regresión de los predictores en los marcadores ajustados de los lugares de muestreo, de la misma forma que lo realizábamos en el caso continuo. Los marcadores del biplot $\mathbf{H}$ pueden ser calculados como en la ecuación (5.14) y el biplot para las predicciones como en (5.15).

Para el caso binario también será posible representar un triplot combinando las matrices $\mathbf{P}$, $\mathbf{Q}$ y $\mathbf{H}$. De esta forma nosotros tenemos representada en un biplot la parte de las variables binarias explicadas por los predictores.

Una forma alternativa para realizar el proceso es centrar la matriz **Z**, a continuación se realizará el cálculo de las Componentes Principales y se ajustará una Regresión Logística para cada columna de **Y** usando las puntuaciones de las componentes, en orden, como variables independientes para recalcular el biplot logístico. Luego obtenemos un biplot logistico en el que las respuestas tienen una relación logística para las dimensiones y las dimensiones son combinaciones lineales de los predictores.

Los cálculos de los marcadores de los puntos usando los valores originales en lugar de las probabilidades esperadas es más complicado y necesita una atención especial. A continuación lo describiremos con mayor detalle.

Para estimar los parámetros del modelo de la ecuación (5.18) maximizamos la función de coste

$$L = \sum_{i=1}^{I} \sum_{k=1}^{K} [-y_{ik} \log(\pi_{ik}) - (1 - y_{ik}) \log(1 - \pi_{ik})], \tag{5.19}$$

usando el método del descenso del gradiente en lugar del método de Newton-Raphson tradicional para obtener las estimaciones de máximoa verosimilitud. El algoritmo iterativo actualizará los valores de los parámetros en cada iteración como

$$p_{is} = p_{is} - \alpha \frac{\partial L}{\partial p_{is}},$$

$$q_{js} = q_{js} - \alpha \frac{\partial L}{\partial q_{js}},$$

$$b_{j0} = b_{j0} - \alpha \frac{\partial L}{\partial b_{j0}},$$

donde α es una constante y

$$\frac{\partial L}{\partial p_{is}} = \sum_{j=1}^{k} q_{js} (\pi_{ik} - y_{ik}) (s = 1, \ldots, S),$$

$$\frac{\partial L}{\partial q_{js}} = \sum_{j=1}^{k} p_{is} (\pi_{ik} - y_{ik}) (s = 1, \ldots, S),$$

$$\frac{\partial L}{\partial b_{j0}} = \sum_{i=1}^{I} (\pi_{ik} - y_{ik}),$$

serán los gradientes y S la dimensión de la solución. Las ecuaciones se pueden usar, de forma conjunta con restricciones fijadas inicialmente, en una rutina de optimización para obtener la solución del problema. Si alguno de los parámetros son conocidos con anterioridad, solo se usarán las ecuaciones que sean necesarias.

5.5. Software para los Análisis de la Redundancia

Igual que en el caso del biplot, esta técnica, para datos continuos, está ampliamente extendida y son muchos los software que incluyen el Análisis de la Redundancia. A diferencia del biplot, los software que permiten realizar este tipo de análisis, por regla general, tienen un coste para el usuario.

Ninguno de ellos contiene el Análisis de la Redundancia para datos de respuesta binaria, ya que es parte de la innovación de este trabajo. Se han desarrollado las funciones necesarias y se ha incluido en el paquete MultBiplotR (Vicente-Villardon, 2021), tanto para datos de respuesta continua como para datos de respuesta binaria.

5.5.1. Paquetes comerciales

Para todos los software de esta sección es necesario obtener licencias de pago.

Podría incluirse SPSS que también presenta la posibilidad de realizar el RDA, pero solo empleando la sintaxis o macros complementarias.

El XLSTAT se presenta como complemento a las hojas de cálculo de Excel.

STATISTICA (StatSoft Europe, 2022)

Este software presenta una interfaz gráfica sencilla para realizar multitud de análisis univariantes y multivariantes con grandes volúmenes de datos. Al tener integrados R y

Python, facilita la compatibilidad con otro software.

Este software se encuentra incluido en TIBCO® Data Science.

OriginPro (OriginLab Corporation, 2022)

El software OriginPro posee una interfaz gráfica que contiene un gran número de análisis multivariantes y representaciones gráficas asociadas y es ampliamente utilizado por un gran número de empresas.

El menú para realizar el RDA en este software también incluye la posibilidad de presentar estadísticas básicas, valores y vectores propios y las puntuaciones.

Aunque es posible realizar representaciones gráficas asociadas a estos análisis, no incluirá las representaciones biplot.

CANOCO (Šmilauer, 2012)

El programa desarrollado en el ámbito de la Ecología y ya presentado en capítulos anteriores para la realización de biplots, permite realizar un gran número de análisis multivariantes ampliamente utilizadas en su campo de aplicación.

Uno de los análisis que son muy utilizados en el ámbito de la Ecología es el Análisis de la Redundancia para datos continuos, y por lo tanto, puede encontrarse dentro de CANOCO, así como su representación biplot asociada.

XLSTAT: Complemento estadístico de Excel (Addinsoft, 2022)

XLSTAT es una de las herramientas básicas que complementan a Excel para realizar análisis estadísticos un poco más complejos. Este paquete incluye desde herramientas

que facilitan la preparación de las bases de datos, hasta pruebas multivariantes no paramétricas o métodos de agrupación y minería de datos.

Para realizar el RDA en este software será necesario utilizar, dentro de las funciones avanzadas, el menú de datos multibloque. El gráfico RDA que presenta, es similar al biplot si presentamos los individuos, lo que denominan RDA triplot.

5.5.2. Paquetes de R (R Core Team, 2021)

Por último, se presentan algunos paquetes de R que permiten realizar el Análisis de la Redundancia.

easyCODA (Greenacre, 2018)

El paquete easyCODA posee la función "*RDA*" que permite calcular el Análisis de la Redundancia, tanto ponderado como no ponderado, de una tabla de datos composicionales de las muestras.

Este paquete solo puede ser utilizado para datos composicionales, en caso de no poseer datos composicionales deben ser transformados previamente. La función "*PLOT.RDA*" permitirá realizar una representación biplot de los resultados.

vegan (Oksanen *et al.*, 2017)

Igual que en las ocasiones anteriores en las que se ha presentado el paquete vegan, la función para el Análisis de la Redundancia ("`rda`"), necesita trabajar introduciendo el modelo que queremos estudiar.

Aunque la función permite introducir como predictores variables categóricas, serán empleadas dentro del modelo como variables dummy.

La representación gráfica realizada también será un biplot que contiene, en el caso de predictores categóricos, cada una de las categorías como variable.

MultBiplotR (Vicente-Villardon, 2021)

En el paquete MultBiplotR, que también se ha presentado en el capítulo 3 para realizar las representaciones biplot de un gran número de técnicas de análisis multivariantes, se ha incluido las funciones necesarias para la realización de la nueva técnica presentada en este capítulo y su representación biplot asociada.

Se utilizará la función "`ConstrainedLogisticBiplot`" los cálculos del Análisis de la Redundancia para datos de respuesta binaria y su representación asociada a la que hemos denominado biplot logístico restringido. Será necesario aportar una matriz de variables binarias como respuesta y una matriz de datos continuos como variables predictoras.

5.6. Ejemplo Análisis de la Redundancia para datos binarios

Con el fin de ilustrar las técnicas presentadas en este capítulo , se ha seleccionado un conjuntos de datos sobre la abundancia de arañas en un área de dunas de los Países Bajos.

Se realizarán los análisis con el software R (R Core Team, 2021), utilizando el paquete MultBiplotR (Vicente-Villardon, 2021), que contiene las funciones para aplicar esta nueva técnica.

5.6.1. Arañas

Este ejemplo fue publicado originalmente por Aart y Smeenk-Enserink (1974). Esta base de datos ya ha sido empleada en diversos artículos para ilustrar técnicas de ordenación, cabe destacar entre ellos el publicado por Braak (1986) en su artículo sobre el Análisis Canónico de Correspondencias.

Esta base de datos contiene la distribución de un conjunto de arañas lobo en una zona de dunas de los Países Bajos. Este tipo de arañas, también denominadas tarántulas europeas, poseen un gran tamaño (19-30 mm), siendo las hembras más grandes que los machos. Se encuentran en lugares generalmente cálidos y secos, con piedras y ramas, aunque algunas especies se han adaptado a climas más húmedos y pueden llegar a sumergirse en el agua. Crean su madriguera en una grieta donde hibernan los meses más fríos del año. Las hembras pueden tener hasta 4 años de vida, generalmente están cerca de la madriguera y la protegen con ramas y seda creando un embudo que ayuda a los machos a seleccionar a su pareja. Por el contrario, los machos viven en torno a 2 años y fallecen al alcanzar su madurez sexual, se desplazan por el terreno para cazar a sus presas y alimentarse. Poseen 8 ojos en tres filas, y una gran sensibilidad a las vibraciones, su gran capacidad visual y su sensibilidad les ayudan en la caza de insectos pequeños de los que se alimentan. Por regla general huyen de los animales grandes. Aunque son venenosas, su veneno no es muy potente y, generalmente, las reacciones en humanos son pequeñas.

Base de datos

Los datos originales contenían el recuento de la abundancia de 12 especies de arañas lobo capturadas en 100 puntos de muestreo diferentes, utilizando trampas, en un área de dunas de los Países Bajos.

Hemos utilizado las abundancias recogidas en 28 de los 100 puntos, donde se midieron también las variables ambientales. Las 12 especies que se han estudiado se recogen

en la tabla 5.1 con los códigos utilizados en los gráficos.

Código	**Especie**		**Código**	**Especie**
Arctl	*Arctosa lutetiana*		Trct	*Trochosa terricola*
Prdl	*Pardosa lugubris*		Alpcn	*Alopecosa cuneata*
Zrsp	*Zora spinimana*		Prdm	*Pardosa monticola*
Prdn	*Pardosa nigriceps*		Alpcc	*Alopecosa accentuata*
Prdp	*Pardosa pullata*		Alpf	*Alopecosa fabrilis*
Allb	*Aulonia albimana*		Arctp	*Artosa perita*

Tabla 5.1: Especies de arañas con la codificación realizada en la aplicación práctica.

Antes de procesar los datos, serán convertidos en presencias (1) y ausencias (0) para ajustarlos al tipo de datos que son necesarios para la técnica propuesta en este capítulo. Las presencias y las ausencias de cada especie de araña en cada uno de los puntos de muestreo serán nuestras variables respuesta.

Nuestros predictores serán las variables ambientales que permiten explicar la presencia o ausencia de cada especie. En los 28 puntos de muestreo se han medido seis variables ambientales recogidas en la siguiente lista, la codificación utilizada en los resultados se encuentra entre paréntesis después del nombre de la variable:

- Contenido de agua (Watcont)
- Arena (Barsand)
- Cobertura de musgo (Covmoss)
- Reflexión de la luz (Ligrefl)
- Ramitas caídas (Falltwi)
- Hierbas de cobertura (Coverher)

Objetivos del ejemplo

Para el ejemplo de las especies de arañas los objetivos serán los siguientes:

Objetivo 1. Comparar los resultados obtenidos con el biplot con y sin restricciones.

Objetivo 2. Analizar las relaciones entre las diferentes especies de arañas lobo estudiadas.

Objetivo 3. Estudiar el comportamiento de las variables ambientales medidas en cada uno de los puntos de muestreo del estudio.

Objetivo 4. Relacionar la presencia y ausencia de cada uno de los tipos de arañas y las variables ambientales.

Objetivo 5. Identificar los posibles grupos de lugares de muestreo con características similares.

Objetivo 6. Analizar cada uno de los grupos de lugares de muestreo estableciendo sus características medidas a través de las variables ambientales y las especies presentes en cada uno de ellos.

Metodología

Primero aplicaremos el biplot logístico sin restricciones. En la versión no restringida del biplot obtendremos las puntuaciones para los lugares de muestreo y las especies sin usar explícitamente la información aportada por las variables ambientales, es decir, incluirá las respuestas que miden la abundancia de las especies de arañas y se proyectarán sobre ellas las variables ambientales.

A continuación, realizaremos la versión restringida de este mismo biplot, en ella obtendremos las puntuaciones de los lugares de muestreo y las especies de arañas como combinaciones lineales de las variables ambientales.

Como medida de la bondad de ajuste para cada una de las especies usaremos la $pseudo - R^2$ de Nagelkerke y el porcentaje de individuos de correctamente clasificados en cada una de las columnas de la matriz binaria. Será utilizada con ambas versiones del biplot, la restringida y la no restringida.

Resultados

Comenzaremos presentando el biplot sin restricciones, figura 5.2, en la que, como mencionábamos anteriormente, se ha construido el biplot logístico de las especies de arañas, representadas con vectores negros, con los puntos de muestreo en color azul, y proyectando sobre ella las variables ambientales en color rojo.

Observamos que los puntos de muestreo están dispersos por el gráfico, los grupos de puntos de muestreo más grandes son de dos o tres lugares. En cuanto a las especies de arañas observamos que las escalas están muy concentradas en el centro en casi todos los casos y la variabilidad es pequeña, lo que quiere decir que es posible distinguir claramente entre las presencias y las ausencias de las distintas especies.

Podría desarrollarse detalladamente cada una de las relaciones, pero nos centraremos en el gráfico creado para el caso restringido.

A continuación encontraremos, en la figura 5.3, el biplot asociado al análisis de la redundancia para datos de respuesta binaria desarrollado en este capítulo. Será de interés estudiar el comportamiento tanto de los espacios de muestreo, representados en azul, como las especies de las arañas, presentadas como vectores negros, y las variables ambientales que, igual que en el biplot anterior, se representan en vectores de color rojo.

En el triplot restringido de la figura 5.3 observamos como los puntos de muestreo quedan agrupados de diferentes bloques en función de sus características ambientales. Las escalas de las especies de las arañas se encuentran más dispersas y, por lo tanto, la variabilidad es mayor. Por último, antes de detallar los resultados obtenidos de este bi-

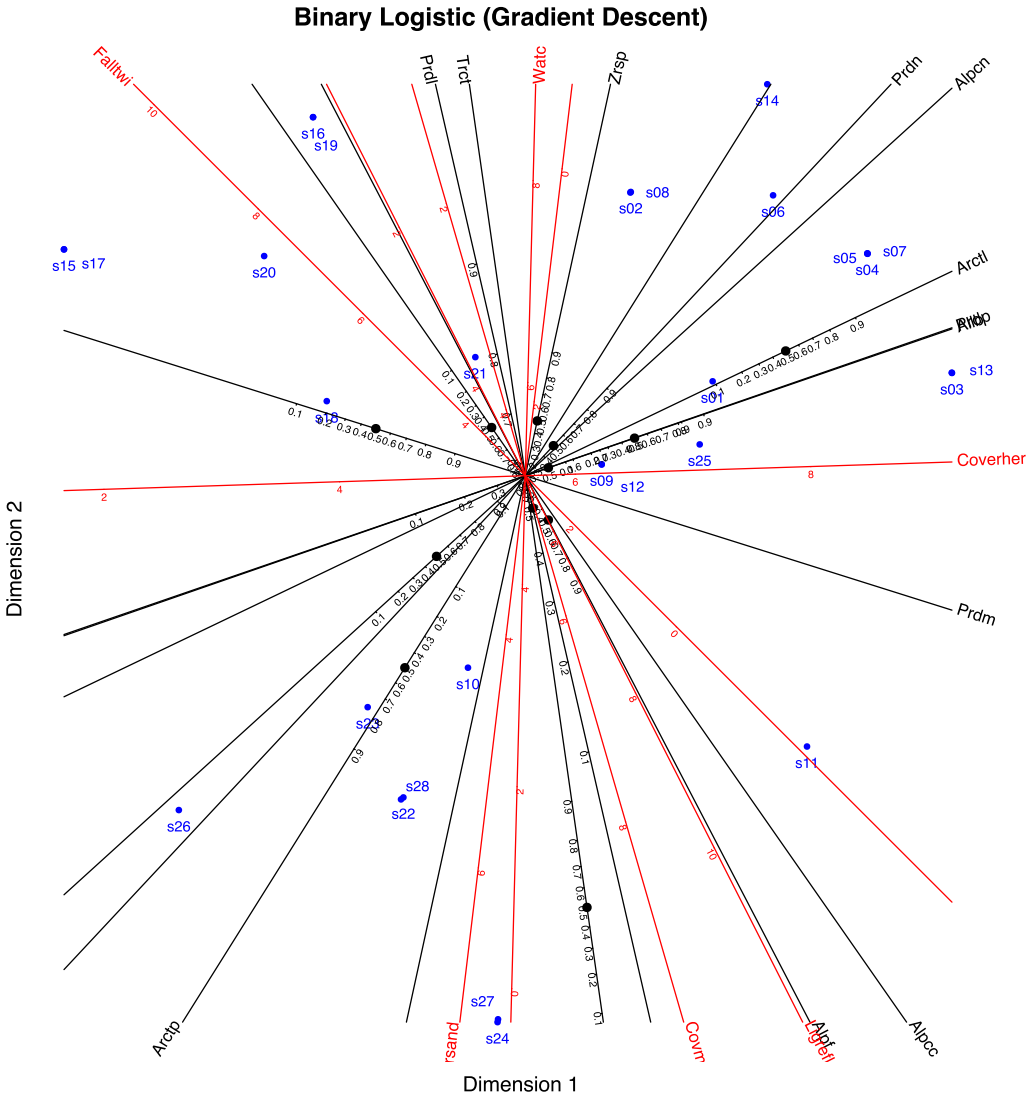

Figura 5.2: Biplot sin restricciones de las especies de arañas con las variables ambientales

plot, destacaremos que la disposición de las variables ambientales respecto a las especies de arañas es diferente.

Comenzaremos estableciendo las relaciones entre las especies de arañas presentadas. En primer lugar, podemos destacar que *Artosa lutetiana* (Arctl), *Pardosa nigriceps* (Prdn), *Pardosa pullata* (Prdp) y *Aulonia albimana* (Allb) están directamente relacionadas, con una relación fuerte entre ellas. También se puede establecer que la relación entre *Pardosa monticola* (Prdm), *Alopecosa cuneata* (Alpcn) y *Zora spinimana* (Zrsp) es directa, se han ordenado de forma que la primera sea la que tiene una relación más fuerte y la última

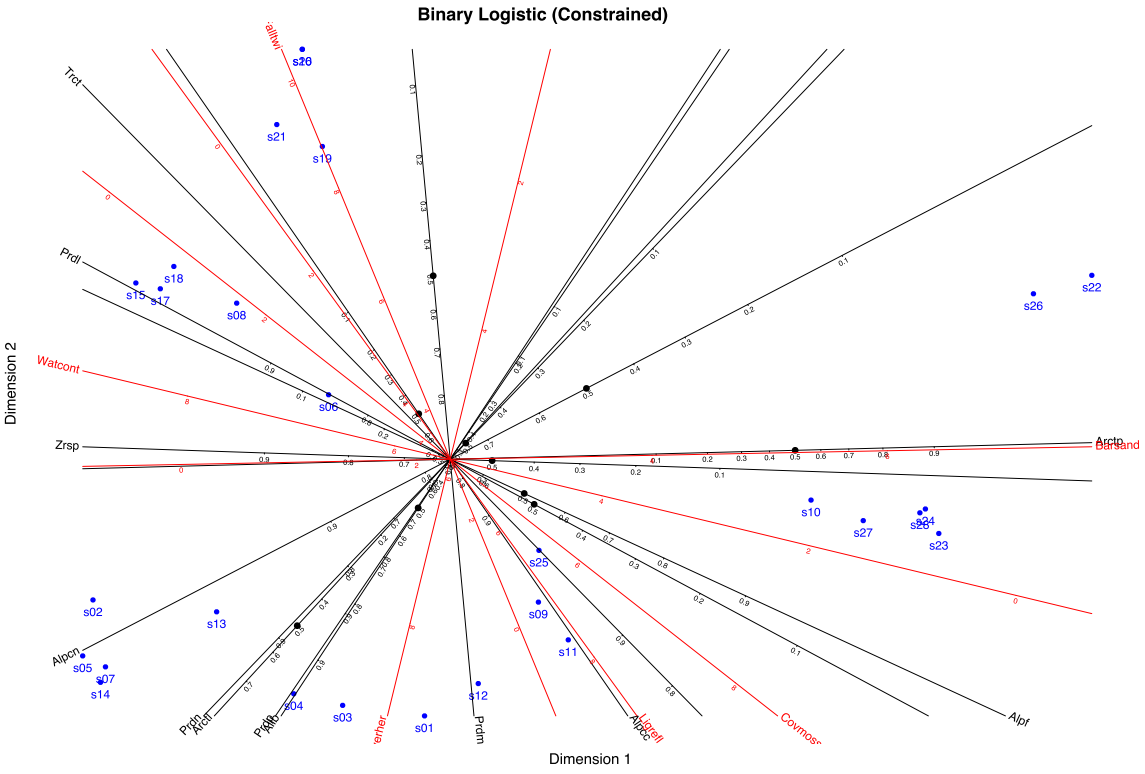

Figura 5.3: Biplot restringido de las especies de arañas con las variables ambientales

la de menor relación. La abundancia de este primer grupo de arañas (Arctl, Prdn, Prdp y Allb) son independientes de la presencia de *Trochosa terricola* (Trct), tiene una relación directa, pero casi nula, con *Alopecosa accentuata* (Alpcc) y *Pardosa lugubris* (Prdl), e indirecta, pero muy débil, con *Alopecosa fabrilis* (Alpf).

Es importante destacar el comportamiento de la *Artosa perita* (Arctp) que tiene una relación muy estrecha e inversa con la *Zora spinimana* (Zrsp), es decir, la presencia de una de estas especies conlleva que la probabilidad de que se encuentre la otra es muy baja. La presencia de esta especie es independiente de si se encuentra en ese punto la *Pardosa monticola* (Prdm). Aunque la relación no es muy fuerte, está directamente relacionada con *Alopecosa fabrilis* (Alpf) y *Alopecosa accentuata* (Alpcc) e inversamente relacionada con todas las demás.

En muchos casos la presencia de una especie no está relacionada con la abundancia de otra, por ello en la tabla 5.2 se recogen aquellas especies que se puede observar en el

gráfico que no tiene relación entre sí o la relación es muy débil.

Relaciones independientes		
Artosa lutetiana (Arctl)	-	*Trochosa terricola* (Trct)
Pardosa lugubris (Prdl)	-	*Pardosa pullata* (Prdp)
Pardosa lugubris (Prdl)	-	*Aulonia albimana* (Allb)
Zora spinimana (Zrsp)	-	*Pardosa monticola* (Prdm)
Pardosa nigriceps (Prdn)	-	*Trochosa terricola* (Trct)
Pardosa pullata (Prdp)	-	*Alopecosa fabrilis* (Alpf)
Aulonia albimana (Allb)	-	*Alopecosa fabrilis* (Alpf)
Alopecosa cuneata (Alpcn)	-	*Alopecosa accentuata* (Alpcc)
Pardosa monticola (Prdm)	-	*Artosa perita* (Arctp)

Tabla 5.2: Resumen de la independencia entre las presencias y ausencias de las diferentes tipos de arañas

En cuanto a las variables ambientales podemos observar que el contenido de agua y las hierbas de cobertura del sitio de muestreo son independientes. El contenido de agua del terreno está directamente relacionado con las ramitas caídas, aunque su relación no es muy fuerte, sin embargo con la arena desnuda tiene una relación muy estrecha e inversa. Con el resto de las variables ambientales la relación del contenido de agua también es inversa, aunque más débil que en los casos anteriores.

También es posible establecer la relación entre las especies de arañas y las variable ambientales. A continuación, caracterizaremos cada una de las especies en función de las variables ambientales medidas.

***Artosa lutetiana* (Arctl)** La relación con la cantidad de hierba de cobertura es fuerte y directa. Un poco más débil, pero también directa, es la relación con el contenido de agua. Los reflejos de luz y la cobertura de musgo no están muy relacionados con esta especie de araña, la primera de forma directa y la segunda de forma inversa.

Las variables restantes se relacionan de forma inversa con este tipo de araña, más fuerte con la arena descubierta que con las ramitas caídas.

Hay probabilidades altas de que esté presente esta especie de araña cuando el contenido de agua es alto (mayor que 6,5), el de arena desnuda es bajo (menor que 1) y el de la cobertura de musgo es medio (4). Tendrá valores medios de la reflexión de la luz (5 a 6,5), bajos de ramitas caídas (0 a 2) y altos en hierbas de cobertura (mayores que 8).

***Pardosa lugubris* (Prdl)** La relación con el contenido de agua y con las ramitas caídas es directa, mientras con la cobertura de musgo, la reflexión de la luz y la arena desnuda es inversa, en ambos casos se han ordenado las relaciones de más fuertes a más débiles. La relación con las hierbas de cobertura es inversa, pero prácticamente nula.

Es probable la presencia de esta especie cuando los valores del contenido de agua son medios o altos (5 a 8), los de arena desnuda son bajos (0 a 3), hay medio bajo contenido de cobertura de musgo (0 a 5) y medios bajos de la reflexión de la luz y de ramitas caídas (2 a 6), los valores de las hierbas de cobertura son medios (5 a 6).

***Zora spinimana* (Zrsp)** Esta especie está estrecha y directamente relacionada con el contenido de agua, su relación también es directa con las ramitas caídas y la cobertura de hierba, aunque sus relaciones son más débiles, e incluso, las hierbas de cobertura son casi independientes. El resto de variables están inversamente relacionadas con este tipo de araña, la arena desnuda está intimamente relacionada, seguida de la cobertura de musgo y los reflejos de luz.

Existe una alta probabilidad de encontrar esta araña en zonas con un contenido de agua medio alto (5 a 8), poca arena desnuda (menos de 3) y una cobertura de musgo media baja (2 a 5). La reflexión de la luz, las ramitas caídas y la cobertura de hierba en las zonas donde se encuentra esta araña también será media baja (3 a 5).

***Pardosa nigriceps* (Prdn)** Este tipo de araña está relacionada de forma directa con la cobertura de hierbas y el contenido de agua. La relación con la arena desnuda y las

ramas caídas es inversa. En ambos casos se han ordenado las variables de mayor a menor relación con la araña. Las dos variables restantes presentan una relación muy débil con esta especie, prácticamente independiente, la reflexión de la luz de forma directa y la cobertura de musgo de forma inversa.

El contenido de agua en las zonas en las que reside esta especie de araña es medio alto (5 a 7), la arena desnuda es escasa (menor a 2) y la cobertura de musgo es media (4). Los valores que recibe la reflexión de la luz son medios (5), los de las ramitas caídas son bajos (1 a 3) y los de la cobertura de hierba serán altos (mayores a 6).

***Pardosa pullata* (Prdp)** Está directamente relacionada con las hierbas de cobertura, el contenido de agua y la reflexión de la luz, de mayor a menor relación, e inversamente relacionado con la arena desnuda y la ramitas caídas. Es independiente de la cobertura de musgo.

El contenido de agua es medio alto (5 a 7), el de arena desnuda es bajo (menos de 2,5), y la cobertura de musgo media (4 a 5). También será media la reflexión de la luz (5) y las ramitas caídas (5 a 6) y alta la cobertura de hierba (mayor que 6).

***Aulonia albimana* (Allb)** La forma de relacionarse de esta especie de araña con las variables ambientales es idéntica a la de la especie anterior.

Este tipo de araña se encuentra en lugares con un contenido de agua medio (6), la arena desnuda es baja (1 a 2) y la cobertura de musgo es media (4 a 5). También tendrá valores medios la reflexión de luz (5 a 6), bajos las ramas caídas (menos de 2) y altos en la cobertura de hierba (mayores de 6).

***Trochosa terricola* (Trct)** Este tipo de tarántula está directamente relacionada con las ramitas caídas y el contenido de agua. Su relación es inversa con la reflexión de la luz, la cobertura de musgo y la de hierba y con la arena desnuda. Igual que en los casos anteriores, están ordenadas las variables de mayor a menor relación.

La probabilidad de presencia de esta especie será mayor del 80 % cuando los valores del contenido de agua sean medios altos (mayores que 3), medio bajo contenido de arena desnuda (menor que 4), medio alto de cobertura de musgo y reflexión de la

luz (menor que 7), medios altos valores de ramitas caídas (mayores que 0) y medios de cobertura de hierba (menor que 8).

***Alopecosa cuneata* (Alpcn)** Esta especie se encuentra relacionada directamente con la cobertura de hierba y el contenido de agua e inversamente con la arena desnuda y la cobertura de musgo. Los reflejos de luz y las ramitas caídas son casi independientes, su escasa relación con este tipo de araña es inversa.

El contenido de agua en las zonas en las que se encuentran este tipo de arañas es medio alto (5 a 7), la arena desnuda será escasa (1 a 3), la cobertura de musgo medio (3 a 5), al igual que la reflexión de la luz (4 a 5) y las ramitas caídas (5). Las hierbas de cobertura sin embargo tendrán valores medios altos (5 a 8).

***Pardosa monticola* (Prdm)** En este caso, la relación será directa, de mayor a menor proximidad, con las hierbas de cobertura, la reflexión de la luz, la cobertura de musgo y la arena desnuda, e inversa con las ramitas caídas y el contenido de agua. Las probabilidades de que esté presente este tipo de araña serán altas cuando los valores del contenido de agua sea medio (6), las de la arena desnuda sean bajas (2) y los de la cobertura de musgo medios bajos (3 a 4). En estas zonas también habrá valores medios bajos en los reflejos de luz (3 a 4), medio altos en las ramitas caídas (4 a 6) y medios bajos de hierbas de cobertura (3 a 5).

***Alopecosa accentuata* (Alpcc)** Este tipo de araña está estrecha y directamente relacionada con los reflejos de luz de la zona en la que se encuentran. Su relación también es directa con ambas coberturas, tanto la de musgo como la de hierba, y con la arena desnuda, sin embargo, está inversamente relacionada con las ramitas caídas y el contenido de agua.

Las probabilidades de que en el lugar se encuentre este tipo de araña son altas si el contenido de agua es medio alto (5 a 6), la arena descubierta es media baja (2 a 3) y la cobertura de musgo es medio alto (4 a 6). Serán también medios altos los valores de la reflexión de luz (4 a 6) en este caso, medias bajos los de las ramitas caídas (1 a 4) y medios altos en las hierbas de cobertura (5 a 7).

***Alopecosa fabrilis* (Alpf)** Las relaciones directas de este tipo de arañas de mayor a me-

nor son con la cobertura de musgo, los reflejos de la luz y la arena desnuda. Las relaciones inversas de estas tarántulas, igual que en casos anteriores, de mayor a menor, serán con el contenido de agua y las ramitas caídas. La variable restante, cobertura de hierba, tiene una débil relación directa, pueden considerarse independientes.

Para tener una alta probabilidad de encontrar este tipo de especie el contenido de agua debe ser medio bajo (2 a 4), la arena desnuda también debe ser media baja (3 a 5), la cobertura de musgo será media alta (5 a 8), habrá altos reflejos de la luz (mayor que 6), pocas ramitas caídas (menor que 2) y un valor medio alto de hierbas de cobertura (6 a 7).

***Artosa perita* (Arctp)** La relación de esta araña con la arena desnuda es muy estrecha y directa. También se relaciona directamente con la cobertura de musgo y la reflexión de la luz. Se relaciona de forma inversa con el contenido de agua, las ramitas caídas y la cobertura de hierba.

En este caso, la especie se encuentra en zonas con bajo contenido de agua (menor que 2), cantidades medias altas de arena desnuda (mayor que 5), alta cobertura de musgo (mayor que 7), alta reflexión de la luz (mayor que 7), pocas ramitas caídas (menor que 1) y una media baja cobertura de hierba (3 a 5).

Los puntos de muestreo se distribuyen en siete grandes grupos distribuidos a lo largo del plano. A continuación, describiremos los puntos de muestreo que pertenecen a cada uno de ellos, las variables ambientales que los caracterizan y cuáles de las especies de arañas será más probable encontrar en cada uno de los grupos.

Grupo 1 Compuesto por los puntos de muestreo s10, s23, s24, s27 y s28.

Se caracteriza por tener bajo contenido de agua (valores menores que 2), una cantidad media alta de arena desnuda (valores entre 5 y 7), una gran cantidad de cobertura de musgo (valores mayores que 7) y de reflexión de la luz (mayores que 8), no habrá ramitas caídas (valores iguales a 0) y una cantidad media de hierbas de cobertura (valores entre 5 y 6).

Será más probable encontrar *Alopecosa accentuata*, *Alopecosa fabrilis*, *Artosa perita*, *Pardosa monticola* y *Trochosa terricola* que *Aulonia albimana*, *Alopecosa cuneata*, *Arctosa lutetiana*, *Pardosa nigriceps*, *Pardosa pullata*, *Pardosa lugubris* y *Zora spinimana*.

Grupo 2 Compuesto por los puntos de muestreo s09, s11 y s25.

Este grupo se caracteriza por tener en su terreno un contenido de agua medio (valores alrededor de 4), una cantidad de arena desnuda media baja (valores en torno a 3), una cobertura de musgo media (valores en torna a 6), una reflexión de la luz media alta (valores entre 6 y 8), pocas ramitas caídas (valores menores que uno) y una gran cobertura de hierba (valores mayores que 7).

Las probabilidades de encontrar *Aulonia albimana*, *Alopecosa fabrilis*, *Pardosa monticola*, *Pardosa nigriceps*, *Pardosa pullata* y *Trochosa terricola* serán más altas que las de encontrar el resto de las especies.

Grupo 3 Compuesto por los puntos de muestreo s01, s03, s04 y s12.

Las zonas donde se encuentran este grupo de lugares de muestreo tiene un contenido de agua medio (valores entre 4 y 6), poca arena desnuda (valores entre 1 y 3), una cobertura de musgo media (valores entre 4 y 6), una reflexión de la luz media alta (valores entre 6 y 8), sin ramitas caídas (con un valor de 0) y con una gran cobertura de hierba (valores mayores a 8).

La mayoría de las especies de araña se pueden encontrar en las zonas de muestreo de este grupo, tienen probabilidades altas todas excepto *Alopecosa fabrilis* y *Pardosa lugubris* que tienen una probabilidad entorno al 50 % y *Artosa perita* que la probabilidad de encontrarla es casi nula.

Grupo 4 Compuesto por los puntos de muestreo s02, s05, s07, s13 y s14.

En este conjunto de zonas se observan niveles altos del contenido de agua (valores mayores que 7) y no presenta arena desnuda (aproximadamente 0), tampoco habrá una gran cobertura de musgo (valores entre 1 y 3), ni un gran número de ramitas caídas (valores entre 2 y 3). La reflexión de la luz será media (valores entre 4 y 5), sin embargo, tendrá una gran cobertura de hierba (valores mayores a 8).

Las especies de arañas que es más probable encontrar en estas zonas son *Aulonia albimana*, *Alopecosa cuneata*, *Pardosa lugubris*, *Pardosa monticola*, *Pardosa nigriceps*, *Pardosa pullata*, *Trochosa terricola*, *Zora spinimana*.

Grupo 5 Compuesto por los puntos de muestreo s06, s08, s15, s17 y s18.

El contenido de agua en este caso es alto (mayor que 7), mientras que presentan bajos niveles de arena desnuda (menor a 1), de cubierta de musgo (menor que 3) y de reflexión de la luz (menor que 3), habrá una cantidad media alta de ramitas caídas (valores entre 5 y 7) y una cobertura de hierba media (valores en torno a 5).

Es bastante probable que en estas zonas se encuentren las especies *Pardosa lugubris*, *Pardosa nigriceps*, *Pardosa pullata*, *Trochosa terricola* y *Zora spinimana*. La especie *Pardosa monticola* tiene una probabilidad en torno al 50 % en la mayor parte del grupo, sin embargo, en el punto de muestreo s06 aumenta hasta el 80 %. Es poco probable encontrar el resto de especies en este grupo de localizaciones.

Grupo 6 Compuesto por los puntos de muestreo s16, s19, s20, s21.

Este conjunto de zonas se caracteriza por tener gran contenido de agua (valores entre 7 y 8), poca arena desnuda (aproximadamente 1), cobertura de musgo (entre 0 y 1), reflexión de la luz (valores de 0) y cobertura de hierbas (entre 1 y 2), sin embargo los valores de ramitas caídas son muy altos (9).

Solo cuatro especies tienen altas probabilidades de estar en este grupo de zonas, *Alopecosa cuneata*, *Pardosa lugubris*, *Trochosa terricola* y *Zora spinimana*, el resto es poco probable encontrarlas aquí.

Grupo 7 Compuesto por los puntos de muestreo s22, s26.

Al contrario que el grupo anterior, en estas localizaciones el contenido de agua es bajo (valores aproximadamente de 1) y la arena desnuda tiene valores altos (por encima de 8). Posee una gran cobertura de musgo (valores mayores que 8), alta reflexión de la luz (valores alrededor de 8), pocas ramitas caídas (valores en torno a 2) y pocas hierbas de cobertura (valores cercanos a 2).

Será probable encontrar las especies *Alopecosa accentuata*, *Alopecosa fabrilis*, *Artosa perita*, *Pardosa monticola* y *Trochosa terricola* en estas zonas, sin embargo, el

resto de especies, *Alopecosa cuneata*, *Arctosa lutetiana*, *Aulonia albimana*, *Pardosa lugubris*, *Pardosa nigriceps*, *Pardosa pullata* y *Zora spinimana*, es poco probable que se presenten en este tipo de puntos de muestreo.

Para estudiar comparación del biplot restringido de la figura 5.3, que se ha desarrollado en este ejemplo, frente al biplot sin restricciones de la figura 5.2, se presentan a continuación las medidas de ajuste de Nagelkerke, el porcentaje de individuos correctamente clasificados y la varianza explicada de las variables ambientales.

En la tabla 5.3 se puede observar la comparación de la bondad de ajuste de Nagelkerke y el porcentaje de individuos correctamente clasificados para ambos modelos. Se observa que las medidas de bondad de ajuste para las variables binarias son ligeramente mejores para la versión sin restricciones, es decir, captura mejor la estructura de la matriz binaria.

En la tabla 5.4 observamos el porcentaje de varianza de las variables ambientales explicadas por las dimensiones de ordenación de la especie . Se puede comprobar que ahora la versión restringida captura mucho mejor la relación de las presencias y ausencias de las especies con las variables ambientales, porque contiene la parte de la estructura de las especies captada por el ambiente.

Conclusiones del ejemplo

- El biplot restringido aporta mayor información sobre la relación entre los dos conjuntos que el modelo sin restricciones. Se observa que los resultados aportan más información respecto a las variables ambientales aunque se pierda parte de la variabilidad explicada para las especies de arañas.

- La *Artosa perita* es la especie que posee el comportamiento más diferenciado del resto de los tipos de araña relacionándose de forma inversa con la mayor parte de ellas. Destaca su estrecha e inversa relación con la *Zora spinimana*. La relación entre la presencia de *Pardosa nigriceps*, *Arctosa lutetiana*, *Pardosa pullata* y *Aulonia*

	Nagelkerke		% Correctos	
Especie	Sin restricciones	Restringido	Sin restricciones	Restringido
Arctl	1.00	0.77	1.00	0.86
Prdl	0.61	0.64	0.82	0.89
Zrsp	0.98	0.74	1.00	0.93
Prdn	0.96	0.57	0.96	0.68
Prdp	0.86	0.72	0.93	0.89
Allb	1.00	0.74	1.00	0.82
Trct	0.99	0.66	1.00	0.89
Alpcn	0.99	0.63	1.00	0.93
Prdm	0.99	0.75	1.00	0.89
Alpcc	0.98	0.91	1.00	0.93
Alpf	0.96	0.74	1.00	0.86
Arctp	0.99	0.99	1.00	1.00

Tabla 5.3: Ajuste de las columnas en los modelos sin restricciones y restringido. Estas medidas de relación entre las variables observadas como función de las puntuaciones del sitio de muestreo para las primeras dos dimensiones

albimana es muy estrecha e independiente de la presencia de *Trochosa terricola*.

- Se observa que el contenido de agua tiene una relación inversa con la arena desnuda. La cobertura de musgo, la de hierba y la reflexión de la luz están muy relacionadas de forma directa entre ellas y de forma inversa con las ramitas caídas. El contenido de agua también se relaciona de forma directa con las caídas y las hierbas de cobertura, aunque con esta última variable su relación es débil.

- La *Trochosa terricola* es una especie que se encuentra en la mayor parte de las zonas de muestreo ya que se puede adaptar a todas las variables ambientales, sin embargo la *Artosa perita* es más probable encontrarla en zonas con mucha arena desnuda, la *Pardosa nigriceps*, *Arctosa lutetiana*, *Pardosa pullata* y *Aulonia albimana* en zonas con mucha cobertura de hierba, la *Alopecosa accentuata* en lugares con

	Watcont	Barsand	Covmoss	Ligrefl	Falltwi	Coverher
Sin restricciones	76.38 %	52.72 %	54.13 %	75.12 %	65.07 %	54.53 %
Restringido	90.11 %	69.05 %	73.90 %	89.92 %	90.30 %	93.27 %

Tabla 5.4: Porcentaje de varianza de las variables ambientales explicada por la ordenación

una gran reflexión de la luz, la *Alopecosa fabrilis* cuando se encuentre una gran cobertura de musgo, la *Zora spinimana* si hay un gran contenido de agua y de ramitas caídas. Si el terreno mezcla el contenido de agua con un número menor de ramitas caídas las probabilidades de encontrar *Pardosa lugubris* son altas y si el agua se combina con la cobertura de hierba la presencie que es altamente probable encontrar es *Alopecosa cuneata*. Por último, si hay un nivel medio de todas las variables ambientales es altamente probable encontrar la última especie, *Pardosa monticola*.

- Se han identificado 7 grupos de lugares de muestreo. El grupo 1 está formado por las zonas s10, s23, s24, s27 y s28, el grupo 2 por s09, s11 y s25, el grupo 3 por s01, s03, s04 y s12, el grupo 4 por s02, s05, s07, s13 y s14, el grupo 5 por s06, s08, s15, s17 y s18, el grupo 6 por s16, s19, s20 y s21 y el grupo 7 por s22 y s26.

- Cada uno de los grupos tiene unas características concretas.

  El grupo 1 se caracteriza por tener una gran cobertura de musgo y de reflexión de la luz y la presencia de *Alopecosa accentuata*, *Alopecosa fabrilis*, *Artosa perita*, *Pardosa monticola* y *Trochosa terricola*.

  El grupo 2 está caracterizado por una gran cobertura de hierba y reflexión de la luz, las arañas que es más probable encontrar en esta zona son *Aulonia albimana*, *Alopecosa fabrilis*, *Pardosa monticola*, *Pardosa nigriceps*, *Pardosa pullata* y *Trochosa terricola*.

  En el grupo 3 destaca la gran cobertura de hierba que posee y se pueden encontrar la mayoría de las especies, sin embargo es casi nula la probabilidad de encontrar *Artosa perita*.

  El grupo 4 posee un gran contenido de agua y una gran cobertura de hierba y se

suelen encontrar las especies *Aulonia albimana*, *Alopecosa cuneata*, *Pardosa lugubris*, *Pardosa monticola*, *Pardosa nigriceps*, *Pardosa pullata*, *Trochosa terricola*, *Zora spinimana*.

En el grupo 5 solo se distingue el alto contenido de agua, las arañas con más probabilidades de estar presentes son *Pardosa lugubris*, *Pardosa nigriceps*, *Pardosa pullata*, *Trochosa terricola* y *Zora spinimana*.

El grupo 6 se caracteriza por una gran cantidad de agua y un alto número de ramitas caídas, *Alopecosa cuneata*, *Pardosa lugubris*, *Trochosa terricola* y *Zora spinimana* son las especies más probables en estas zonas.

Por último, el grupo 7 destaca por tener una gran cantidad de arena desnuda, de reflexión de la luz y de cobertura de musgo, caracterizan a este grupo las especies *Alopecosa accentuata*, *Alopecosa fabrilis*, *Artosa perita*, *Pardosa monticola* y *Trochosa terricola*.

Capítulo 6

Regresión de Mínimos Cuadrados Parciales para datos de respuesta binaria

Notación

I : Número de individuos.

J : Número de variables predictoras.

K : Número de variables respuesta.

$\mathbf{X}_{(I\times J)}$: Matriz de variables explicativas con I individuos y J variables.

$\mathbf{Y}_{(I\times K)}$: Matriz de variables respuesta con I individuos y K variables.

R : Rango de la matriz $\mathbf{X}$.

S : Rango reducido de las soluciones $\mathbf{X}$.

$\mathbf{P}_{(J\times S)}$: Matriz de cargas de los predictores de $\mathbf{X}$ con J filas y S columnas.

$\mathbf{T}_{(I\times S)}$: Matriz de puntuaciones con I filas y S columnas.

$\mathbf{Q}_{(J\times S)}$: Matriz de cargas de las respuestas $\mathbf{Y}$ con K filas y S columnas.

$\mathbf{E}_{(I\times J)}$: Matriz de residuales de los predictores $\mathbf{X}$.

$\mathbf{F}_{(I\times K)}$: Matriz de residuales de las respuestas $\mathbf{Y}$.

$\mathbf{B}$: Matriz de coeficientes de regresión.

$\mathbf{\Pi}$: Valores esperados de $\mathbf{Y}$ cuando las respuestas son binarias.

$\mathbf{q}_0$: Vector con la intersección para cada variable en el caso de la BLR-PLS.

π_{ij} : Probabilidad esperada que el individuos i tienen en la variable j.

L : Función de coste.

6.1. Introducción

El Análisis de la Redundancia no es la única técnica que utiliza un conjunto de variables predictoras o explicativas para explicar o predecir el comportamiento de un conjunto de variables respuesta.

La Regresión de Mínimos Cuadrado Parciales (PLS-R) es un método muy conocido y utilizado para modelar las relaciones entre dos conjuntos, en especial en el ámbito de las aplicaciones de las ciencias químicas o de las aplicaciones industriales (Anzanello y Fogliatto, 2014) cuando ambos conjuntos de variables son continuos.

Cuando el número de predictores es pequeño y no existen efectos redundantes, es posible utilizar la Regresión Múltiple o Multivariante (MLR) o el Análisis de la Redundancia (RDA) para modelar el comportamiento de las respuestas. Sin embargo, si las condiciones de la aplicación no se cumplen, estos modelos no son apropiados. Un ejemplo particularmente interesante es cuando el número de observaciones es mucho menor que el de predictores, en este caso, los parámetros estimados para la Regresión Lineal Multivariante no existen. Esto ocurre, por ejemplo, en los estudios genómicos, donde un conjunto de datos de expresión génica se puede usar para predecir el tipo de tumor.

La PLS-R es un método que permite construir modelos cuando hay un gran número de predictores, o estos están altamente correlacionados, como se puede ver en Wold *et al.* (2006); Firinguetti *et al.* (2017). Por ejemplo, la PLS-R se utiliza principalmente para predicciones y no se considera un procedimiento apropiado para comprender la relación existente entre las variables, aunque, junto con una representación gráfica adecuada, es posible proporcionar información sobre la estructura de los datos. Esta representación,

como ya hemos afirmado en otros puntos del documento, será el biplot. Oyedele y Lubbe (2015) propuso una representación biplot para las soluciones PLS cuando las respuestas son de tipo continuo, aunque es posible encontrar versiones menos formales publicadas con anterioridad, por ejemplo en Vargas *et al.* (1999) entre otros autores. Recientemente se ha aplicado el PLS-Biplot a datos de efectividad de equipos (Silva *et al.*, 2020).

Para respuestas binarias se utiliza el Análisis Discriminante PLS (PLS-DA) (Barker y Rayens, 2003) que, básicamente, ajusta una Regresión PLS a una variable dummy o ficticia. El razonamiento que existe detrás de este procedimiento es similar a la relación existente entre el Análisis Canónico de Correlaciones y el Análisis Discriminante, esta segunda metodología es un caso particular de la primera, donde uno de los conjuntos de datos está formado por variables dummy o indicadores.

En el contexto de la regresión podríamos usar la Regresión Logísitca (LR), en lugar de una Regresión Lineal Multivariante, para resolver el problema discriminante, capturando de esa manera la naturaleza binaria de las respuestas. aunque se trata de una solución adecuada en muchos casos, la LR tiene las mismas limitaciones que MLR en cuanto al número de individuos y variables y la colinealidad, con una dificultad añadida cuando existe el problema de separación.

El objetivo de este capítulo es generalizar la PLS-R tradicional para el caso en el que existen múltiples respuestas binarias empleando funciones logísticas en lugar de funciones lineales. De la misma forma que la Regresión Lineal puede ser una alternativa del Análisis Discriminante, la técnica propuesta en este capítulo puede ser una alternativa al PLS-DA. Se puede encontrar una generalización para una única variable respuesta binaria en Bastien *et al.* (2005), pero si hay varias variables respuesta, es necesario reducir la dimensión para los datos binarios, es decir, queremos desarrollar una versión de un PLS-2 (PLS con más de 2 variables respuesta) para el caso en el que las variables dependientes son binarias. A este modelo, le asociaremos una representación gráfica adecuada.

Proponemos un método de PLS-R para el caso en que las variables dependientes son un conjunto de respuestas binarias, usando ajustes logísticos en lugar de lineales que recojan a la naturaleza de las respuestas. Denominaremos a este método Regresión Logística Binaria de Mínimos Cuadrado Parciales (PLS-BLR) y será una generalización, para incluir varias respuestas, del modelo propuesto en Bastien *et al.* (2005).

Igual que en el capítulo anterior, será posible asociar representaciones gráficas que contengan las variables predictoras, las respuestas y los individuos de los datos originales sobre el mismo plano. Este triplot será utilizado para evaluar de forma visual la calidad de las predicciones y reconocer las variables que están asociadas a ellas.

La principal diferencia con las metodologías descritas en el capítulo anterior está en la reducción de la dimensionalidad. En el Análisis de la Redundancia se realizará la reducción de la dimensión en la matriz de variables repuestas, mientras en la Regresión de Mínimos Cuadrados Parciales presentada en este capítulo, se realizará la reducción de la dimensión de ambas matrices, la de predictores y la de respuestas.

En la figura 6.1 se encuentra un esquema que resume el procedimiento realizado con las técnicas que se desarrollarán en este capítulo.

Comenzaremos describiendo la base del método, en la sección 6.2 describiremos la Regresión de Mínimos Cuadrados Parciales para datos de respuesta continua (PLS-R). A continuación, en la sección 6.3 desarrollaremos la Regresión Logística de Mínimos Cuadrados Parciales para datos de respuesta binaria (PLS-BLR) como una generalización del PLS-R para datos de respuesta binaria. La sección 6.4 contiene los biplot o triplot asociados a PLS-R (6.4.1) y a PLS-BLR (6.4.2).

Continuaremos mostrando el software que nos permite realizar este tipo de análisis en la sección 6.5. Finalizaremos el capítulo mostrando un ejemplo de con datos reales (sección 6.6).

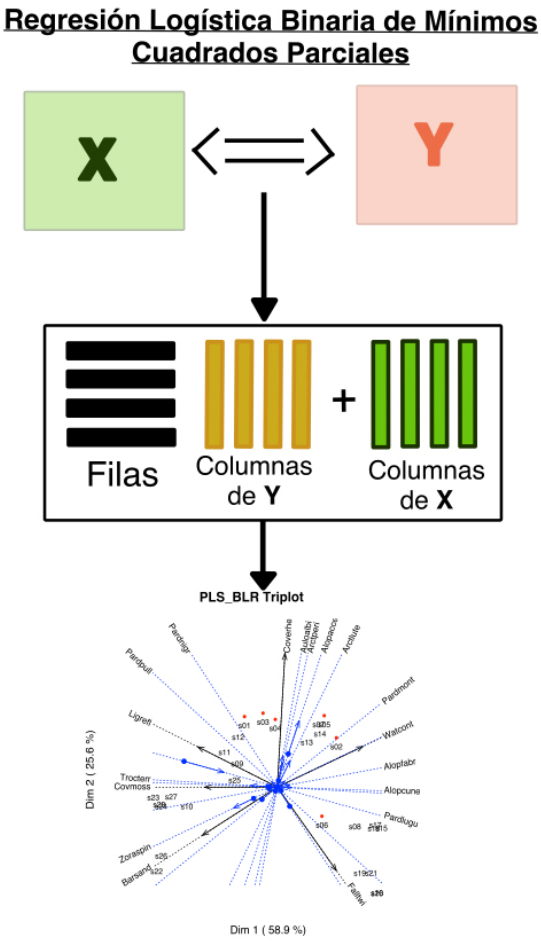

Figura 6.1: Esquema de la realización de la Regresión de Mínimos Cuadrados Parciales

6.2. Regresión de Mínimos Cuadrados Parciales para Datos de respuesta Continua

Partiremos de una matriz de predictores $\mathbf{X}_{I\times J}$ con J variables continuas y una matriz de repuestas $\mathbf{Y}_{I\times K}$ con K variables continuas, ambas medidas en I individuos. Por lo tanto, tendremos J predictores, K respuestas e I individuos.

En esta sección, describiremos la Regresión de Mínimos Cuadrados Parciales cuando ambos conjuntos de datos, tanto los predictores como las respuestas son variables nu-

méricas continuas.

El objetivo de PLS-R es predecir **Y** a partir de **X** y describir la estructura común a ambas matrices. La diferencia con la Regresión Múltiple es que **X** no tiene porque ser de rango completo. Este es el caso, por ejemplo, cuando el número de variables independientes es mayor que el número de individuos o las variables están altamente correlacionadas. Una posible solución para este problema sería usar las Componentes Principales como predictores, sin embargo, no hay garantía de que las Componentes Principales sean óptimas para explicar las respuestas.

Los modelos de PLS buscan dos conjuntos de nuevas variables que sean combinaciones lineales de los predictores, uno y de las respuestas el otro, de forma que las nuevas variables capten lo mejor posible las relaciones entre ambas matrices de datos.

PLS-R combina características de la MLR y PCA obteniendo las componentes de **X** que tienen mayor relevancia en para predecir **Y** mejorando, por tanto, la regresión sobre las Componentes Principales.

6.2.1. Algoritmo NIPALS

El objetivo del PLS es encontrar una combinación lineal de los predictores de **X** que mejor prediga las respuestas de **Y**.

En esta sección vamos a desarrollar el algoritmo NIPALS clásico que se puede encontrar en Wold (1975).

Cuando los datos son continuos, ambos conjuntos de datos están generalmente centrados por columnas y probablemente estandarizados. Tratamos de reducir la dimensión en los dos conjuntos simultáneamente, capturando las relaciones entre ellos lo mejor que sea posible.

Para reducir la dimensión de la matriz $\mathbf{X}$ podemos descomponerla como

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} = E[\mathbf{X}] + \mathbf{E} = \hat{\mathbf{X}} + \mathbf{E}, \tag{6.1}$$

donde $\mathbf{P}$ es una matriz de cargas con J filas por S columnas, $\mathbf{T}$ es una matriz de puntuaciones con I filas y S columnas y $\mathbf{E}$ una matriz de residuales de tamaño $I \times J$, donde S es el número de dimensiones que son de interés para el investigador.

De la misma forma, es posible descomponer la matriz $\mathbf{Y}$ como

$$\mathbf{Y} = \mathbf{T}\mathbf{Q}^T + \mathbf{F} = E[\mathbf{Y}] + \mathbf{F} = \hat{\mathbf{Y}} + \mathbf{F}, \tag{6.2}$$

donde $\mathbf{Q}$ es una matriz de cargas $K \times S$, $\mathbf{T}$ es una matriz de puntuaciones de dimensiones $I \times S$ y $\mathbf{F}$ es una matriz de residuales $I \times K$. Ambas descomposiciones son similares a un PCA o un biplot de cada matriz.

Si queremos descomposiciones separadas, podemos usar el algoritmo NIPALS para las Componentes Principales como se describe en Wold *et al.* (1987). El conjunto de puntuaciones comunes $\mathbf{T}$ se obtiene a partir de $\mathbf{X}$ y se utiliza como predictor para construir el modelo de $\mathbf{Y}$.

Para la matriz $\mathbf{X}$, el algoritmo hace mínima la suma de cuadrados de los residuales, es decir se trata de minimizar la función de pérdida,

$$L = \sum_{i=1}^{I} \sum_{j=1}^{J} e_{ij}^2. \tag{6.3}$$

Partimos de un conjunto de puntuaciones $\mathbf{t}_{(s)}$ para los individuos que pueden ser los valores de cualquiera de las variables o incluso un conjunto de puntuaciones aleatorias. A partir de estas, calculamos las cargas $\mathbf{p}_{(s)}$, volvemos a calcular las puntuaciones $\mathbf{t}_{(s)}$ y alternamos ambos pasos hasta que el procedimiento converja, es decir, hasta que las

cargas y las puntuaciones no cambien de un paso a otro. Llevaremos a cabo el procedimiento de forma recursiva, es decir, vamos estimando las componentes de una en una, calculando la última sobre los residuales de las anteriores. El resultado final será similar al obtenido a partir de la DVS para construir un biplot, si bien, cuando el número de dimensiones de la solución final es elevado, pueden acumularse más errores de redondeo.

Podemos describir el procedimiento en el siguiente algoritmo (Algoritmo 1).

Algorithm 1 Algoritmo de Componentes de X

procedure X-COMPONENTS($\mathbf{X}$, S)
  for $s = 1 \longrightarrow S$ **do**
    $\mathbf{t}_{(s)} \longleftarrow \mathbf{x}_{(j)}$ para algún j ▷ Iniciar: $\mathbf{t}_{(s)}$
    repeat
      $\mathbf{p}_{(s)} \longleftarrow \mathbf{X}^T\mathbf{t}_{(s)} / \parallel \mathbf{X}^T\mathbf{t}_{(s)} \parallel$ ▷ Actualizar: $\mathbf{p}_{(s)}$
      $\mathbf{t}_{(s)} \longleftarrow \mathbf{X}\mathbf{p}_{(s)}$ ▷ Actualizar: $\mathbf{t}_{(s)}$
    until $\mathbf{t}_{(s)}$ no cambia
    $\mathbf{X} \longleftarrow \mathbf{X} - \mathbf{t}_{(s)}\mathbf{p}_{(s)}^T$ ▷ Actualizar: $\mathbf{X}$
  return $\mathbf{T} = \left[\mathbf{t}_{(1)}, \ldots, \mathbf{t}_{(S)}\right]$ y $\mathbf{P} = \left[\mathbf{p}_{(1)}, \ldots, \mathbf{p}_{(S)}\right]$

De la misma forma, para la matriz $\mathbf{Y}$ el algoritmo hace mínima la suma de cuadrados de los residuales, es decir se trata de minimizar la función de pérdida,

$$L = \sum_{i=1}^{I}\sum_{k=1}^{K} f_{ik}^2. \tag{6.4}$$

De la misma forma que antes, utilizaremos el siguiente algoritmo (Algoritmo 2) para obtener una descomposición del tipo deseado.

Algorithm 2 Algoritmo de Componentes de Y

procedure Y-COMPONENTS($\mathbf{Y}, S$)
  for $s = 1 \longrightarrow S$ **do**
    $\mathbf{t}_{(s)} \longleftarrow \mathbf{y}_{(j)}$ para algún j ▷ Iniciar: $\mathbf{t}_{(s)}$
    repeat
      $\mathbf{q}_{(s)} \longleftarrow \mathbf{Y}^T\mathbf{t}_{(s)} / \parallel \mathbf{Y}^T\mathbf{t}_{(s)} \parallel$ ▷ Actualizar: $\mathbf{q}_{(s)}$
      $\mathbf{t}_{(s)} \longleftarrow \mathbf{Y}\mathbf{q}_{(s)}$ ▷ Actualizar: $\mathbf{t}_{(s)}$
    until $\mathbf{t}_{(s)}$ no cambia
    $\mathbf{Y} \longleftarrow \mathbf{Y} - \mathbf{t}_{(s)}\mathbf{q}_{(t)}^T$ ▷ Actualizar: $\mathbf{Y}$
  return $\mathbf{T} = \left[\mathbf{t}_{(1)}, \dots, \mathbf{t}_{(S)}\right]$ y $\mathbf{Q} = \left[\mathbf{q}_{(1)}, \dots, \mathbf{q}_{(S)}\right]$

Se puede observar que la única diferencia entre los dos algoritmos se encuentra en la sustitución de $\mathbf{X}$ por $\mathbf{Y}$ y $\mathbf{P} = \left[\mathbf{p}_{(1)}, \dots, \mathbf{p}_{(S)}\right]$ por $\mathbf{Q} = \left[\mathbf{q}_{(1)}, \dots, \mathbf{q}_{(S)}\right]$.

En este caso, $\hat{\mathbf{X}}$ y $\hat{\mathbf{Y}}$ son, respectivamente, las mejores aproximaciones de rango reducido (S)) de $\mathbf{X}$ e $\mathbf{Y}$, ya que proporcionan la misma solución que la DVS.

De acuerdo con Wold *et al.* (1987), $\mathbf{p}_{(s)}$ es un vector propio de de $\mathbf{X}^T\mathbf{X}$ y el algoritmo es una variante del Power Method descrito, por ejemplo, en Golub y Van Loan (2013), para la diagonalización de matrices. Una de las ventajas de este método frente a la SVD es que mientras la Descomposición en Valores Singulares calcula todas las componentes principales de forma simultánea, el algoritmo NIPALS lo realiza de forma secuencial. Por ello, para grandes conjuntos de datos, la SVD no es viable realizar los cálculos, mientras con el algoritmo NIPALS es mucho más eficiente. Por otra parte, el algoritmo NIPALS pierde la ortogonalidad porque se redondean los errores, y, por lo tanto, es útil para calcular solo unas pocas componentes (Andrecut, 2009).

Como lo que nosotros queremos es relacionar ambos conjuntos de datos, vamos a construir una combinación de ambos procedimientos, para $\mathbf{X}$ (Algoritmo 1) y para $\mathbf{Y}$ (Algoritmo 2), con el fin de obtener el algoritmo NIPALS para la Regresión PLS. En la combinación de ambos algoritmos usaremos un único conjunto de puntuaciones, obte-

niendo finalmente la combinación de variables predictoras que mejor explica una combinación de las variables respuesta.

El procedimiento final se muestra en el siguiente algoritmo (Algoritmo 3). Como antes, se trata de un procedimiento alternado sobre ambas matrices que obtiene las componentes de una en una, la ultima sobre los residuales de las anteriores para ambas matrices.

Algorithm 3 Algoritmo NIPALS

procedure XY-COMPONENTS($\mathbf{X}, \mathbf{Y}, S$)
  for $s = 1 \longrightarrow S$ **do**
    $\mathbf{t}_{(s)} \longleftarrow \mathbf{y}_{(j)}$ para algún j ▷ Iniciar: $\mathbf{u}_{(s)}$
    repeat
      $\mathbf{p}_{(s)} \longleftarrow \mathbf{X}^T\mathbf{t}_{(s)} / \parallel \mathbf{X}^T\mathbf{t}_{(s)} \parallel$ ▷ Actualizar: $\mathbf{p}_{(s)}$
      $\mathbf{t}_{(s)} \longleftarrow \mathbf{X}\mathbf{p}_{(s)}$ ▷ Actualizar: $\mathbf{t}_{(s)}$
      $\mathbf{q}_{(s)} \longleftarrow \mathbf{Y}^T\mathbf{t}_{(s)} / \parallel \mathbf{Y}^T\mathbf{t}_{(s)} \parallel$ ▷ Actualización: $\mathbf{q}_{(s)}$
      $\mathbf{t}_{(s)} \longleftarrow \mathbf{Y}\mathbf{q}_{(s)}$ ▷ Actualizar: $\mathbf{t}_{(s)}$
    until $\mathbf{t}_{(s)}$ no cambia
    $\mathbf{X} \longleftarrow \mathbf{X} - \mathbf{t}_{(s)}\mathbf{p}_{(s)}^T$ ▷ Actualizar: $\mathbf{X}$
    $\mathbf{Y} \longleftarrow \mathbf{Y} - \mathbf{t}_{(s)}\mathbf{q}_{(s)}^T$ ▷ Actualizar: $\mathbf{Y}$
  return $\mathbf{T} = \left[\mathbf{t}_{(1)}, \ldots, \mathbf{t}_{(S)}\right]$, $\mathbf{P} = \left[\mathbf{p}_{(1)}, \ldots, \mathbf{p}_{(S)}\right]$ y $\mathbf{Q} = \left[\mathbf{q}_{(1)}, \ldots, \mathbf{q}_{(S)}\right]$

Con este algoritmo, $\mathbf{P}^T\mathbf{P} = \mathbf{I}$ y $\mathbf{P}$ define un conjunto de vectores ortogonales en la misma forma que en un PCA de la matriz de predictores. También $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$. $\mathbf{T}$ contiene las puntuaciones de $\mathbf{X}$ que mejor explican un conjunto de respuestas. Entonces las puntuaciones para $\mathbf{X}$ se pueden calcular como

$$\mathbf{T} = \mathbf{XP}. \tag{6.5}$$

Un problema importante de estos algoritmos y los siguientes, es que pueden tener varios puntos de inicio y, por lo tanto, es posible llegar a diferentes soluciones. Este algoritmo produce secuencias de valores decrecientes de las sumas de cuadrados de los

residuales, y debe converger, al menos, en un mínimo local.

Si queremos los coeficientes de la regresión para obtener las variables de **Y** en función de las variables de **X**

$$E[\mathbf{Y}] = \hat{\mathbf{Y}} = \mathbf{XB}. \tag{6.6}$$

Teniendo en cuenta que $\hat{\mathbf{Y}} = \mathbf{TQ}^T$, tendremos que

$$\hat{\mathbf{Y}} = \mathbf{TQ}^T = \mathbf{XPQ}^T. \tag{6.7}$$

Por lo tanto, es posible escribir los coeficientes de la regresión en términos de las variables originales de la forma:

$$\mathbf{B} = \mathbf{PQ}^T. \tag{6.8}$$

B contiene, entonces, los coeficientes de regresión que ponen las variables respuesta directamente en función de las variables predictoras.

Las bondades de ajuste de la explicación pueden medirse utilizando descomposiciones de las sumas de cuadrados en las partes explicadas y residuales mediante análogos del coeficiente de determinación para la matriz completa o para cada una de las respuestas por separado.

6.3. Regresión de Mínimos Cuadrados Parciales para Datos de respuesta Binaria

Si las respuestas son binarias, el Modelo Lineal, como ya hemos mencionado en varios lugares en este trabajo, no es el más adecuado, por ello será necesario adaptar la

ecuación de la regresión 6.2 usando transformaciones $logit$ para trabajar con datos binarios.

Bastien *et al.* (2005) propone un modelo PLS para una única respuesta binaria (el equivalente al modelo PLS-1). En esta sección extenderemos el modelo para varias variables binarias, incluyendo un reducción de la dimensión para datos binarios, obteniendo un modelo PLS-2 para variables binarias basado en respuestas logísticas que ha sido publicado por Vicente-Gonzalez y Vicente-Villardon (2022) recientemente. Este método también será una alternativa para el PLS-DA.

Aquí usaremos un procedimiento similar al método descrito en el Algoritmo 3, con las adaptaciones necesarias para incluir datos binarios. Los valores esperados serán ahora probabilidades y usaremos el $logit$ como función de enlace.

Teniendo en cuenta que ahora las respuestas son binarias y por analogía con lo detallado en el capítulo del biplot logístico (Sección 3.3); llamando $\mathbf{\Pi}$ a los valores esperados de $\mathbf{Y}$ ($E[\mathbf{Y}]$), podemos escribir ahora

$$\text{logit}(\mathbf{\Pi}) = \mathbf{1}\mathbf{q}_0^T + \mathbf{T}\mathbf{Q}^T. \tag{6.9}$$

Esta ecuación es una generalización de la ecuación (6.2), excepto porque es necesario añadir un vector $\mathbf{q}_0$ con la intersección para cada variable debido a que, a diferencia del caso anterior, las variables no pueden ser centradas directamente. Cada probabilidad π_{ij} se puede escribir como

$$\pi_{ik} = \frac{e^{\left(q_{k0} + \sum_{s=1}^{S} t_{ks} q_{ks}\right)}}{1 + e^{\left(q_{k0} + \sum_{s=1}^{S} t_{ks} q_{ks}\right)}}. \tag{6.10}$$

Antes de realizar la generalización del Algoritmo 3, es necesario generalizar cada uno de las partes para adaptarlas a datos de respuesta binaria. Es posible mantener el Algoritmo 1 como está, y realizar las modificaciones para las respuestas en el Algoritmo 2. A continuación, se desarrolla la generalización para las componentes separadas del conjunto

binario.

6.3.1. Componentes Separadas para Respuestas Binarias

Para las respuestas, en lugar de usar las sumas de cuadrados del residual, emplearemos la función de coste

$$L = \sum_{i=1}^{I} \sum_{k=1}^{K} [-y_{ik} \log(\pi_{ik}) - (1 - y_{ik}) \log(1 - \pi_{ik})]. \tag{6.11}$$

Obsérvese que esta función es fundamentalmente reciproca a la de máxima verosimilitud usada en la regresión logística estándar para realizar su ajuste. Aquí, interpretaremos la función como función de coste a minimizar, en lugar de como función de verosimilitud a maximizar.

Buscaremos los parámetros $\mathbf{T}$, $\mathbf{Q}$ y $\mathbf{q}_0$ que minimicen la función de coste de la ecuación (6.11). Debido a que este tipo de problemas de optimización no tiene una única forma de buscar su solución, utilizaremos un método iterativo que obtenga una una secuencia de valores decrecientes de la función perdida en cada iteración. Emplearemos, de forma general, el método del descenso del gradiente para la búsqueda de los parámetros. Existen múltiples métodos de optimización final que podemos usar en este contexto como el gradiente conjugado, BFGS, L-BFGS, el gradiente por lotes, el gradiente estocástico, que no serán objeto de estudio de este trabajo. Desde el punto de vista práctico utilizaremos paquetes de optimización pre-programados a los que le proporcionaremos la forma de calcular el coste y los gradientes como mostramos a continuación.

Emplearemos el método del descenso del gradiente, de forma recursiva para realizar los cálculos, es decir, calcularemos los parámetros de cada dimensión separadamente manteniendo los de las anteriores constantes. En la forma general del algoritmo, la ac-

tualización para cada parámetro sería la siguiente:

$$q_{k0} = q_{k0} - \alpha \frac{\partial L}{\partial q_{k0}},$$

$$t_{ks} = t_{ks} - \alpha \frac{\partial L}{\partial t_{ks}}$$

$$q_{ks} = q_{ks} - \alpha \frac{\partial L}{\partial q_{ks}},$$

para un α elegido.

Los gradientes para cada parámetro son

$$\frac{\partial L}{\partial q_{k0}} = \sum_{i=1}^{I} (\pi_{ik} - y_{ik}),$$

$$\frac{\partial L}{\partial t_{is}} = \sum_{k=1}^{K} q_{ks}(\pi_{ik} - y_{ik}),$$

$$\frac{\partial L}{\partial q_{ks}} = \sum_{i=1}^{I} t_{is}(\pi_{ik} - y_{ik}).$$

Podremos entonces, organizar los cálculos en un algoritmo alternado que calcule los parámetros para filas $\mathbf{t}_{(s)} = (t_{1s}, \dots, t_{Is})$ y columnas $\mathbf{q}_{(s)} = (q_{1s}, \dots, q_{Ks})$ para cada dimensión s de forma alterna fijando los parámetros ya obtenidos para cada una de las dimensiones previas, como una forma de obtener las componentes no correlacionadas. Antes, será necesario calcular la constante $\mathbf{q}_{(0)} = (q_{10}, \dots, q_{K0})$ por separado.

El procedimiento sería el que se observa en el Algoritmo 4.

Algorithm 4 Algoritmo para calcular la descomposición de la matriz de datos binarios Y

procedure Y-BINARY-COMPONENTS($\mathbf{Y}, S$)
  Elegir α
  $\mathbf{q}_{(0)} = random$ ▷ Iniciar: $\mathbf{q}_{(0)}$
  repeat
    $q_{k0} \longleftarrow q_{k0} - \alpha \sum_{i=1}^{I} (p_{ik} - y_{ik}), (k = 1, \dots, K)$ ▷ Actualización: $\mathbf{q}_{(0)}$
    $\pi_{ik} \longleftarrow \frac{e^{q_{k0}}}{1+e^{q_{k0}}}; (i = 1, \dots, I; k = 1, \dots, K)$ ▷ Actualizar: $\mathbf{\Pi}$
  until $\mathbf{q}_{(0)}$ no cambie
  for $s = 1 \longrightarrow S$ **do**
    $\mathbf{t}_{(s)} \longleftarrow random$ ▷ Iniciar: $\mathbf{t}_{(s)}$
    $\mathbf{q}_{(s)} \longleftarrow random$ ▷ Iniciar: $\mathbf{q}_{(s)}$
    repeat
      repeat
        $q_{ks} \longleftarrow q_{ks} - \alpha \sum_{k=1}^{I} t_{is}(\pi_{ik} - y_{ik})$ ▷ Actualizar: $\mathbf{q}_{(s)}$
        $\pi_{ik} = \frac{e^{\left(q_{k0} + \sum_{l=1}^{s} t_{kl} q_{kl}\right)}}{1+e^{\left(q_{k0} + \sum_{l=1}^{s} t_{kl} q_{kl}\right)}}$ ▷ Actualizar: $\mathbf{\Pi}$
      until $\mathbf{q}_{(s)}$ no cambie
      repeat
        $t_{is} \longleftarrow t_{is} - \alpha \sum_{k=1}^{K} q_{ks}(\pi_{ik} - y_{ik})$ ▷ Actualización: $\mathbf{t}_{(s)}$
        $\pi_{ik} = \frac{e^{\left(q_{k0} + \sum_{l=1}^{s} t_{kl} q_{kl}\right)}}{1+e^{\left(q_{k0} + \sum_{l=1}^{s} t_{kl} q_{kl}\right)}}$ ▷ Actualizar: $\mathbf{\Pi}$
      until $\mathbf{t}_{(s)}$ no cambie
      $L \longleftarrow \sum_{i=1}^{I} \sum_{k=1}^{K} [-y_{ik} \log(\pi_{ik}) - (1 - y_{ik}) \log(1 - \pi_{ik})]$
    until L no cambie
  return $\mathbf{q}_{(0)} = (q_{10}, \dots, q_{K0})$, $\mathbf{Q} = \left[\mathbf{q}_{(1)}, \dots, \mathbf{q}_{(S)}\right]$, $\mathbf{T} = \left[\mathbf{t}_{(1)}, \dots, \mathbf{t}_{(S)}\right]$

Otros procedimientos, como el gradiente conjugado, evitan la elección de α, sin embargo, se trata de procedimientos más complejos. Desde el punto de vista practico, es posible utilizar rutinas de optimización que mínimicen la función de coste de una manera mucho más eficiente, pero como mencionábamos anteriormente, no es el objetivo de este trabajo. Babativa-Márquez y Vicente-Villardón (2021) desarrolla y describe procesos

de optimización más sofisticados.

En la practica, no es necesario seleccionar un valor particular de α, usando paquetes generales de optimización en los que solo hay que proporcionar las funciones de coste y el gradiente para las rutinas.

Puede haber otro inconveniente al ajustar un Modelo Logístico denominado como problema de separación: cuando hay un hiperplano en el espacio generado por las $\mathbf{T}$ que separa las presencias y ausencias, el estimador no existe (Albert y Anderson, 1984) y tiende a infinito. Incluso cuando la separación no es perfecta (cuasi-separación), el estimador es altamente inestable. La solución habitual es usar una versión penalizada de la función de coste como en Heinze y Schemper (2002). Podemos usar una penalización Ridge (le Cessie y van Houwelingen, 1992) como describiremos luego en el algoritmo conjunto.

6.3.2. Algoritmo Regresión Logística Binaria para Mínimos Cuadrados Parciales

Habiendo establecido los métodos para estimar las componentes separadas, podremos unir ambos procesos para obtener un algoritmo para la PLS-BLR que generalice el procedimiento NIPALS visto anteriormente.

En primer lugar, tendremos que calcular la constante para la parte binaria, ya que no es posible centrar los datos como ocurre en el caso continuo. A continuación, calcularemos las componentes PLS de forma recursiva combinando los Algoritmos 1 y 4 en el Algoritmo 5 de la siguiente forma:

Algorithm 5 PLS Binary Regression

1: **procedure** PLS-BLR($\mathbf{Y}, S$)
2: Elegir α
3: $\mathbf{q}_{(0)} = random$ ▷ Iniciar: $\mathbf{q}_{(0)}$
4: **repeat**
5: $q_{k0} \leftarrow q_{k0} - \alpha \sum_{i=1}^{I} (p_{ik} - y_{ik}), (k = 1, \dots, K)$ ▷ Actualizar: $\mathbf{q}_{(0)}$
6: $\pi_{ik} \leftarrow \frac{e^{q_{k0}}}{1+e^{q_{k0}}}; (i = 1, \dots, I; k = 1, \dots, K)$ ▷ Actualizar: $\mathbf{\Pi}$
7: **until** $\mathbf{q}_{(0)}$ no cambie
8: **for** $s = 1 \longrightarrow S$ **do**
9: $\mathbf{t}_{(s)} \leftarrow random$ ▷ Iniciar: $\mathbf{t}_{(s)}$
10: $\mathbf{q}_{(s)} \leftarrow random$ ▷ Iniciar: $\mathbf{q}_{(s)}$
11: **repeat**
12: $\mathbf{p}_{(s)} \leftarrow \mathbf{X}^T\mathbf{t}_{(s)} / \parallel \mathbf{X}^T\mathbf{t}_{(s)} \parallel$ ▷ Actualizar: $\mathbf{p}_{(s)}$
13: $\mathbf{t}_{(s)} \leftarrow \mathbf{X}\mathbf{p}_{(s)}$ ▷ Actualizar: $\mathbf{t}_{(s)}$
14: **repeat**
15: $q_{ks} \leftarrow q_{ks} - \alpha \sum_{k=1}^{I} u_{is}(\pi_{ik} - y_{ik})$ ▷ Actualizar: $\mathbf{q}_{(s)}$
16: $\pi_{ik} = \frac{e^{\left(q_{k0} + \sum_{l=1}^{s} u_{kl} q_{kl}\right)}}{1+e^{\left(q_{k0} + \sum_{l=1}^{s} u_{kl} q_{kl}\right)}}$ ▷ Actualizar: $\mathbf{\Pi}$
17: **until** $\mathbf{q}_{(s)}$ no cambie
18: **repeat**
19: $t_{is} \leftarrow t_{is} - \alpha \sum_{k=1}^{K} q_{ks}(\pi_{ik} - y_{ik})$ ▷ Actualizar: $\mathbf{t}_{(s)}$
20: $\pi_{ik} = \frac{e^{\left(q_{k0} + \sum_{l=1}^{s} t_{kl} q_{kl}\right)}}{1+e^{\left(q_{k0} + \sum_{l=1}^{s} t_{kl} q_{kl}\right)}}$ ▷ Actualizar: $\mathbf{\Pi}$
21: **until** $\mathbf{t}_{(s)}$ no cambie
22: $L \leftarrow \sum_{i=1}^{I} \sum_{k=1}^{K} [-y_{ik} \log(\pi_{ik}) - (1 - y_{ik}) \log(1 - \pi_{ik})]$
23: **until** L no cambie
24: $\mathbf{X} \leftarrow \mathbf{X} - \mathbf{t}_{(s)}\mathbf{p}_{(s)}^T$ ▷ Actualizar: $\mathbf{X}$
return $\mathbf{q}_{(0)} = (q_{10}, \dots, q_{K0}), \mathbf{Q} = \left[\mathbf{q}_{(1)}, \dots, \mathbf{q}_{(S)}\right], \mathbf{P} = \left[\mathbf{p}_{(1)}, \dots, \mathbf{p}_{(S)}\right]$ y $\mathbf{T} = \left[\mathbf{t}_{(1)}, \dots, \mathbf{t}_{(S)}\right]$

El procedimiento del Algoritmo 5 puede tener el mismo problema de separación que el Algoritmo 4. Actualmente, en un gran número de conjuntos de datos en los que se logra un buen ajuste para las respuestas, es probable que ocurra la separación. En ese

caso, **Q** puede tender a infinito.

Una penalización cuadrática se puede usar en la función de coste

$$L = \sum_{i=1}^{I} \sum_{k=1}^{K} [-y_{ik} \log(\pi_{ik}) - (1 - y_{ik}) \log(1 - \pi_{ik})] + \lambda \sum_{k=1}^{K} \sum_{s=1}^{S} q_{ks}^2. \tag{6.12}$$

La adaptación del gradiente es sencilla

$$\frac{\partial L}{\partial q_{ks}} = \sum_{i=1}^{I} t_{is}(\pi_{ik} - y_{ik}) + 2\lambda q_{ks}. \tag{6.13}$$

Ahora debemos ajustar otro parámetro (λ). La estrategia habitual es probar diferentes valores para seleccionar el óptimo, aunque cualquier valor positivo puede ser suficiente para resolver el problema. Será necesario emplear distintos valores para encontrar el λ óptimo, por ejemplo, $\lambda = (0.1, \ldots, 100)$ y resolver la ecuación (6.13).

El algoritmo puede tener también problemas con el inicio aleatorio. Debería producir valores decrecientes de la función de coste en ambos conjuntos y converger al menos en un mínimo local. Se pueden probar varios puntos de partida para encontrar la mejor solución posible. Desde la experiencia, es posible afirmar que las soluciones desde los diferentes puntos de inicio producen predicciones similares de las respuestas.

6.3.3. Modelo de Regresión Logística

Como en el caso continuo, podemos hacer una regresión de las columnas de **Y** en **T** con una regresión logística. Como se puede ver en el Algoritmo 4, $\mathbf{q}_0$ y **Q** contiene los coeficientes de regresión.

En términos de las variables originales

$$\text{logit}(\mathbf{\Pi}) = \mathbf{1}\mathbf{q}_0^T + \mathbf{T}\mathbf{Q}^T = \mathbf{1}\mathbf{q}_0^T + \mathbf{X}\mathbf{P}\mathbf{Q}. \tag{6.14}$$

Luego

$$\text{logit}(\mathbf{\Pi}) = \mathbf{1}\mathbf{q}_0^T + \mathbf{X}\mathbf{B}^T, \tag{6.15}$$

con

$$\mathbf{B} = \mathbf{P}\mathbf{Q}^T, \tag{6.16}$$

es decir, **B** son los coeficientes de regresión relativos a las variables observadas.

6.4. Representación Biplot asociada a Mínimos Cuadrados Parciales

En esta sección, revisaremos la construcción de un biplot para la visualización de los resultados. Emplearemos los biplots, descritos en el capítulo 3, que estén relacionados con los métodos propuestos aquí.

La representación final combinará los biplot clásicos para los predictores continuos, junto a un biplot logístico para las respuestas binarias, ambas comparten las puntuaciones para los individuos.

6.4.1. Representaciones Biplot asociadas a la Regresión de Mínimos Cuadrados Parciales para Datos de respuesta Continua

La factorización PLS de la matriz **X** de la ecuación (6.1) también define un biplot que puede ayudar en la exploración de nuestros datos, usando **T** y **P**, los marcadores para filas y columnas respectivamente. Utilizaremos la aproximación que mejor predice la respuesta en lugar de la mejor aproximación de bajo rango. Es posible construir un biplot de interpolación usando los mismos datos.

En el Algoritmo 3 observamos que **T** contiene las coordenadas para ajustar las variables respuesta de forma que, usando **Q**, podamos representar dichas variables sobre la representación de los predictores.

Tendremos entonces tres conjuntos de marcadores:

T Las puntuaciones para los individuos

P Los marcadores para los predictores

Q Los marcadores para las respuestas

Como ya habíamos mencionado en el capitulo anterior, la representación de tres matrices de forma simultánea sobre el mismo plano puede ser denominado triplot.

La proyección de las puntuaciones sobre las direcciones de los predictores y respuestas permitirán aproximar los valores esperados como se ha descrito en el capítulo 3.

6.4.2. Representaciones Biplot asociadas a la Regresión Logística de Mínimos Cuadrados Parciales para Datos de respuesta Binaria

Igual que ocurría en la sección 6.4.1, es posible definir un biplot a partir de la factorización PLS de la matriz **X** de la ecuación (6.1), usando los marcadores fila **T** y los marcadores columnas **P**, que ayudará a la exploración de los datos. Esta representación permite, en lugar de realizar la mejor aproximación de bajo rango, explorar la aproximación que mejor predice las respuestas.

En los pasos del 15 al 18 del Algoritmo 5 se puede observar claramente que **T** son también las coordenadas que mejor ajustan las variables binarias de forma que usando $\mathbf{q}_0$ y **Q** se pueden usar para construir un biplot logístico en la representación de los predictores.

Como ocurría en el caso de las respuestas continuas, existen tres conjuntos de marcadores:

T Las puntuaciones para los individuos.

P Los marcadores para los predictores.

Q ($\mathbf{q}_0$) Los marcadores para las respuestas dicotómicas.

De nuevo, la representación de las tres matrices de forma simultánea puede ser denominada triplot.

La proyección de las puntuaciones sobre las direcciones de las variables binarias y numéricas permitirán aproximar las probabilidades y los valores esperados respectivamente.

Existe otro posible biplot,

$$\mathbf{B} = \mathbf{P}\mathbf{Q}^T,$$

que aproxima los coeficientes de regresión.

El producto interno de los *marcadores de X*, **P**, y los *marcadores de Y*, **Q**, aproximan los coeficientes de regresión, es decir, el coeficiente b_{ij} de la variable X_i en la regresión logística para explicar Y_j es

$$b_{ij} = \mathbf{p}_i^T \mathbf{q}_j.$$

Proyectando todos los *marcadores de X* sobre los *marcadores de Y*, por ejemplo la respuesta j-ésima, tendremos la importancia relativa de cada predictor en la explicación de las respuestas.

El biplot BLR-PLS es una herramienta útil para explorar las relaciones entra las variables respuestas y predictoras, como podemos ver en los ejemplos de la siguiente sección. Su interpretación se realizará igual que en los casos anteriores.

6.5. Software para la Regresión de Mínimos Cuadrados Parciales

Igual que en los capítulos anteriores se realizará un resumen de algunos de los software que se pueden utilizar la realizar las regresiones de Mínimos Cuadrados Parciales.

En este caso, como ocurría en el capítulo 5, el software para datos de respuesta binaria solo se encontrará dentro del paquete MultBiplotR (Vicente-Villardon, 2021) del software estadístico R (R Core Team, 2021), ya que, al ser parte de las innovaciones de este trabajo, no se puede encontrar en el resto de los software estadísticos presentados.

6.5.1. Software comercial

Comenzaremos revisando los software con licencia más conocidos que permiten realizar la Regresión PLS.

Algunos de ellos permiten asociar representaciones gráficas, aunque la mayor parte de esos gráficos no son los biplots presentados en este trabajo.

SPSS (IBM Corp., 2021)

El programa SPSS ha introducido en sus versiones más recientes dentro del menú de Regresión la posibilidad de realizarla empleando el método PLS. Este menú incluye la posibilidad de especificar valores de referencia, incluir modelos más complejos o realizar representaciones gráficas, entre otras.

Nuestro conjunto de datos en SPSS debe tener todas las variables respuesta, que serán introducidas en el apartado de variables dependientes, y todas las variables predictoras, que se introducirán en el apartado de variables independientes.

Para poder utilizarlo es necesario tener descargados los complementos de R y Python de SPSS y descargar el módulo asociado a este tipo de análisis.

SAS(SAS Institute Inc., 2022)

Este software tiene integrados varios procedimientos que utilizan los Mínimos Cuadrados Parciales en un único menú. Entre ellos se encuentra la Regresión descrita para datos continuos en este capítulo. Igual que el software anterior permitirá introducir en el modelo variables ficticias.

A diferencia del anterior, permitirá ajustar un modelo para cada grupo de observaciones y realizar un contraste del ML utilizado.

Creará automáticamente gráficos para el estudio de las técnicas englobadas en el menú.

XLSTAT (Addinsoft, 2022)

Este complemento de Excel, además del RDA, permite realizar Regresiones PLS.

Es posible encontrar este tipo de Análisis dentro de las herramientas de XLSTAT en el apartado de modelado de datos, debemos elegir el menú de "*Partial Least Squares Regression*". Dentro de este menú hay más métodos de cálculo, por ello es necesario que el método elegido sea "*PLS-R*". Para fijar el número de dimensiones será necesario entrar en las opciones del menú.

Los resultados de este tipo de análisis serán, entre otros, correlaciones, tablas de observaciones o las puntuaciones. Entre los gráficos obtenidos están representaciones biplot asociadas a estas técnicas.

PLS_Toolbox (Eigenvector Research, 2022)

Esta herramienta asociada a Matlab, con una amplia interfaz gráfica, es una de las herramientas más extendidas dentro de la Quimiometría que contiene métodos multivariantes y herramientas de aprendizaje automático (machine learning).

El software recibe el nombre de las técnicas de Mínimos Cuadrados Parciales, con un gran número de usos dentro del paquete, sin embargo existe otro gran número de técnicas que se pueden realizar con ella.

6.5.2. Paquetes de R (R Core Team, 2021)

Como en los casos anteriores, se van a recoger algunos de los paquetes con los que se puede realizar este tipo de técnicas.

pls (Liland *et al.*, 2022)

El paquete pls es el más utilizado para realizar la mayor parte de los análisis que están relacionados con los Mínimos Cuadrados Parciales.

La función para realizar la PLSR será "`plsr`" o "`mvr`". Es necesario aportar la formula del modelo que buscamos contrastar, el número de dimesiones que queremos retener y los datos con los que vamos a trabajar. De forma opcional, es posible cambiar el método de cálculo de los Mínimos Cuadrados Parciales del Kernel, que está implementado por defecto, a Wide Wernel, NIPALS (como hemos utilizado en este documento) o SIMPLS.

En la representación gráfica de este tipo de resultados es posible elegir los métodos biplot.

caret (Kuhn, 2022)

Este paquete complementará al anterior, permitiendo la realización del Análisis Discriminante PLS y del Análisis Discriminante Sparse PLS.

Se aportará la matriz de datos con la que se quiere trabajar y el resto de argumentos serán los mismos que en el caso anterior.

MultBiplotR (Vicente-Villardon, 2021)

En el paquete MultBiplotR hemos incluido las funciones necesarias para realizar las técnicas explicadas en este capítulo.

La función "*PLSR*" permite realizar la regresión PLS para datos de respuesta continua descrita en 6.2, será necesario aportar la matriz de respuestas y la matriz de predictores. También es posible señalar el número de dimensiones que queremos identificar o la transformación inicial que queremos realizar en los datos. Para representar el Biplot asociado a este análisis, (sección 6.4.1), será necesario realizar los cálculos con la función "*Biplot.PLSR*" y después representar este resultado. Será posible definir los mismos argumentos que en el resto de biplots del paquete MultBiplotR.

Utilizaremos la función "*PLSRBin*" cuando disponemos más de una respuesta binaria, como se ha descrito en la sección 6.3. Igual que en el caso continuo, será necesario introducir en la función la matriz de variables respuesta y la matriz de variables predictoras. Igual que en el caso anterior estos resultados pueden ser representados utilizando los cálculos de la funció "*Biplot.PLSRBIN*", que corresponde a las representaciones biplot explicadas en la sección 6.4.2.

6.6. Ejemplo de Regresión de Mínimos Cuadrados Parciales para Datos Binarios

En esta sección ilustraremos el rendimiento de los métodos propuestos con ejemplos en dos conjuntos de datos diferentes.

El primero corresponderá a la clasificación de vinos españoles de acuerdo con su origen y año. El segundo estará relacionado con la predicción de la presencia de algunas especies de arañas a partir de las características ambientales de los lugares de muestreo. Se trata del mismo conjunto de datos utilizado para ilustrar el Análisis de la Redundancia para datos binarios

Todos los cálculos se han realizado con el software estadístico R (R Core Team, 2021), con el paquete MultBiplotR (Vicente-Villardon, 2021). Como ya hemos mencionado en la descripción de los paquetes, se han desarrollado las funciones específicas para los métodos propuestos. El código para producir los resultados principales se incluirá en los ejemplos de las funciones del paquete.

Los resultados se han obtenido con el método del gradiente conjugado dentro del paquete general de optimización *optimir* (Nash, 2019). A la función *optimir* del mencionado paquete basta con proporcionarle los procedimientos que calculan la función de coste y los gradientes para que proporcione una solución. Hemos seleccionado esta forma de proceder porque los paquetes preprogramados suelen ser bastante eficientes en la búsqueda del óptimo.

6.6.1. Vinos

Este ejemplo se ha utilizado en un gran número de ocasiones para ilustrar diferentes técnicas debido a la simplicidad y la claridad de sus resultados.

Originalmente los datos se utilizaron en un estudio publicado por Rivas-Gonzalo *et al.* (1993) y consiste en un conjunto de vinos de las denominaciones de origen Ribera de Duero y Toro de dos cosechas diferentes (1986 y 1987).

La Denominación de Origen (DO) es un reconocimiento otorgado por el Consejo Regulador de cada zona cuando el vino se hace con uvas de esa misma zona, cumple con unos estándares de calidad, tiene características propias determinadas por la localización geográfica y han trascurrido 5 años desde que se ha reconocido como vino de calidad con indicación geográfica. También pueden recibir el reconocimiento de Denominación de Origen Protegida (DOP), este reconocimiento se otorga a nivel europeo que unifica las DO de cada país. Por último, destacaremos las Denominaciones de Origen Calificados (DOCa), este reconocimiento se reserva para aquellos vinos que han mantenido su DO durante al menos 10 años y cumplen unas características específicas muy altas de calidad dentro de ellas, en España existen dos DOCa, Rioja y Priorat.

En Castilla y León se encuentran las DOP de Ribera del Duero y Toro, con las que trabajaremos, aunque en los últimos años se han ido incluyendo un mayor número de ellas.

Base de datos

Se analizarán 45 vinos jóvenes de Ribera del Duero y Toro de los años 1986 y 1987. Usaremos características químicas que, a diferencia del análisis sensorial común, que es más subjetivo, permiten caracterizar los vinos de una forma más rigurosa. Se han medido parámetros enológicos convencionales, fenoles y variables relacionadas con el color.

Los vinos se obtienen directamente de las bodegas de los Consejos Reguladores.

Una breve descripción de las variables se puede encontrar en la tabla 6.1. La descripción completa de las variables y el conjunto de datos se muestra en el artículo original.

Label	Description
Year	Un factor con niveles *1986* y *1987*
Origin	Un factor con niveles *Ribera* y *Toro*
Group	Un factor con niveles *R86*, *R87*, *T86* y *T87*
A	Contenido alcohólico (porcentaje)
VA	Acidez Volátil - g Ácido acético/l
TA	Acidez total valorable - g ácido tartárico/l
FA	Acidez fija - g ácido tartárico/l
pH	pH
TPR	Fenoles totales - g ácido gálico/l - Folin
TPS	Fenoles totales - Somers
V	Sustancias reactivas a la vainillina - mg catequina/l
PC	Procianidinas - mg cianidina/l
ACR	Antocianinas totales - mg/l - método 1
ACS	Antocianinas totales - mg/l - método 2
ACC	Malvidina - malvidina-3-glucósido mg/l
CI	Densidad del color - Sudraud
CI2	Densidad del color - Glories
H	Tono de vino Color
I	Grado de ionización - Porcentaje
CA	Edad química
VPC	ratio V/PC

Tabla 6.1: Descripción de las variables

Objetivos del ejemplo

Para el ejemplo de las Denominaciones de Origen de los vinos los objetivos serán los siguientes:

Objetivo 1. Determinar cuales son las variables que mejor predicen el origen y el año de los vinos sometidos a estudio.

Objetivo 2. Establecer las variables más relacionadas con cada una de las respuestas, Origen y Año.

Objetivo 2.1. Analizar cuales son las variables que están más relacionadas con la predicción del origen del vino.

Objetivo 2.2. Examinar las variables que tienen una mayor relación con el Año mejorando su predicción.

Objetivo 2.3. Estudiar qué predictores están relacionados con ambas variables de forma simultanea.

Objetivo 3. Establecer las características que identifican a cada vino tanto por su origen como por su año.

Objetivo 3.1. Describir las características de los vinos de Toro respecto a los vinos de Ribera de Duero.

Objetivo 3.2. Especificar las características que cambian en función del año y/o el origen.

Objetivo 4. Presentar la calidad de las predicciones realizadas, tanto para el año como para el origen y de forma conjunta.

Metodología

El artículo original empleaba la Regresión Logística y un HJ-biplot (de forma separada) para buscar las diferencias entre los dos orígenes del vino.

En este caso, las predicciones son altamente colineales, y hay un problema de separación. Por lo tanto, el estimador de máxima verosimilitud en una Regresión Logística clásica no existe, y el método que se propone en este capítulo será más adecuado para el estudio de la relación entre las variables binarias (Origen y Año) y las características químicas, ya que obtenemos las combinaciones lineales de los predictores que mejor expliquen las respuestas.

Como hemos descrito anteriormente, este análisis puede ser representado gráficamente a través de un biplot. Para este caso, en nuestras representaciones los predictores

se presentarán en vectores de color negro, las variables respuesta en vectores de color azul y cada muestra con un punto de otro color.

Resultados

Comenzaremos estudiando la varianza de los predictores explicada por la descomposición. Se presenta en la Tabla 6.2.

	Valor Propio	Var. Expl.	Acumulado
Comp. 1	259.63	32.78	32.78
Comp. 2	194.34	24.54	57.32
Comp. 3	46.33	5.85	63.17
Comp. 4	57.17	7.22	70.39
Comp. 5	44.20	5.58	75.97
...	...	...	...

Tabla 6.2: Varianza de los predictores explicada en la dimensión reducida

Las dos primeras dimensiones explican el 57,32 % de la varianza de los predictores. Esto significa que hay una importante parte de la variabilidad que explica las respuestas. El resto de las componentes PLS son mucho menos importantes para la predicción.

La representación gráfica de las dos primeras dimensiones se muestra en la figura 6.2. La representación incluye no solo las predicciones sino también las respuestas.

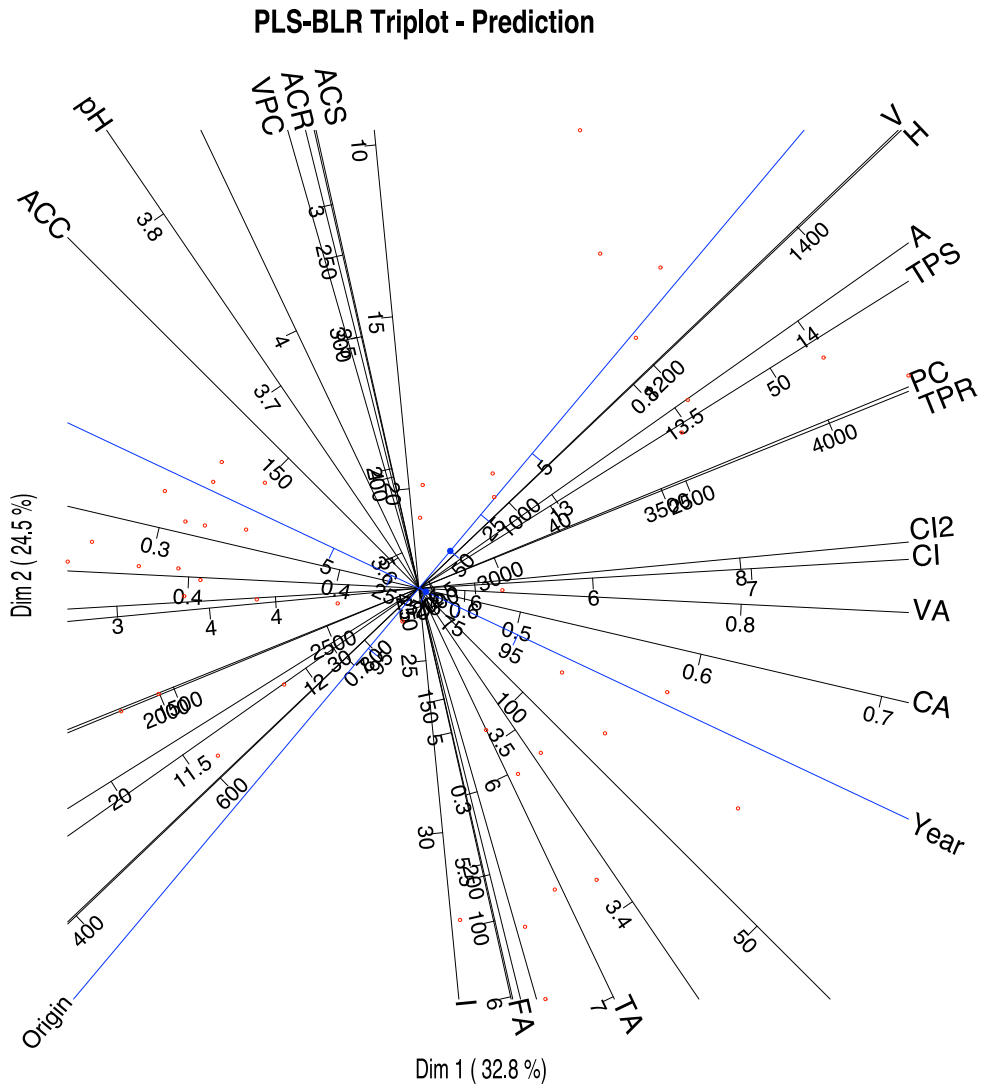

Figura 6.2: Triplot de Predicción PLS para los datos de los vinos

La tabla 6.3 muestra algunas medidas de la calidad de representación para las variables binarias, incluyendo una prueba para la comparación del modelo nulo, tres medidas de pseudo-R^2 y el porcentaje clasificaciones correctas.

	Deviance	g.l.	P-val	Nagelkerke	Cox-Snell	MacFaden	% Correct
Año	40.3754	2	0.0000	0.7930	0.5923	0.6530	86.6667
Denominación	16.4241	2	0.0003	0.4556	0.3058	0.3281	93.3333
Total	56.7995	4	0.0000	0.6577	0.4680	0.5077	90.0000

Tabla 6.3: Medidas de ajuste para cada respuesta

Todos los porcentajes de clasificación correcta superan el 86 % por lo que el modelo tiene capacidad para discriminar las presencias y las ausencias en las dos variables estudiadas. El porcentaje de clasificación correcta es un poco más alto en el caso de la denominación.

La figura puede ayudar para la identificación de las variables más importantes para las predicciones. Podemos seleccionar en el gráfico únicamente las variables con alta calidad de representación, por ejemplo mayor que 0,6 (Figura 6.3).

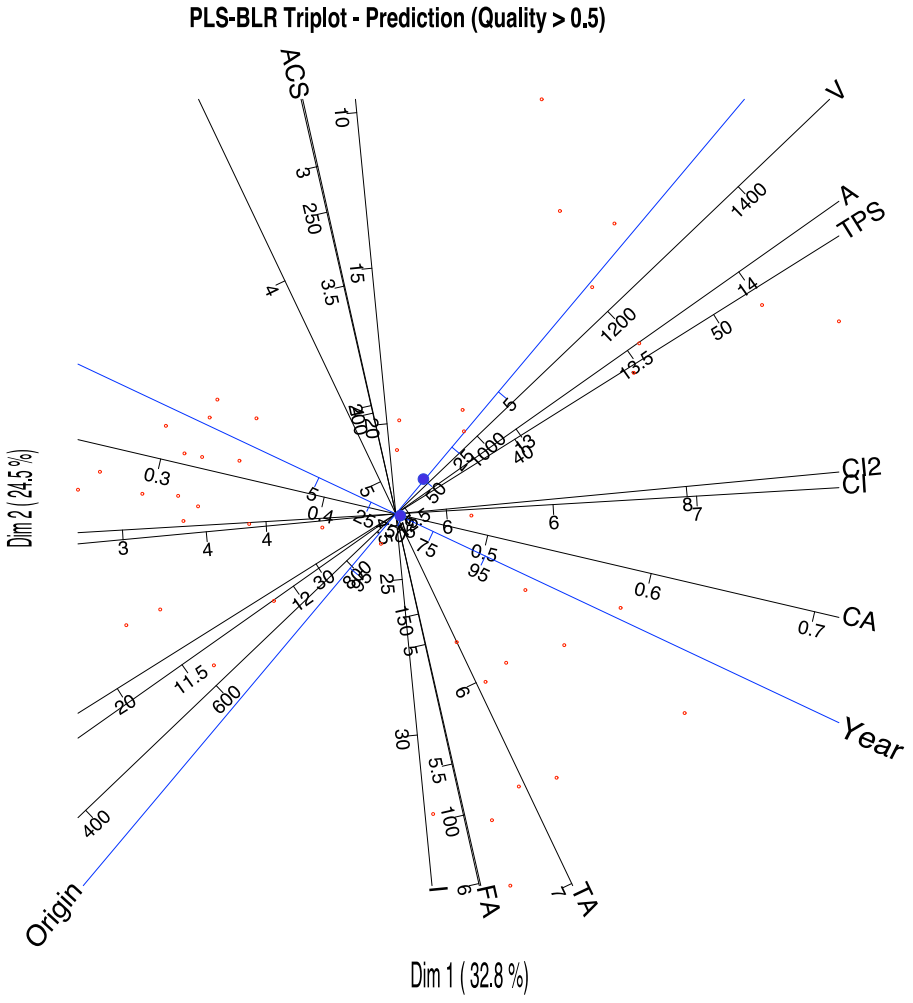

Figura 6.3: Triplot PLS-BLR para los datos del vino que muestra sólo las variables con una predicción superior a 0,6

La calidad de representación de la primera y las dos primeras componentes se recogen

en la tabla 6.4.

	Comp. 1	Comp. 1 + Comp. 2
A	56.40	83.02
VA	33.13	32.96
TA	16.32	65.46
FA	5.15	78.39
pH	14.50	34.87
TPR	48.85	57.54
TPS	58.48	80.88
V	33.20	60.38
PC	48.21	57.11
ACR	2.13	26.96
ACS	4.46	65.60
ACC	32.76	52.63
CI	67.69	68.43
CI2	67.17	68.51
H	16.11	29.14
I	1.00	78.30
CA	83.77	84.44
VPC	0.74	7.11

Tabla 6.4: Calidad acumuladas de las columnas (porcentaje de la variabilidad de la variable correspondiente a la primera dimensión y a la suma de las dos primeras)

Las variables con calidades de representación más altas serán aquellas cuya información está mejor recogida en el gráfico y, por tanto, aquellas que son interpretables para las clasificaciones obtenidas. En orden de acuerdo con su capacidad de predicción serían la Edad química (CA), el Contenido alcohólico (A), los Fenoles totales - Somers (TPS), la Acidez fija (FA), el Grado de Ionización (I), la Densidad del color (CI y CI2), las Antiocianinas totales - método 2 (ACS), la Acidez total valorable (TA), las Sustancias reactivas a la vainillina (V), los Fenoles totales - Folin (TPR), las Procianididas (PC), la Malvidina (ACC), el pH (pH), la Acidez Volátil (VA), el Tono del color del vino (H), las Antiocianinas totales - método 1 (ACR) y el ratio V/PC (VPC).

Podemos tener un gráfico más sencillo eliminando las escalas de cada variable y cambiando las líneas por vectores como en la figura 6.4.

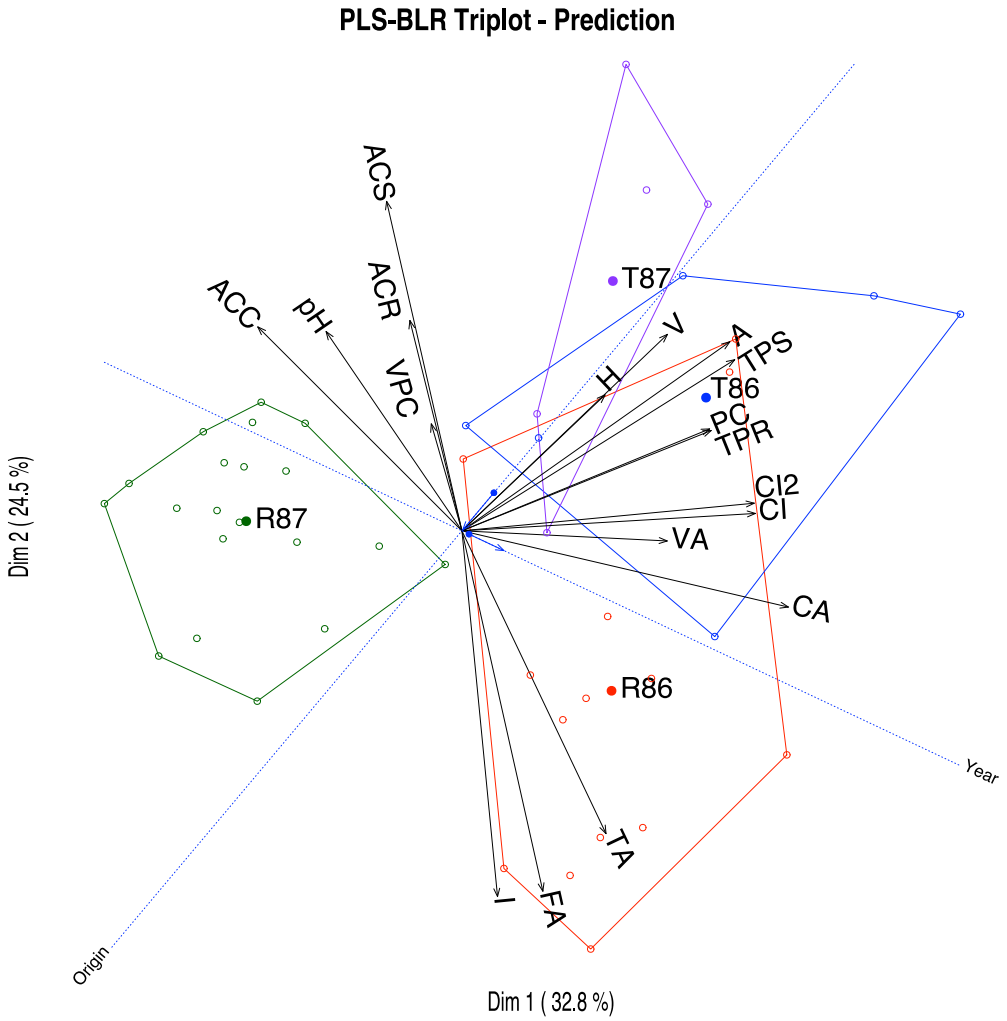

Figura 6.4: Triplot PLS-BLR para los datos del vino con grupos representados como convex hulls

Las escalas deben ser usadas cuando queremos conocer los valores aproximados para cada vino o grupo de vinos, pero para una interpretación general, el último gráfico también es útil y más legible. El paquete de software permite la selección de variables para limpiar la imagen final. Este es el objetivo de explorar el triplot en la pantalla del ordenador, porque a veces es difícil realizar esto mismo con un gráfico en papel.

Una inspección más detallada del gráfico muestra que V, H, A, TPS, PC y TPR están

relacionadas de forma más estrecha con el Origen (Denominación) de los vinos, estas variables tomarán valores más altos para los vinos de Toro que para los de Ribera de Duero. Las variables más relacionadas con el Año son ACC y CA, la primera tendrá valores altos en el caso de los vinos del año 1987 y la segunda en el caso de los vinos de 1986. Las variables CI y CI2, que son en realidad dos medidas del color, se asocian a ambas variables, Origen y Año, siendo mayor en los vinos de Toro del primer año. ACS está asociada con el segundo año y Toro, y FA y TA con el primer año de Ribera de Duero.

En resumen, los vinos de Toro son más oscuros, con mayor graduación alcohólica, con más fenoles y procianidinas, en comparación con los de Ribera de Duero. El color también cambia con el año, así como la edad química y las malvidinas. Los antocianos y la acidez cambian tanto con el origen como con el año.

Obtenemos el 90 % de correctamente clasificados, el 86,67 % para el año y el 93,33 % para el origen. Junto con los valores de la pseudo R^2, podemos afirmar que la predicción es precisa (observe la tabla 6.3). Se puede comprobar gráficamente en la figura 6.5, donde las lineas de puntos son las direcciones que mejor predicen la probabilidad, las flechas muestran la dirección en la que se incrementan las probabilidades y la linea perpendicular limita las regiones de predicción. Las flechas comienzan en el punto de predicción 0,5 y terminan en el punto de predicción 0,75. Si clasificamos un individuo en un grupo cuando la probabilidad esperada es mayor que 0,5, podemos observar que la mayoría de los puntos está en la región de predicción correcta.

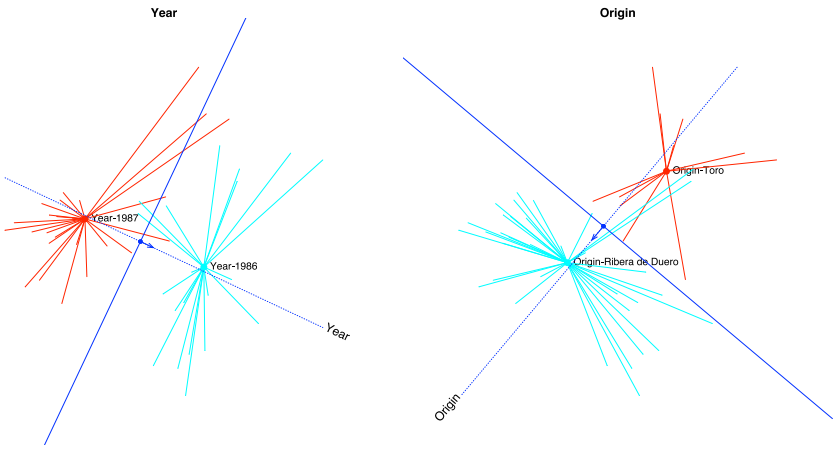

Figura 6.5: Regiones de predicción para cada variable por separado

Finalmente, hemos colocado en el gráfico los convex hulls que contienen los puntos para cada combinación de origen y año así como las regiones que predicen las mismas combinaciones. Podemos observar que la mayoría de los puntos se encuentran en la región predicción correcta, vea la figura 6.6

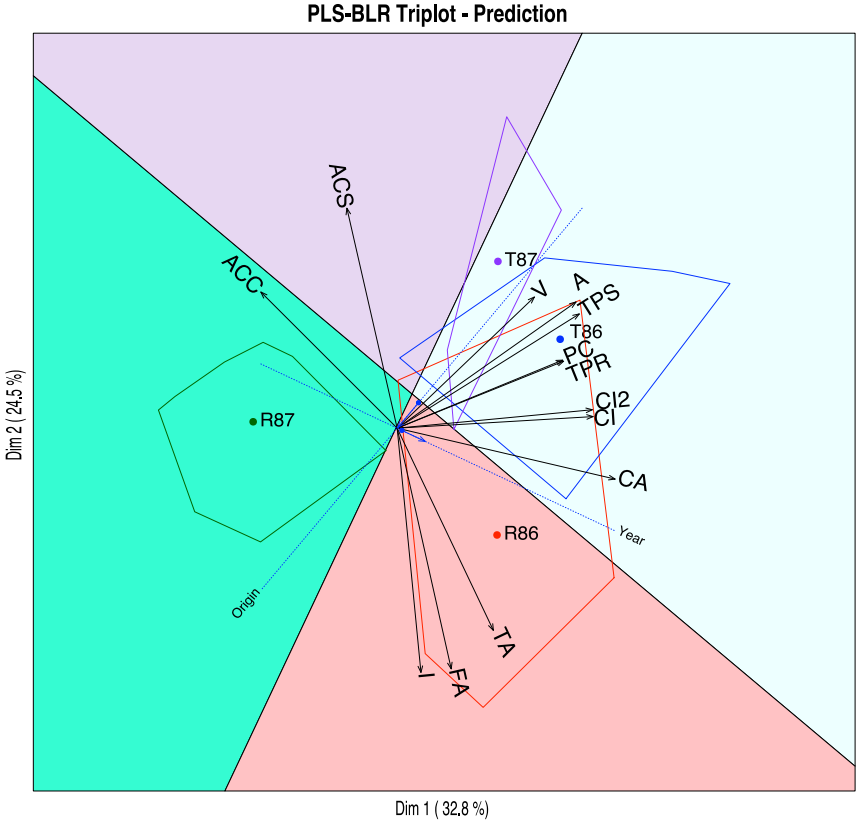

Figura 6.6: Regiones de predicción para la combinación de ambas variables

Conclusiones del ejemplo

- Las variables que mejor predicen las respuestas son la Edad química (CA), el Contenido alcohólico (A), los Fenoles totales - Somers (TPS), la Acidez fija (FA), el Grado de Ionización (I), la Densidad del color (CI y CI2), las Antiocianinas totales - método 2 (ACS), la Acidez total valorable (TA), las Sustancias reactivas a la vainillina (V), los Fenoles totales - Folin (TPR), las Procianididas (PC) y la Malvidina (ACC).

- Las variables más relacionadas con el Origen de los vinos son las Sustancias reactivas a la vainillina (V), el Tono del color del vino (H), el Contenido alcohólico (A), los Fenoles totales - Somers (TPS), las Procianididas (PC) y los Fenoles totales - Folin (TPR).

- Las variables que tienen una mayor relación con el Año de los vinos son la Malvidina (ACC) y la Edad química (CA).

- Las medidas del color CI y CI2 y las Antiocianinas totales - método 2 (ACS) se relacionan tanto con el Origen como con el Año de los vinos.

- Los vinos de la Denominación de Origen Toro se caracterizan por un color más oscuro, mayor graduación alcohólica y poseen más fenoles y procianidinas en comparación a los vinos de la Denominación de Origen Ribera de Duero.

- El color de un vino, la edad química y las malvidinas puede variar con el año. De la misma forma los antocianos y la acidez cambiarán tanto por el origen como por el año del vino.

- Se puede afirmar por el porcentaje de correctamente clasificados y las pseudo R^2 utilizadas la predicción realizada es correcta y precisa. Utilizando ambas variables predictoras se clasifican correctamente el 90 % de los vinos, utilizando solo su DO un 93,33 % y un 86,67 % solo con el año.

6.6.2. Arañas

Para este segundo ejemplo emplearemos la base de datos publicada por Aart y Smeenk-Enserink (1974) que hemos empleado anteriormente en la sección 5.6.1. Como ya habíamos mencionado en ese momento, este ejemplo ha sido utilizado por diversos autores para ilustrar diferentes técnicas ordinales, por ejemplo Braak (1986) en su artículo original sobre el Análisis de Correspondencias Canónico.

El ejemplo empleará la abundancia de las 12 especies diferentes de arañas lobo que se encuentran en la tabla 6.5.

			
Arctosa lutetiana	*Trochosa terricola*	*Pardosa lugubris*	*Alopecosa cuneata*
			
Zora spinimana	*Pardosa monticola*	*Pardosa nigriceps*	*Alopecosa accentuata*
			
Pardosa pullata	*Alopecosa fabrilis*	*Aulonia albimana*	*Arctosa perita*

Tabla 6.5: Especies de arañas utilizadas en el ejemplo

Base de datos

La matriz original contenía las abundancias de 12 especies de arañas lobo recogidas en 100 sitos (empleando trampas) en un área de dunas en los Países Bajos. Disponemos

de las abundancias en 28 lugares en los que también se disponía de un conjunto de variables ambientales.

Los nombres utilizados en el ejemplo se pueden encontrar en la tabla 5.1. Las variables ambientales son Contenido de agua (Watcont), Arena desnuda (Barsand), Cobertura de musgo (Covmoss), Reflexión de la luz (Ligrefl), Ramitas caídas (Falltwi) y Hierbas de cobertura (Coverher).

Los datos iniciales se han convertido al formato binario. Nuestras respuestas serán presencias o ausencias de las especies de araña. Se pueden observar los datos en la tabla 6.6.

Lugar	Arctl	Prdl	Zrsp	Prdn	Prdp	Allb	Trct	Alpcn	Prdm	Alpcc	Alpf	Arctp
s01	0	0	1	1	1	1	1	1	1	1	0	0
s02	0	1	1	1	1	1	1	1	1	0	0	0
s03	1	1	1	1	1	1	1	1	1	1	1	0
s04	1	1	1	1	1	1	1	1	1	1	0	0
s05	1	1	1	1	1	1	1	1	1	1	0	0
s06	1	0	1	1	1	1	1	1	1	0	0	0
s07	1	1	1	1	1	1	1	1	1	1	0	0
s08	0	1	1	1	1	1	1	1	1	0	0	0
s09	0	0	0	1	1	0	1	1	1	1	0	0
s10	0	0	0	0	0	0	1	0	1	1	1	0
s11	0	0	0	0	1	1	1	1	1	1	1	0
s12	0	0	0	1	1	0	1	1	1	1	0	0
s13	1	1	1	1	1	1	1	1	1	1	1	0
s14	1	1	1	1	1	1	1	1	1	0	0	0
s15	0	1	1	0	0	0	1	0	0	0	0	0
s16	0	1	1	1	0	0	1	1	0	0	0	0
s17	0	1	1	0	0	0	1	0	0	0	0	0
s18	0	1	0	0	0	0	1	1	0	0	0	0
s19	0	1	1	1	0	0	1	1	0	0	0	0
s20	0	1	1	0	0	0	1	1	0	0	0	0
s21	0	1	1	0	1	0	1	1	1	0	0	0
s22	0	0	0	0	0	0	1	0	1	1	1	1
s23	0	1	0	0	0	0	1	0	1	1	1	1
s24	0	0	0	0	0	0	0	0	1	1	1	1
s25	0	1	1	1	0	1	1	1	1	1	1	0
s26	0	0	0	0	0	0	1	0	0	1	1	1
s27	0	0	0	0	0	0	0	0	1	1	1	1
s28	0	0	0	0	0	0	1	0	1	1	1	1

Tabla 6.6: Datos Arañas: Especies de Araña Lobo

Nuestros predictores serán las seis variables ambientales medidas en los mismos puntos de muestreo que deben explicar la presencia o ausencia de estas especies. Se pueden observar en la tabla 6.7.

Lugar	Watcont	Barsand	Covmoss	Ligrefl	Falltwi	Coverher
s01	5	0	7	8	0	9
s02	8	0	2	3	3	9
s03	6	0	5	8	0	9
s04	6	0	5	6	0	9
s05	8	0	0	5	0	9
s06	9	5	5	1	7	6
s07	8	0	1	5	0	9
s08	6	0	2	1	9	6
s09	5	0	9	7	0	6
s10	4	8	7	8	0	5
s11	4	0	9	8	0	7
s12	5	0	8	8	0	8
s13	9	3	1	7	3	9
s14	8	0	4	2	0	9
s15	9	0	1	1	9	5
s16	8	0	1	0	9	0
s17	9	0	1	2	9	5
s18	8	0	0	2	9	5
s19	7	0	3	0	9	2
s20	8	0	1	0	9	0
s21	7	0	1	0	9	2
s22	1	7	9	8	0	0
s23	0	6	9	9	0	6
s24	2	7	9	9	0	5
s25	3	7	2	5	0	8
s26	0	9	4	9	0	2
s27	0	5	8	8	0	6
s28	0	7	8	8	0	6

Tabla 6.7: Datos Arañas: Variables ambientales

Objetivos del ejemplo

Para el ejemplo del PLS para datos de respuestas binarias de las especies de las arañas los objetivos planteados serán los siguientes:

Objetivo 1. Determinar cuales son las variables ambientales que mejor predicen la presencia o ausencia de las especies de arañas.

Objetivo 2. Establecer los grupos de lugares de muestreo creados empleando el PLS para datos de respuesta binaria.

Objetivo 3. Identificar las variables ambientales que mejor caracterizan a cada especie de araña lobo

Objetivo 4. Presentar la calidad de las predicciones realizadas para cada especie y de forma conjunta.

Metodología

Utilizaremos las técnicas presentadas en este capítulo. Las componentes que obtenemos para los predictores son las combinaciones lineales de las variables que mejor predicen la presencia y la ausencia de la especie.

Ambas, las especies y las variables ambientales, serán representados en el mismo biplot de forma conjunta.

Emplearemos el software estadístico R (R Core Team, 2021), concretamente el paquete MultBiplotR (Vicente-Villardon, 2021), para realizar tanto los cálculos como las representaciones gráficas que se muestran en este ejemplo.

Resultados

Comenzaremos calculando la varianza explicada por las variables ambientales. La tabla 6.8 contiene el recuento de la variabilidad de los datos ambientales explicados por

las componentes PLS calculadas. Las dos primeras componentes recogen el 84,53 % de la varianza.

	Valores Propios	Var Exp.	Acumulada
Comp. 1	95.44	58.91	58.91
Comp. 2	41.49	25.61	84.53

Tabla 6.8: Varianza Explicada

En este análisis, todas las variables ambientales están bien representadas. La calidad de representación de cada una de ellas se puede observar en la tabla 6.9.

	Comp. 1	Comp. 2
Watcont	90.44	92.52
Barsand	70.95	71.87
Covmoss	66.20	66.22
Ligrefl	79.35	97.13
Falltwi	43.65	93.61
Coverher	0.82	87.90

Tabla 6.9: Calidad de las columnas

Igual que ocurría en los ejemplos anteriores es necesario analizar las medidas de ajuste del modelo construido. La información para los ajustes de las respuestas se encuentran en la tabla 6.10.

	Deviance	g.l.	P-val	Nagel.	Cox-Sn.	MacF.	% Correct	Sensib.	Espec.
Alopacce	11.89	2	0.00	0.51	0.35	0.38	85.71	85.71	85.71
Alopcune	12.47	2	0.00	0.49	0.36	0.33	85.71	82.35	90.91
Alopfabr	17.65	2	0.00	0.63	0.47	0.47	89.29	88.24	90.91
Arctlute	14.46	2	0.00	0.54	0.40	0.37	71.43	73.33	69.23
Arctperi	20.06	2	0.00	0.68	0.51	0.52	85.71	78.57	92.86
Auloalbi	18.73	2	0.00	0.65	0.49	0.49	85.71	83.33	87.50
Pardlugu	0.54	2	0.77	0.05	0.02	0.04	85.71	84.62	100.00
Pardmont	13.71	2	0.00	0.54	0.39	0.39	89.29	94.74	77.78
Pardnigr	8.12	2	0.02	0.37	0.25	0.26	85.71	85.71	85.71
Pardpull	32.50	2	0.00	0.93	0.69	0.87	96.43	100.00	90.91
Trocterr	18.01	2	0.00	0.64	0.47	0.48	85.71	90.91	82.35
Zoraspin	16.05	2	0.00	0.68	0.44	0.55	89.29	100.00	86.36
Total	184.18	24	0.00	0.60	0.42	0.45	86.31	86.81	85.71

Tabla 6.10: Medidas de ajuste para los datos de las arañas

La mayoría de las especies tienen un buen porcentaje de correctamente clasificados. Los coeficientes pseudo R^2 son aceptables excepto en el caso de la especie *Pardosa lugubris*. Esto puede deberse al hecho de que está presente en la mayoría de los lugares, y por lo tanto, no tiene poder discriminante.

En la figura 6.7, tenemos una representación biplot para los datos de las arañas. Se muestran tres conjuntos de marcadores: para las respuestas (especies de arañas), para los predictores (variables ambientales) y para los individuos (sitios de muestreo).

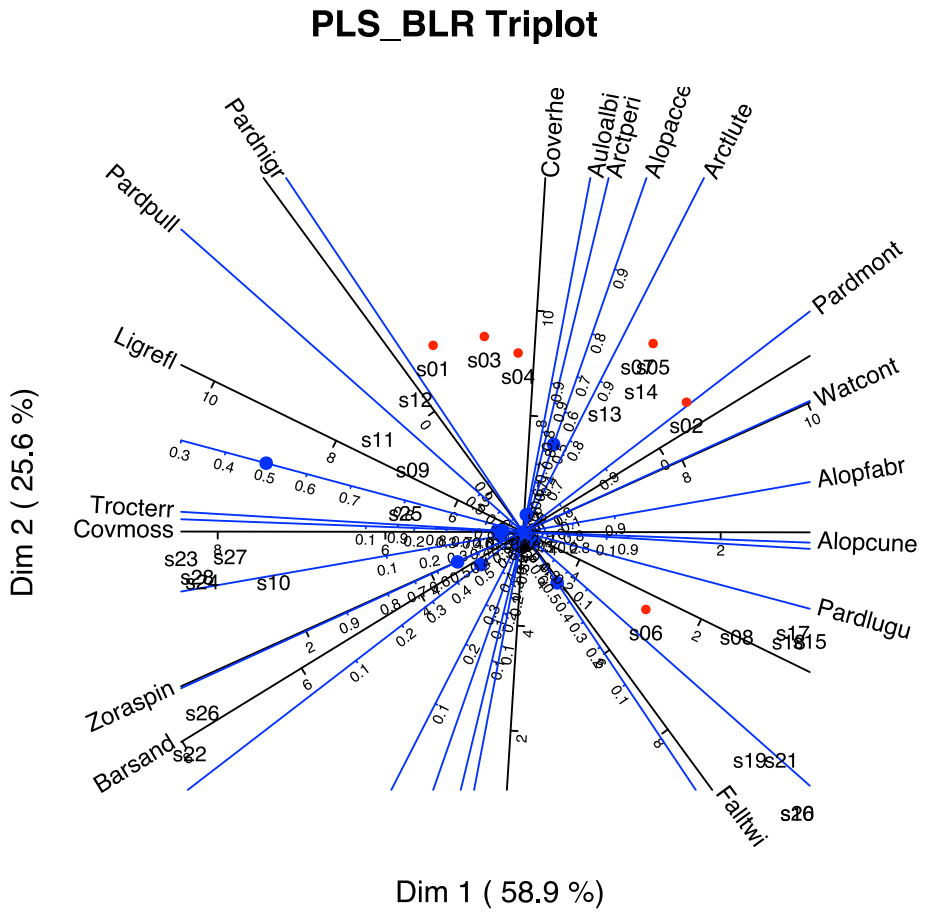

Figura 6.7: Triplot PLS-BLR para los datos de las arañas con escalas para las variables

En el gráfico tenemos una imagen completa del problema. La interpretación es similar a la del caso anterior. Por ejemplo, los valores más altos de *Covermoss* están asociados a una mayor presencia de *Trocterr* y a una menor presencia de *Alopcune*. Una inspección más detenida del biplot permite establecer las relaciones entre los predictores y las respuestas, así como agrupaciones de sitios y sus características principales.

El gráfico también se puede simplificar para facilitar la lectura si no queremos utilizar las escalas de las variables. Hemos ampliado las direcciones de las flechas para colocar las etiquetas fuera. Es posible observarlo en el gráfico de la figura 6.8.

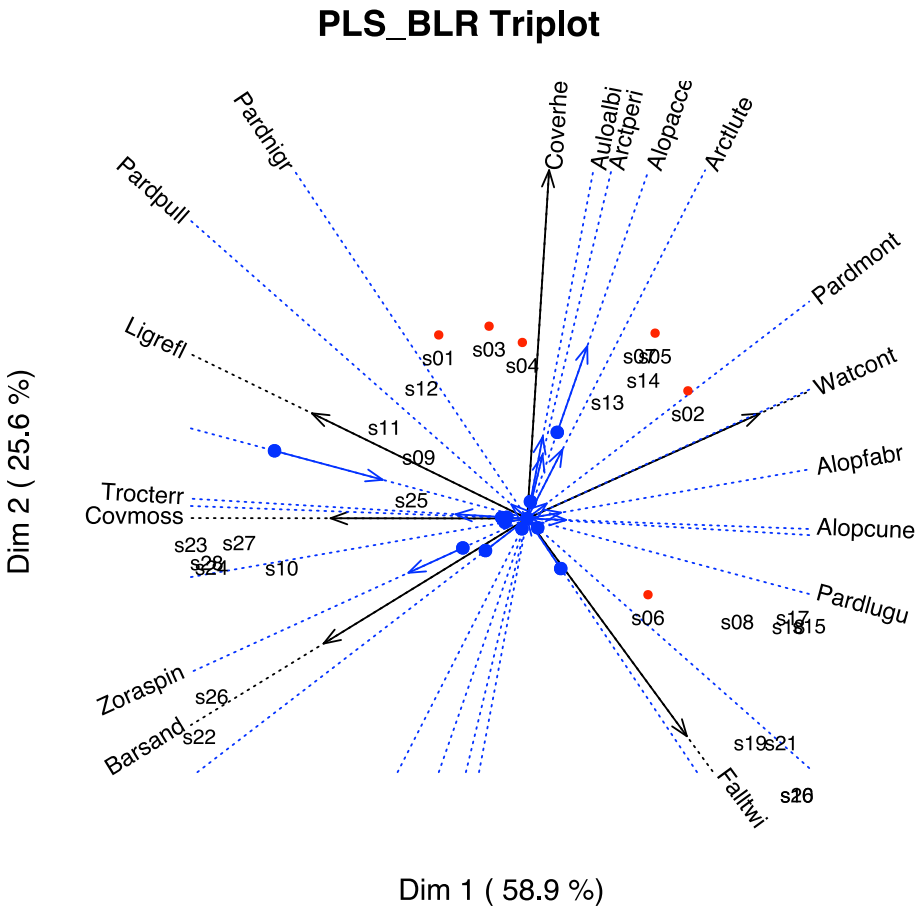

Figura 6.8: Triplot PLS-BLR para los datos de las arañas con flechas

Finalmente, mostraremos las regiones de predicción para cada respuesta de forma separada (Figura 6.9). En el gráfico, los puntos con presencias observadas se han representado en azul, la estrella une cada punto con el centroide de las presencias. Las ausencias se han coloreado en rojo. Como antes, la línea de puntos es la dirección que mejor predice la probabilidad esperada y la flecha muestra la dirección de las probabilidades crecientes. La línea perpendicular es la separación entre las regiones que predicen la presencia y la ausencia, siendo el lado de la flecha la predicción de la presencia.

Podemos ver que la mayoría de los valores observados se encuentran en las regiones de predicción correctas. Esto significa que la técnica propuesta capta correctamente la estructura de los datos y las relaciones entre las especies y las variables ambientales.

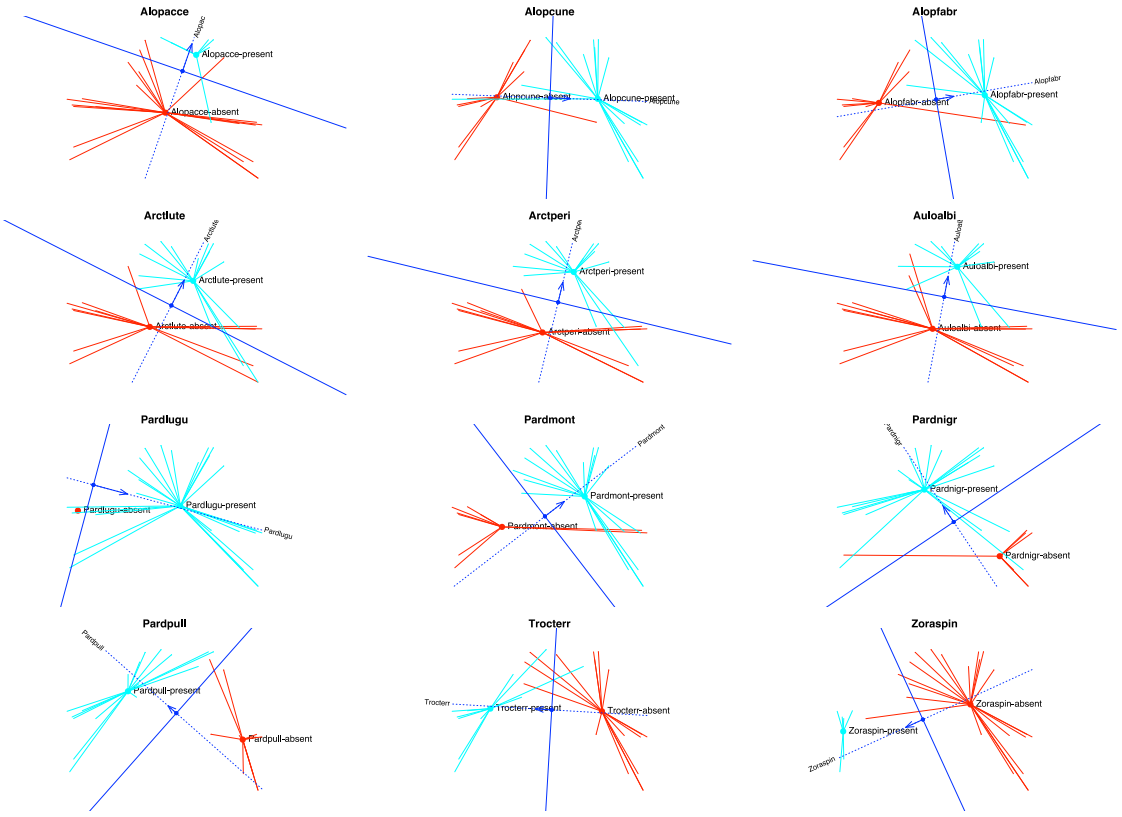

Figura 6.9: Regiones de predicción para cada especie

Conclusiones del ejemplo

- Las variables ordenadas en función de su poder de predicción son la reflexión de la luz, las ramitas caídas, el contenido de agua, la cobertura de hierbas, la arena descubierta y la cobertura de musgo.
- Se han identificado los 7 grupos de lugares de muestreo que se habían presentado con el RDA para datos de respuesta binaria. El grupo 1 está formado por las zonas s10, s23, s24, s27 y s28, el grupo 2 por s09, s11 y s25, el grupo 3 por s01, s03, s04 y s12, el grupo 4 por s02, s05, s07, s13 y s14, el grupo 5 por s06, s08, s15, s17 y s18, el grupo 6 por s16, s19, s20 y s21 y el grupo 7 por s22 y s26.
- Se observa que el contenido de agua de la zona de muestreo está relacionada directamente con las especies *Pardosa monticola*, *Alopecosa fabrilis*, *Alopecosa cuneata* y *Pardosa lugubris*. La arena desnuda está íntimamente relacionada con la especie *Zora spinimana*. La cubierta de musgo presenta una relación muy estrecha con la

Trochosa terricola. La reflexión de la luz muestra una relación importante con las especies *Pardosa pullata* y *Pardosa nigriceps*. Las ramitas caídas mantienen una relación muy estrecha, pero inversa con *Pardosa pullata* y *Pardosa nigriceps*. Y por último, la cobertura de hierba está muy relacionada con las especies *Aulonia albimana*, *Arctosa perita*, *Alopecosa accentuata* y *Arctosa lutetiana*.

- Todas las medidas de calidad presentan que los resultados son precisos. La especie que tiene un mayor porcentaje de correctamente clasificados es *Pardosa pullata* con un 96,43 % y la especie que menor porcentaje de correctamente clasificados posee es *Arctosa lutetiana* con un 71,43 %. A nivel general se observa que el porcentaje de clasificados correctamente es del 86,31 %. La única especie que tiene un p-valor no significativo es *Pardosa lugubris*, sin embargo posee un buen porcentaje de correctamente clasificados (85,71 %), una sensibilidad de 84,62 % y una especificidad 100 %.

Capítulo 7

STATIS Dual para varias matrices de datos binarios: STATIS tetracórico

Notación

K : Número de ocasiones del estudio.

I_k : Número de individuos de la matriz de estudio k.

J : Número de variables en todos los estudios.

$\mathbf{X}_{k_{(I_k \times J)}}$: Matriz de variables explicativas con I_k individuos y J variables.

$\mathbf{U}_{k_{(I \times R)}}$: Matriz de vectores propios de $\mathbf{X}_k\mathbf{X}_k^T$.

$\mathbf{V}_{k_{(J \times R)}}$: Matriz de vectores propios de $\mathbf{X}_k^T\mathbf{X}_k$.

$\mathbf{\Lambda}_k$: Valores propios no nulos de $\mathbf{X}_k\mathbf{X}_k^T$ y $\mathbf{X}_k^T\mathbf{X}_k$.

$\mathbf{C}_k$: Matriz de correlaciones de la ocasión k.

$\mathbf{Y}_k$: Coordenadas de los individuos sobre las Componentes Principales en el estudio k.

$\mathbf{P}, \bar{\mathbf{P}}$: Matriz de saturaciones.

$\mathbf{S}$: Matriz de covarianzas entre estudios.

$\mathbf{R}$: Matriz de correlaciones RV.

$\mathbf{E}$: Matriz de coordenadas de la interestructura.

$\bar{\mathbf{C}}$: Matriz compromiso.

$\mathbf{B}, \mathbf{B}_k$: Parámetros de las variables en los modelos de respuesta

$\mathbf{A}, \mathbf{A}_k$: Parámetros de las variables en los modelos de respuesta

$\hat{\mathbf{Y}}_k$: Representación de los individuos en el compromiso.

π_{ij} : Probabilidad esperada del individuos i en la variable j.

L : Función de pérdida.

7.1. Introducción

El propósito general de los métodos STATIS-ACT (Lavit *et al.*, 1994; des Plantes, 1976) es extraer información común a un conjunto de tablas de datos con los mismos individuos y representarlas como una configuración euclídea o mapa de puntos (o vectores) de la misma manera que en Análisis de Componentes Principales (ACP) o Análisis de Coordenadas Principales (ACoP). Si el objeto es analizar las variables y las estructuras de correlación entre ellas usaremos un Análisis Factorial (AF). Cuando disponemos de tablas en las que medimos un conjunto de variables comunes y queremos obtener una estructura consenso de todas ellas, usaremos el denominado STATIS-Dual.

El método fue diseñado inicialmente para trabajar con individuos comunes a todas la tablas, pero aquí nos centraremos en la versión dual, que trabaja con variables comunes a todas ellas. Una descripción completa para matrices de datos continuos podemos encontrarla en Abdi *et al.* (2012).

Cuando disponemos de varias tablas de datos binarios, los métodos clásicos para datos continuos no son adecuados. Si los individuos son los mismos en todas las tablas, podemos usar un STATIS basado en distancias, también conocido como DISTATIS (Abdi *et al.*, 2005). El procedimiento consiste en calcular una matriz de distancias a partir de para un coeficiente de similaridad para datos binarios. Las distancias se convierten en productos escalares, como en ACoP, y se trabaja a partir de ellos como en el STATIS tradicional.

Cuando disponemos de variables comunes, y estamos interesados en la asociación

entre las mismas, podríamos usar un coeficiente que, en lugar de la similaridad, muestre la asociación entre las variables.

En este trabajo proponemos la utilización de la matriz de correlaciones tetracóricas para cada tabla y desarrollamos las adaptaciones necesarias para el método.

En la sección 7.2 desarrollamos los conceptos fundamentales cuando tenemos datos continuos. En la sección 7.3 extendemos los conceptos para trabajar con datos binarios.

En la sección 7.4 realizaremos un resumen de algunos de los software que pueden ser utilizados para realizar este tipo de técnicas. Por último, en la sección 7.5 realizamos una aplicación a datos reales de los conceptos desarrollados.

7.2. STATIS Dual para una matriz de datos continuos

Disponemos de varias matrices de datos $\mathbf{X}_k (k = 1, \dots, K)$ correspondientes a K diferentes ocasiones o estudios, cada una de ellas con I_k filas y J columnas, es decir, todos los estudios comparten las mismas variables.

7.2.1. Análisis de cada ocasión

Omitiremos el índice k por simplicidad. Cada ocasión contiene la medida de un conjunto de variables $(X_1, \dots X_J)$ sobre I individuos y puede ser analizada separadamente con un ACP, un ACoP o incluso con un biplot obtenido de la DVS de la matriz de datos.

$$\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T, \tag{7.1}$$

donde $\mathbf{U}$ contiene los vectores propios de $\mathbf{X}\mathbf{X}^T$, $\mathbf{V}$ los vectores propios de $\mathbf{X}^T\mathbf{X}$ y $\mathbf{\Lambda} = diag(\lambda_1, \lambda_2, \dots)$ las raíces cuadradas de los valores propios no nulos de ambas matrices, que son iguales, ordenados en magnitud decreciente.

Recuérdese que $\mathbf{X}\mathbf{X}^T$ contiene los productos escalares (euclídeos) que usamos para obtener las Coordenadas Principales (que coinciden con las coordenadas sobre las Componentes Principales). Por otra parte $\mathbf{X}^T\mathbf{X}$ contiene las covarianzas utilizadas para el cálculo de las Componentes Principales. Si los datos están centrados y estandarizados, contienen las correlaciones entre las variables salvo un factor de escala relacionado con el tamaño muestral. Sin pérdida de generalidad, utilizaremos la matriz de correlaciones directamente

$$\mathbf{C} = \frac{1}{I}\mathbf{X}^T\mathbf{X}. \tag{7.2}$$

Para conseguir ésto basta con sustituir la matriz de datos original por $\frac{1}{\sqrt{I}}\mathbf{X}$, es decir,

$$\mathbf{X} \equiv \frac{1}{\sqrt{I}}\mathbf{X}. \tag{7.3}$$

Utilizamos esta matriz por conveniencia, para que los análisis posteriores se basen directamente en la matriz de correlaciones. Las componentes son las mismas ya que los vectores propios de $\mathbf{X}^T\mathbf{X}$ y de $\mathbf{C}$ son los mismos. La diferencia en el factor de escala se traslada a los valores propios.

Resumiendo los resultados:

- $\mathbf{V}$ define las Componentes Principales.
- Las coordenadas de los individuos sobre las Componentes Principales son

$$\mathbf{Y} = \mathbf{X}\mathbf{V} = \mathbf{U}\mathbf{\Lambda}. \tag{7.4}$$

- La DVS está directamente relacionada con la descomposición en valores y vectores propios de $\mathbf{C}$, ya que,

$$\mathbf{C} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T. \tag{7.5}$$

- De esta forma

$$\mathbf{P} = \mathbf{V}\mathbf{\Lambda}, \tag{7.6}$$

contiene las correlaciones entre las componentes y las variables observadas, conocidas como *saturaciones* o *factores de carga* en el contexto del modelo factorial lineal.

$$X_j = p_{j1}F_1 + p_{j2}F_2 + \dots a_{jS}F_S(j = 1, \dots, J), \tag{7.7}$$

siendo $(F_1, \dots, F_S)$ los factores comunes. Para cada individuo

$$z_{ij} = p_{j1}f_{i1} + p_{j2}f_{i2} + \dots p_{jS}f_{iS}, (i = 1, \dots, I; j = 1, \dots, J), \tag{7.8}$$

o, en forma matricial,

$$\mathbf{X} = \mathbf{F}\mathbf{P}^T, \tag{7.9}$$

donde $\mathbf{P} = (p_{js})$ es la matriz factorial que contiene los coeficientes de las J variables en los S factores y $\mathbf{F} = (p_{is})$ contiene las puntuaciones factoriales de cada individuo.

- Las puntuaciones factoriales pueden tomarse como

$$\mathbf{F} = \mathbf{U}. \tag{7.10}$$

- Es posible construir representaciones biplot del tipo JK-Biplot o RMP-Biplot (Gabriel, 1971; Gabriel y Odoroff, 1990) combinando $\mathbf{Y}$ como marcadores para las filas y $\mathbf{V}$ como marcadores para las columnas.

- Y es posible construir representaciones biplot del tipo GH-Biplot o CMP-Biplot (Gabriel, 1971; Gabriel y Odoroff, 1990) combinando $\mathbf{F} = \mathbf{U}$ como marcadores para las filas y $\mathbf{P}$ como marcadores para las columnas.

7.2.2. STATIS Dual

Cada una de las K ocasiones o estudios será representada por el objeto

$$\mathbf{C}_k = \frac{1}{I_k}\mathbf{X}_k^T\mathbf{X}_k.$$

Inter-Structura

En primer lugar buscaremos la *Interestructura*, es decir, trataremos de establecer las similitudes y diferencias entre la distintas ocasiones ya que, si son similares entre ellas será mas sencillo buscar la estructura común a todas ellas.

Calculamos el producto escalar entre ocasiones, es decir, la covarianza entre los estudios mediante

$$cov_{kl} = Cov\,(\mathbf{C}_k, \mathbf{C}_l) = (\mathbf{C}_k \mathbf{C}_l)\,. \tag{7.11}$$

Tenemos entonces una matriz $\mathbf{S} = (s_{kl})$ de covarianzas entre los estudios que puede convertirse en correlaciones

$$rv_{kl} = \frac{cov_{kl}}{\sqrt{cov_{kk} cov_{ll}}}.$$

Se suelen denominar coeficientes de correlación *RV* y pueden organizarse en una matriz.

$$\mathbf{R} = (rv_{kl}). \tag{7.12}$$

Una representación euclídea de esta matriz puede obtenerse de su descomposición en valores y vectores propios

$$\mathbf{R} = \mathbf{L}\Delta\mathbf{L}_T, \tag{7.13}$$

tomando

$$\mathbf{E} = \mathbf{L}\Delta^{1/2}, \tag{7.14}$$

como coordenadas en la representación. Cuanto mayor sea la correlación entre los estudios, mayor será el primer valor propio y mejor la posible estructura común. En este caso los coeficientes del primer vector propio serán de similar magnitud y del mismo signo.

El primer vector propio representa el estudio hipotético más correlacionado con todos los estudios simultáneamente. Los cosenos de los ángulos entre los vectores que representan a los estudios, aproximan la correlación entre los mismos.

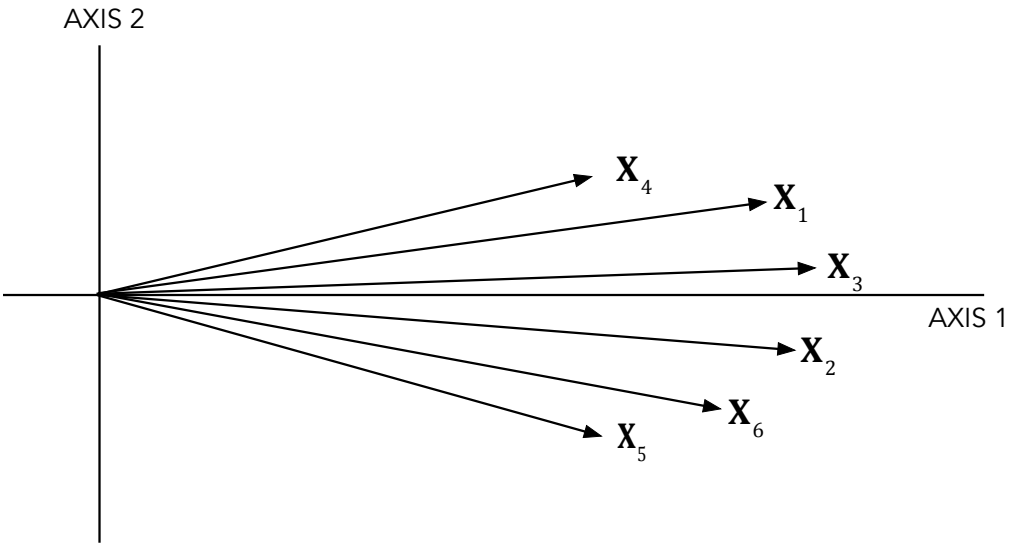

Figura 7.1: Representación típica de la *Interestructura en STATIS*

Compromiso

Buscamos un objeto compromiso que recoja la estructura común y que será una combinación lineal de los objetos individuales

$$\bar{\mathbf{C}} = \sum_{k=1}^{K} \alpha_k \mathbf{C}_k. \tag{7.15}$$

Puede demostrase que las ponderaciones $(\alpha_1, \ldots, \alpha_K)$ de la combinación se obtienen de reescalar los coeficientes del primer vector propio de $\mathbf{R}$. El objeto compromiso es el objeto hipotético más correlacionado simultáneamente con los objetos de las ocasiones individuales. Nótese que el objeto compromiso es del mismo tipo que los objetos de cada ocasión y, por tanto, puede representarse en el mismo espacio que los objetos separados.

Una configuración *compromiso* o *consenso* se obtiene ahora de las descomposición en valores y vectores propios del objeto compromiso

$$\bar{\mathbf{C}} = \bar{\mathbf{V}}\bar{\mathbf{\Lambda}}^2\bar{\mathbf{V}}^T,$$

tomando

$$\bar{\mathbf{P}} = \bar{\mathbf{V}}\bar{\mathbf{\Lambda}}.$$

Obsérvese que la representación del compromiso contiene una matriz de saturaciones similar a la de las ocasiones individuales.

El mapa euclídeo puede completarse con trayectorias para las variables proyectando cada ocasión individual en el consenso. Teniendo en cuenta que

$$\bar{\mathbf{P}} = \bar{\mathbf{V}}\bar{\mathbf{\Lambda}} = \bar{\mathbf{C}}\bar{\mathbf{V}}\bar{\mathbf{\Lambda}}^{-1}.$$

Las coordenadas para el *k-ésimo* estudio se obtienen de las ecuaciones

$$\bar{\mathbf{P}}_k = \mathbf{C}_k\bar{\mathbf{V}}\bar{\mathbf{\Lambda}}^{-1}.$$

Si todos los puntos están cercanos a los del compromiso quiere decir que hay una fuerte estructura común.

En definitiva, hemos buscado una estructura factorial común a todas las ocasiones, aquella que está más cerca de las estructuras de todas las ocasiones individuales. La ventaja de buscarlas de esta forma en lugar de hacerlo con la matriz concatenada de todas las individuales es que, de esta forma, eliminamos las posibles diferencias en la localización de las distintas matrices que pueden conducir a la obtención de una estructura espúrea.

Desde otro punto de vista, $\bar{\mathbf{V}}$ y todos los $\mathbf{V}_k$ para cada ocasión son conjuntos de vectores ortonomarles que definen subespacios del espacio $\mathcal{J}$*-dimensional* en el que podemos

representar todos los objetos. Podemos entonces entender $\bar{\mathbf{V}}$ como un conjunto de componentes comunes que recoge la estructura consenso de todas las ocasiones. Usando las ecuaciones en 7.4 podemos proyectar todas las matrices individuales en este subespacio con

$$\bar{\mathbf{Y}}_k = \mathbf{X}_k \bar{\mathbf{V}}, \tag{7.16}$$

de forma que obtenemos una representación de todos los individuos en un espacio de referencia común. Un razonamiento similar en el contexto de los denominados *meta-biplots* puede encontrarse en Martín-Rodríguez *et al.* (2002). Hay que hacer notar que el biplot representa los datos de cada matriz estandarizados, es decir, la comparación entre ocasiones tiene que hacerse relativa a la media de cada ocasión en unidades de la desviación típica.

7.3. STATIS Dual para una matriz de datos binarios: Tetra-STATIS dual

Disponemos ahora de matrices $\mathbf{X}_k (k = 1, \ldots, K)$ de datos binarios en lugar de datos continuos y los métodos que hemos descrito en la sección anterior no son adecuados ya que al tratarse de métodos lineales, como se ha afirmado en capítulos anteriores, no son óptimos para datos binarios de la misma manera que, por ejemplo, la Regresión Lineal no es adecuada cuando la respuesta es binaria. Adaptaremos los métodos para que se pueda trabajar con respuestas binarias utilizando funciones logísticas, de forma análoga a las anteriores Regresiones Lineales que se han generalizado mediante Regresiones Logísticas.

el esquema para la realización de este análisis será muy similar al explicado anteriormente, como se puede observar en la figura 7.2.

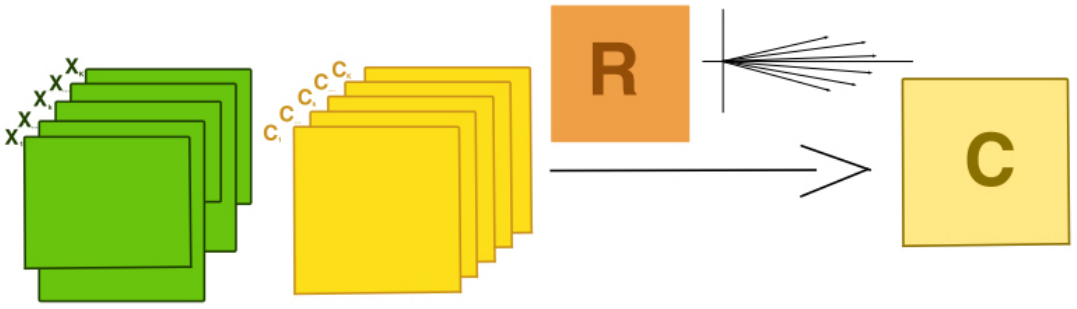

Figura 7.2: Esquema de la realización del STATIS Dual

7.3.1. Análisis de cada ocasión

El caso separado para cada ocasión es un poco más complejo que el caso continuo, ya que no es posible calcular directamente la DVS y las relaciones entre filas y columnas no están tan claras. La relación entre una generalización de la DVS con la descomposición en valores y vectores propios de la matriz de correlaciones no es tan clara ya que ambas proceden de modelos diferentes.

Usaremos dos aproximaciones al análisis individual de cada ocasión, la primera es la función de respuesta (teoría de la respuesta al ítem (TRI) o rasgo latente) y la segunda es la de una variable subyacente. En la primera aproximación obtendremos un modelo que generaliza la DVS mientras que, en la segunda, tendremos la descomposición de una matriz de correlaciones. Una descripción similar para variables ordinales podemos encontrarla en Jöreskog y Moustaki (2001).

Modelos de función de respuesta

Expresa la probabilidad de respuesta como función de los factores latentes y supone que las respuestas a diferentes variables son independientes dados los factores latentes.

Sea $\mathbf{X}_k = (x_{ij}^k)$ una matriz $I_k \times J$ de datos binarios en la que se han medido J variables (o características) binarias $(X_1, \ldots, X_J)$ en I_k individuos en la *k-ésima* ocasión. Los datos

binarios pueden entenderse también como probabilidades observadas, es decir, si una característica está presente, la probabilidad observada es 1 mientras que, si está ausente, es 0. Para este apartado omitiremos los índices k por simplicidad.

Podemos adaptar las DVS con el siguiente modelo bilineal generalizado que, en lugar de la descomposición en tres partes, considera solamente 2.

Sea $\pi_{ij} = E\left(x_{ij}\right)$ la probabilidad esperada de que el individuo i tenga en la característica (variable) j. Podemos definir un modelo bilineal de la forma

$$\pi_{ij} = \frac{e^{\left(b_{j0}+\sum_s b_{js}a_{is}\right)}}{1 + e^{\left(b_{j0}+\sum_s b_{js}a_{is}\right)}}, \tag{7.17}$$

donde a_{is} y b_{js}, $(i = 1, \ldots, I; j = 1, \ldots, J; s = 1, \ldots, S)$ son los parámetros del mismo. Los primeros (**a**'s) pueden entenderse como las puntuaciones de cada una de los individuos sobre los factores latentes mientras que los segundos (**b**'s) son los coeficientes que nos permiten relacionar las variables observadas y las latentes, de forma análoga a la utilizada en capítulos anteriores.

El modelo expuesto es similar a los modelos de la Teoría de la Respuesta al ítem que se describen, por ejemplo, en Baker y Kim (2004) o en Cai *et al.* (2016) o la versión multidimensional en Bonifay (2019). La diferencia fundamental es que aquí consideramos los modelos de forma general sin las restricciones asociadas a la TRI y normalmente multidimensionales. También es similar al Análisis de Componentes Principales para datos binarios en la forma propuesta por de Leeuw (2006) ó Schein *et al.* (2003).

La ecuación (7.17) es una generalización de la DVS utilizando como enlace la función $logit$

$$\text{logit}(\pi_{ij}) = b_{j0} + \sum_{s=1}^{S} b_{js}a_{ij} = b_{j0} + (\mathbf{a}_i)^T\mathbf{b}_j. \tag{7.18}$$

En forma matricial,

$$\text{logit}(\mathbf{\Pi}) = \mathbf{1}_n(\mathbf{b}_0)^T + \mathbf{A}\mathbf{B}^T. \tag{7.19}$$

Donde $\mathbf{\Pi}$ es la matriz de probabilidades esperadas, $\mathbf{b}_0$ es el vector de constantes y $\mathbf{A}$ y $\mathbf{B}$ las matrices de parámetros.

Excepto por el vector de constantes, tendremos una factorización en escala $logit$ que nos permitirá la construcción de un biplot en el que los marcadores de las filas se encuentran en la matriz $\mathbf{A}$ y los de las columnas en $\mathbf{B}$. De nuevo, es necesario mantener esta constante porque la matriz de datos binarios no se puede centrar como ocurre en el caso continuo. Como tenemos un Modelo Generalizado, la geometría es muy similar al caso lineal. La geometría del modelo para construir biplots puede encontrase, por ejemplo, en Vicente-Villardon *et al.* (2006), Vicente-Gonzalez y Vicente-Villardon (2022), Vicente-Villardon y Vicente-Gonzalez (2021), Vicente-Villardón y Hernández-Sánchez (2020) ó Babativa-Márquez y Vicente-Villardón (2021). En el contexto de la psicometría los biplots comparten características en común con las representaciones gráficas propuestas en Ackerman (1996).

La estimación de los parámetros del modelo puede hacerse por diversos métodos. En un artículo reciente Bergner *et al.* (2022) describen algunos de ellos en relación con la TRI. Describimos brevemente una selección de ellos que no pretende ser exhaustiva.

Maxima Verosimilitud Conjunta (Joint Maximum Likelihood): Es probablemente el más antiguo de los métodos. Trata todos los parámetros, tanto para individuos como variables, como desconocidos y los estima simultáneamente mediante optimización. En el contexto de la TRI, los parámetros son inconsistentes cuando el número de individuos tiende a infinito y el de variables (ítemes) se mantiene constante. El método ha sido descrito y modificado recientemente por Chen *et al.* (2019). Es también el método para construcción de un biplot logístico en Vicente-Villardon *et al.* (2006) desde un punto de vista más exploratorio. Además de lo descrito puede presentar también el conocido co-

mo problema de la separación donde el método de máxima verosimilitud no converge si, en el espacio generado por las variables latentes, hay un hiperplano que separa las presencias de las ausencias para alguna de las variables observadas.

Maxima Verosimilitud Marginal (Marginal Maximum Likelihood): Es el preferido en la literatura psicométrica. Evita la inconsistencia suponiendo a los parámetros de los individuos como incidentales e integrando sobre ellos. Usa el algoritmo EM (Expectation-Maximization) y fue propuesto por Bock y Aitkin (1981).

Análisis de Componentes Principales para datos binarios: Por ejemplo, en la forma propuesta por **?**, que utiliza un método iterativo basado en mayorizaciones, o Schein *et al.* (2003), que usa también un algoritmo de mínimos cuadrados alternados.

Método del Gradiente: En el método del gradiente utilizamos la siguiente función de pérdida:

$$L = \sum_{i=1}^{I} \sum_{k=1}^{K} \left[-x_{ij} \log(\pi_{ij}) - (1 - x_{ij}) \log(1 - \pi_{ij}) \right], \tag{7.20}$$

Desarrollaremos el método del gradiente de forma recursiva, es decir, estimaremos una componente en cada paso. Las actualizaciones de cada parámetro mediante el gradiente son son:

$$b_{j0} = b_{j0} - \alpha \frac{\partial L}{\partial b_{j0}} = b_{j0} - \alpha \sum_{i=1}^{I} (\pi_{ij} - x_{ij}) \tag{7.21}$$

$$a_{is} = a_{is} - \alpha \frac{\partial L}{\partial a_{is}} = a_{is} - \alpha \sum_{j=1}^{J} b_{js} (\pi_{ij} - x_{ij}) \tag{7.22}$$

$$b_{js} = b_{js} - \alpha \frac{\partial L}{\partial b_{js}} = b_{js} - \alpha = \sum_{i=1}^{I} a_{is} (\pi_{ij} - x_{ik}) \tag{7.23}$$

para un valor de α.

Podemos organizar los cálculos en un algoritmo conjunto en el que se calculan todos los parámetros simultáneamente o en un algoritmo alternado en el que se calculan inicialmente los b_{j0} y entonces, para cada dimensión, dados los a_{is} calculamos los b_{js} y viceversa. El algoritmo completo se ha descrito en el capítulo anterior y pude consultarse en Vicente-Gonzalez y Vicente-Villardon (2022). Para evitar la elección de α es posible utilizar variaciones del método como el gradiente conjugado, que normalmente están accesibles directamente en paquetes pre-programados de optimización. Un desarrollo completo más reciente de este tipo de métodos puede encontrarse en Babativa-Márquez y Vicente-Villardón (2021).

Biplot Logístico Externo: Cuando las coordenadas de los individuos contenidas en **A** son conocidas, es posible calcular las coordenadas de las variables mediante regresiones logísticas de cada columna de **X** sobre **A**. Los términos independientes de cada regresión son los elementos del vector b_{j0} y los vectores de coeficientes forman las filas de la matriz **A**. Este procedimiento es denominado *biplot logístico externo* por Demey *et al.* (2008). También podría ajustarse por el método del gradiente usando solamente las ecuaciones (7.21) y (7.23).

Modelo de variable subyacente: Factorización de la correlación tetracórica

En adelante supondremos que, para cada una de las variables binarias observadas, hay una variable continua subyacente Z_j, con distribución normal estándar, y que la variable binaria es el resultado de truncar la variable normal con un umbral τ_j de forma que valores menores codifican ausencia (0) y valores mayores, presencia (1).

Suponiendo este modelo subyacente, la correlación entre las variables que generaliza la mostrada en (7.2) se denomina correlación tetracórica y trata de estimar la correlación entre las variables normales subyacentes. Algunos detalles sobe este índice y la forma de estimarlo puede encontrarse por ejemplo en Kirk (1973), Castellan (1966) o Divgi (1979).

Llamaremos ahora **C** a las matrices de correlaciones tetracóricas estimadas y sean τ_j los umbrales estimados para la variable j, que podemos organizar en un vector

$$\boldsymbol{\tau} = (\tau_1, \dots, \tau_J)^T \tag{7.24}$$

que contiene los umbrales estimados de cada variable.

En el caso de datos continuos encontrábamos una clara relación entre la DVS y la descomposición de la matriz de correlaciones. Para el caso de datos binarios, esa relación no es tan clara.

Con el modelo de variables normales subyacentes, podemos factorizar la matriz de correlaciones tetracóricas para obtener la matriz de cargas factoriales. Si denominamos $(Z_1, \dots, Z_J)$ al conjunto de variables continuas subyacentes, con distribución normal estándar, que genera cada una de las binarias, el modelo factorial lineal para S factores comunes, $F_1, \dots, F_s$, se podría escribir como

$$Z_j = p_{j1}F_1 + p_{j2}F_2 + \dots a_{jS}F_S, \tag{7.25}$$

con $j = 1, \dots, J$. Para el í-ésimo individuo tenemos

$$z_{ij} = p_{j1}f_{i1} + p_{j2}f_{i2} + \dots p_{jS}f_{iS}, \tag{7.26}$$

$\mathbf{P} = (p_{js})$ es la matriz factorial que contiene los coeficientes de las J variables en los S factores y $\mathbf{F} = (p_{is})$ contiene las puntuaciones factoriales de cada individuo.

La matriz factorial podemos obtenerla de la descomposición en valores y vectores propios de **C**,

$$\mathbf{C} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T, \tag{7.27}$$

tomando

$$\mathbf{P} = \mathbf{V}\boldsymbol{\Lambda}, \tag{7.28}$$

igual que en la ecuación (7.6) para el caso continuo. Allí podemos encontrar una relación directa entre las cargas y las puntuaciones de los individuos. En el caso binario lo

haremos de una forma indirecta.

Siguiendo a Jöreskog y Moustaki (2001), Bartholomew *et al.* (2011) o Takane y De Leeuw (1987), podemos establecer la relación que existe entre los parámetros de la ecuación (7.17) y los parámetros en las ecuaciones (7.24) y (7.25).

Los parámetros del modelo factorial para las variables subyacentes pueden ponerse en función de los del modelo de respuesta de la siguiente forma:

$$\tau_j = b_{j0}\left(1 + \sum_{s=1}^{S} b_{js}^2\right)^{-1/2}, \tag{7.29}$$

$$p_{js} = b_{js}\left(1 + \sum_{s=1}^{S} b_{js}^2\right)^{-1/2}. \tag{7.30}$$

De una forma similar es posible obtener los parámetros del modelo de respuesta en función de los del factorial:

$$b_{j0} = \tau_j\left(1 + \sum_{s=1}^{S} p_{js}^2\right)^{-1/2}, \tag{7.31}$$

$$b_{js} = p_{js}\left(1 + \sum_{s=1}^{S} p_{js}^2\right)^{-1/2}. \tag{7.32}$$

Obsérvese que los parámetros de cada uno de los modelos no son más que un re-escalado de los parámetros del otro. Es decir, en la representación gráfica los vectores tendrían la misma dirección.

Para este modelo podemos también construir un biplot haciendo la conversión de los parámetros en el modelo de respuesta y usando la ecuación (7.22) para ajustar las coordenadas de los individuos.

7.3.2. Tetra-STATIS Dual

Cada una de las K ocasiones o estudios será representada por el objeto $\mathbf{C}_k$, es decir, la matriz de correlaciones tetracóricas para la ocasión k. La generación del método STATIS-dual es inmediata sustituyendo la matriz de correlaciones de Pearson por la de correlaciones tetracóricas.

Inter-Structura

Calculamos la matriz de correlaciones vectoriales y la descomponemos en valores y vectores propios para obtener una representación de la Interestructura igual que en el caso continuo

$$\mathbf{R} = \mathbf{L}\Delta\mathbf{L}_T, \tag{7.33}$$

tomando

$$\mathbf{E} = \mathbf{L}\Delta^{1/2}. \tag{7.34}$$

como coordenadas en la representación. La interpretación es exactamente la misma que en el caso anterior.

Compromiso

Buscamos un objeto compromiso que recoja la estructura común y que será una combinación lineal de los objetos individuales

$$\bar{\mathbf{C}} = \sum_{k=1}^{K} \alpha_k \mathbf{C}_k. \tag{7.35}$$

Las ponderaciones $(\alpha_1, \ldots, \alpha_K)$ de la combinación se obtienen de reescalar los coeficientes del primer vector propio de $\mathbf{R}$. Como en el caso continuo, el compromiso es el objeto hipotético más correlacionado simultáneamente con todas las ocasiones indivi-

duales.

Una estructura factorial *compromiso* o *consenso* se obtiene ahora de las descomposición en valores y vectores propios del objeto compromiso

$$\bar{\mathbf{C}} = \bar{\mathbf{V}}\bar{\mathbf{\Lambda}}^2\bar{\mathbf{V}}^T,$$

tomando

$$\bar{\mathbf{P}} = \bar{\mathbf{V}}\bar{\mathbf{\Lambda}}.$$

Obsérvese que la representación del compromiso contiene una matriz de saturaciones similar a la de las ocasiones individuales. Las variables se representarán mediante vectores, la forma usual de hacerlo.

El mapa euclídeo puede completarse con trayectorias para las variables proyectando cada ocasión individual en el consenso. Teniendo en cuenta que

$$\bar{\mathbf{P}} = \bar{\mathbf{V}}\bar{\mathbf{\Lambda}} = \bar{\mathbf{C}}\bar{\mathbf{V}}\bar{\mathbf{\Lambda}}^{-1}.$$

Las coordenadas para el *k-ésimo* estudio se obtienen de las ecuaciones,

$$\bar{\mathbf{P}}_k = \mathbf{C}_k\bar{\mathbf{V}}\bar{\mathbf{\Lambda}}^{-1}.$$

Si todos los puntos están cercanos a los del compromiso quiere decir que hay una fuerte estructura común.

En definitiva, hemos buscado una estructura factorial común a todas las ocasiones en la que eliminamos las posibles diferencias en la localización de las distintas matrices que pueden conducir a la obtención de una estructura espúrea.

Dado que la matriz factorial contiene un conjunto de saturaciones, es posible representarla en un círculo de correlaciones como el que mostramos en la aplicación.

La construcción del biplot ya no es tan sencilla como la que tenemos en la ecuación (7.16) ya que no es posible proyectar las matrices de datos directamente. Podemos usar los parámetros de la solución factorial para convertirlos en parámetros de la función de respuesta y, a partir de ellos, estimar las puntuaciones de los individuos por el método del gradiente con la ecuación (7.22), para ello necesitamos calcular primero el vector de constantes $\bar{\mathbf{b}}_0 = (\bar{b}_{10}, \dots, \bar{b}_{J0})^T$ consenso y la matriz de parámetros $\bar{\mathbf{B}} = (\bar{b}_{js})$ consenso de cada variable en cada dimensión.

Para los primeros deberíamos usar la ecuación (7.31) pero resulta que tenemos un conjunto de umbrales para cada ocasión y obtendríamos un conjunto diferente para cada una de ellas.

Si llamamos τ_{jk} los umbrales estimados para la variable j en la ocasión, que podemos k organizar en un vector

$$\boldsymbol{\tau}_k = (\tau_{1k}, \dots, \tau_{Jk})^T, \tag{7.36}$$

a los umbrales de la ocasión k. Si los ordenamos como columnas de una matriz $\mathbf{T} = (\tau_{jk})$ podemos combinar las columnas de $\mathbf{T}$ para obtener un umbral consenso con las mismas ponderaciones que usamos para combinar las correlaciones, esto es,

$$\bar{\boldsymbol{\tau}} = \mathbf{T}\boldsymbol{\alpha}, \tag{7.37}$$

donde $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_K)^T$ son las ponderaciones usadas en la ecuación 7.35.

A partir de estas, escribimos los parámetros como:

$$\bar{b}_{j0} = \bar{\tau}_j \left(1 + \sum_{s=1}^{S} \bar{p}_{js}^2 \right)^{-1/2}, \tag{7.38}$$

$$\bar{b}_{js} = \bar{p}_{js} \left(1 + \sum_{s=1}^{S} \bar{p}_{js}^2 \right)^{-1/2}. \tag{7.39}$$

Finalmente calculamos las puntuaciones de los individuos de la ocasión k iterando

$$\bar{a}_{is}^k = \bar{a}_{is}^k - \alpha \sum_{j=1}^{J} \bar{b}_{js}(\pi_{ij}^k - x_{ij}^k), \tag{7.40}$$

para algún valor de α. En la ecuación, $\bar{a}_{is}^k$ es la puntuación del individuo i en la dimensión s para la ocasión k, x_{ij}^k es el valor observado del individuo i en la variable j para la ocasión k y π_{ij}^k la correspondiente probabilidad esperada. En este caso pueden calcularse todos los parámetros simultáneamente ya que los de las variables son fijos.

7.4. Software

No son muchos los software disponibles para la realización del STATIS actualmente, sin embargo, reservaremos esta sección para realizar una pequeña revisión de algunos de ellos.

La propuesta realizada en este capítulo solo se encuentra dentro del paquete MultBiplotR (Vicente-Villardon, 2021), donde se han incluido las funciones necesarias para la realización del STATIS Dual para matrices de datos binarios y sus representaciones gráficas.

7.4.1. Paquetes comerciales

En esta sección se recogen los software comerciales con los que se pueden realizar este tipo de análisis.

XLSTAT (Addinsoft, 2022)

Es uno de los pocos paquetes comerciales que actualmente mantiene esta técnica dentro de sus menús. Este software es un complemento de Excel, como se ha descrito anteriormente, que amplia el número de técnicas estadísticas que se pueden realizar con el software de base.

XLSTAT permitirá realizar el STATIS y el STATIS Dual dentro de sus menús, y asociará diferentes representaciones gráficas a estos análisis para facilitar su interpretación. Sin embargo, sus representaciones solo serán fiables si el porcentaje de variabilidad asociada a los ejes de nuestra representación es superior al 80 %.

ACT (STATIS Method) Lavit *et al.* (1994)

No es un software muy conocido y es de uso restringido. Permitía realizar los métodos STATIS a través de una interfaz gráfica.

Actualmente, se ha integrado sus funcionalidades dentro del paquete de R (R Core Team, 2021), en el paquete denominado missRows (Gonzalez y Voillet, 2022) del repositorio BioConductor.

ade4 (Thioulouse *et al.*, 1997)

Igual que en el caso anterior, desde su creación el software ade4 ha permitido realizar estos métodos a través de la interfaz gráfica. El software previo aún se encuentra disponible en su página web (`http://pbil.univ-lyon1.fr/ADE-4-old/ADE-4.html`), sin embargo no se mantendrá mucho más tiempo.

Actualmente, las utilidades del software se han implementado en R (R Core Team,

2021). Es posible utilizarlo a través del paquete ade4 (Dray *et al.*, 2022) o de la interfaz gráfica desarrollado en el paquete ade4TkGUI.

Mathematica (Inc., 2022)

Como alternativa a los paquetes mencionados anteriormente, Elizondo y Varela (1998) proponen el uso de Mathematica para la realización tanto del STATIS como del STATIS Dual.

En el artículo mencionado anteriormente describen cada uno de los pasos para realizar los cálculos y representaciones gráficas de los STATIS.

7.4.2. Paquetes de R R Core Team (2021)

Existen diversos paquetes para la realización del STATIS y el STATIS Dual en R, en esta sección se mostrarán algunos de ellos, excluyendo los mencionados en la sección anterior, ya que serán considerados programas comerciales que se han transformado a paquetes de R.

ClustBlock (Llobell *et al.*, 2022)

Este paquete contiene funciones para realizar algoritmos jerárquicos y de partición de bloques de variables.

En relación con este capítulo, se incluye en el paquete funciones para la realización de STATIS para variables cuantitativas divididas en bloques y en datos Free Sorting, junto a sus correspondientes representaciones gráficas.

multigroup (Eslami *et al.*, 2020)

multigroup es un paquete que contiene únicamente la opción Dual del STATIS para datos continuos. Es un paquete de reciente creación que contiene métodos multivariantes para describir, resumir y visualizar datos con una estructura de grupo.

MultBiplotR (Vicente-Villardon, 2021)

Las funciones para realizar los cálculos del STATIS Dual para datos binarios presentados en este capítulo y sus representaciones gráficas han sido implementadas e incluidas en el paquete MultBiplotR (Vicente-Villardon, 2021), como ocurría en los capítulos anteriores.

7.5. Ejemplo de STATIS Dual para varias matrices de datos binarios

Se ilustrará en esta sección el último método propuesto en este documento con la encuesta del CIS ya mencionada en ejemplos anteriores. Como ya se había descrito en el capítulo 3, se utilizará la encuesta que realizó el CIS titulada "Efectos y consecuencias del Coronavirus".

Todos los cálculos se han realizado con las funciones implementadas en R y que han sido incluidas en el paquete MultBiplotR (Vicente-Villardon, 2021).

7.5.1. Estudio de la evolución de la opinión de los residentes en España sobre los efectos y las consecuencias del coronavirus

Uno de los temas de interés para la sociedad actual es, si se ha visto modificado el comportamiento debido a la pandemia vivida desde principios de 2020. Por ello el Centro de Investigaciones Sociológicas (CIS) decidió realizar una encuesta sobre los efectos y las consecuencias del Coronavirus.

La encuesta utilizada se puede encontrar detallada en el catálogo de servicios de la web del CIS (`https://www.cis.es/cis/opencms/ES/2_bancodatos/catalogoencuestas.html`).

En total, adscritas a este nombre se han realizado seis encuestas, existen preguntas comunes a todas ellas y algunas que se han ido adaptando a lo largo de los meses para adecuarse mejor al momento en el que se realzaban. Las encuestas realizadas por el CIS son las siguientes:

Efectos y consecuencias del Coronavirus (I): (Nº 3298) Se realizó entre el 23 y el 31 de octubre de 2020.

Efectos y consecuencias del Coronavirus (II): (Nº 3302) Se realizó entre el 23 y el 26 de noviembre de 2020.

Efectos y consecuencias del Coronavirus (III): (Nº 3305) Se realizó entre el 11 y el 16 de diciembre de 2020.

Efectos y consecuencias del Coronavirus (IV): (Nº 3324) Se realizó entre el 14 y el 19 de mayo de 2021.

Efectos y consecuencias del Coronavirus (V): (Nº 3336) Se realizó entre el 11 y el 30 de septiembre de 2021.

Efectos y consecuencias del Coronavirus (VI): (Nº 3302) Se realizó entre el 14 y el 17 de diciembre de 2021.

En el ejemplo de la sección 3.5, se empleó la encuesta IV para estudiar las preguntas 20 y 21 que no se recogen en las dos últimas encuestas.

Base de datos

En este ejemplo emplearemos las encuestas "Efectos y consecuencias del Coronavirus (I), (V) y (VI)". Estas tres encuestas se han realizado a diferentes individuos, pero comparten dos conjuntos de preguntas que emplearemos en este análisis.

La encuesta "Efectos y consecuencias del Coronavirus (I)" fue realizada a un total de 2861 individuos, sin embargo debido a la presencia de datos perdidos y otros sesgos, se han seleccionado 705 observaciones.

La encuesta "Efectos y consecuencias del Coronavirus (V)" fue realizada a un total de 3097 individuos, sin embargo debido a la presencia de datos perdidos y otros sesgos, se han seleccionado 685 observaciones.

La encuesta "Efectos y consecuencias del Coronavirus (VI)" fue realizada a un total de 2462 individuos, sin embargo debido a la presencia de datos perdidos y otros sesgos, se han seleccionado 572 observaciones.

En todos los casos se han cogido las mismas 18 variables que se detallan a continuación:

TemorEnf La pandemia del covid-19 ha dado lugar a diversas situaciones que pueden afectar a la salud de las personas, ¿podría decirme si desde que se declaró la pandemia del coronavirus ha sentido usted temor a enfermar?

DolorPerdida La pandemia del covid-19 ha dado lugar a diversas situaciones que pue-

den afectar a la salud de las personas, ¿podría decirme si desde que se declaró la pandemia del coronavirus ha sentido usted dolor por la pérdida de algún/a familiar, amigo/a o conocido/a?

PerEmpleo La pandemia del covid-19 ha dado lugar a diversas situaciones que pueden afectar a la salud de las personas, ¿podría decirme si desde que se declaró la pandemia del coronavirus ha sentido usted preocupación por haber perdido su empleo personal o el de algún/a familiar?

MedContacto La pandemia del covid-19 ha dado lugar a diversas situaciones que pueden afectar a la salud de las personas, ¿podría decirme si desde que se declaró la pandemia del coronavirus ha sentido usted inquietud por las medidas que puede limitar los contactos y relaciones cara a cara con sus familiares, amigos/as y vecinos/as?

PosPerEmpleo La pandemia del covid-19 ha dado lugar a diversas situaciones que pueden afectar a la salud de las personas, ¿podría decirme si desde que se declaró la pandemia del coronavirus ha sentido usted miedo por la posibilidad de poder perder su empleo personal o el de algún/a familiar?

AfrontarGastos La pandemia del covid-19 ha dado lugar a diversas situaciones que pueden afectar a la salud de las personas, ¿podría decirme si desde que se declaró la pandemia del coronavirus ha sentido usted intranquilidad por no poder afrontar sus gastos (hipotecas, alquileres, préstamos, suministros, telefonía, etc.)?

VidaAnterior La pandemia del covid-19 ha dado lugar a diversas situaciones que pueden afectar a la salud de las personas, ¿podría decirme si desde que se declaró la pandemia del coronavirus ha sentido usted miedo por no recuperar su vida tal como era antes de la pandemia?

EmprProyect La pandemia del covid-19 ha dado lugar a diversas situaciones que pueden afectar a la salud de las personas, ¿podría decirme si desde que se declaró la pandemia del coronavirus ha sentido usted miedo por no poder emprender ya proyectos vitales como emanciparse, o abrir un negocio, o hacer algún viaje?

TemorFuturo La pandemia del covid-19 ha dado lugar a diversas situaciones que pueden afectar a la salud de las personas, ¿podría decirme si desde que se declaró la pandemia del coronavirus ha sentido usted inquietud o temor ante el futuro?

Aliment. A lo largo de estos meses de pandemia hay personas que han estado reflexionando o pensando sobre distintos aspectos de su vida. Me gustaría saber si usted personalmente en estos meses ha tomado decisiones o ha hecho propósitos para mejorar sus hábitos de alimentación.

ActFisica A lo largo de estos meses de pandemia hay personas que han estado reflexionando o pensando sobre distintos aspectos de su vida. Me gustaría saber si usted personalmente en estos meses ha tomado decisiones o ha hecho propósitos para mejorar su actividad física.

Salud A lo largo de estos meses de pandemia hay personas que han estado reflexionando o pensando sobre distintos aspectos de su vida. Me gustaría saber si usted personalmente en estos meses ha tomado decisiones o ha hecho propósitos para mejorar su salud

RelFamilia A lo largo de estos meses de pandemia hay personas que han estado reflexionando o pensando sobre distintos aspectos de su vida. Me gustaría saber si usted personalmente en estos meses ha tomado decisiones o ha hecho propósitos para mejorar su relación con la familia

RelVecinos A lo largo de estos meses de pandemia hay personas que han estado reflexionando o pensando sobre distintos aspectos de su vida. Me gustaría saber si usted personalmente en estos meses ha tomado decisiones o ha hecho propósitos para mejorar su relación con los/as vecinos/as.

ActVolntariado A lo largo de estos meses de pandemia hay personas que han estado reflexionando o pensando sobre distintos aspectos de su vida. Me gustaría saber si usted personalmente en estos meses ha tomado decisiones o ha hecho propósitos para mejorar su implicación en actividades de voluntariado y de ayuda comunitaria.

Ocio A lo largo de estos meses de pandemia hay personas que han estado reflexionando o pensando sobre distintos aspectos de su vida. Me gustaría saber si usted personalmente en estos meses ha tomado decisiones o ha hecho propósitos para mejorar sus actividades de ocio.

Amistades A lo largo de estos meses de pandemia hay personas que han estado reflexionando o pensando sobre distintos aspectos de su vida. Me gustaría saber si usted personalmente en estos meses ha tomado decisiones o ha hecho propósitos para mejorar sus amistades, sus relaciones sociales.

ActPrin A lo largo de estos mese de pandemia hay personas que han estado reflexionando o pensando sobre distintos aspectos de su vida. Me gustaría saber si usted personalmente en estos meses ha tomado decisiones o ha hecho propósitos para mejorar su trabajo, sus estudios, su actividad principal.

Objetivos del ejemplo

Para este ejemplo se han planteado los siguientes objetivos:

Objetivo 1. Analizar las correlaciones existentes entre las matrices de datos obtenidas de las encuestas (I), (V) y (VI).

Objetivo 2. Estudiar si existe una estructura común para las encuestas de "Efectos y consecuencias del Coronavirus" realizada en tres momentos diferentes.

Objetivo 3. Examinar la Estructura compromiso generada para los datos de las diferentes encuestas analizadas.

Objetivo 4. Representar sobre las variables de la Estructura compromiso las respuestas individuales de cada momento ampliando el estudio de dicha Estructura.

Metodología

Para analizar estos datos emplearemos las técnicas presentadas en este capítulo.

	EncI	EncV	EncVI
EncI	1.0000000	0.9818109	0.9792816
EncV	0.9818109	1.0000000	0.9782357
EncVI	0.9792816	0.9782357	1.0000000

Tabla 7.1: Correlaciones entre ocasiones

Utilizando el software estadístico R (R Core Team, 2021), y en concreto el paquete MultBiplotR (Vicente-Villardon, 2021), donde se han implementado las funciones para el STATIS tetracorico Dual, se realizarán tanto los cálculos como las representaciones gráficas descritas en las secciones anteriores.

El biplot presentado en este ejemplo permitirá estudiar las contribuciones de la estructura compromiso de este tipo de STATIS.

Resultados

Comenzaremos analizando los resultados obtenidos en la correlación entre las encuestas realizadas. Nótese que en el ejemplo denominaremos "EncI" a la encuesta "Efectos y Consecuencias del Coronavirus (I)", "EncV" a la encuesta "Efectos y Consecuencias del Coronavirus (V)" y "EncVI" a la encuesta "Efectos y Consecuencias del Coronavirus (VI)".

Las correlaciones entre los objetos se encuentran en la tabla siguiente (tabla 7.1).

En ella se pueden observar como las correlaciones entre todos los momentos de la encuesta son muy altas.

Es posible observar esta misma información en el gráfico de la Interestructura generado para el Tetra-STATIS Dual, que se encuentra en la figura 7.3.

En ella se puede observar como las tres encuestas tienen una estructura muy pareci-

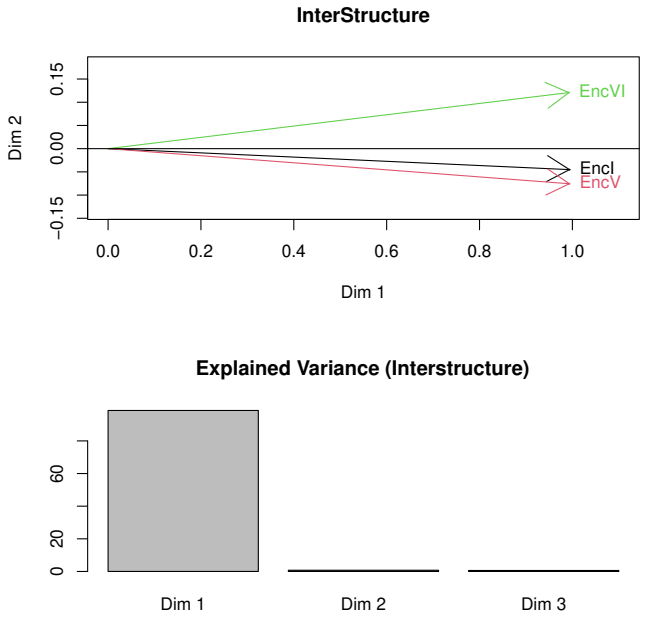

Figura 7.3: Representación de la *Interestructura*

da, aunque la última encuesta realizada presenta una pequeña diferencia con respecto de las dos restantes. Si observamos la tabla 7.1 podemos observar que su correlación es un poquito menor que la correlación entre las encuestas I y V. Esto puede deberse a que la fecha de realización de la encuesta coincidió con la entrada en vigor de nuevas medidas para la prevención del COVID-19 que ampliaba las posibilidades de viajar entre países.

Otro de los puntos de interés en el estudio de varias matrices de forma simultánea, es el análisis del comportamiento de las variables dentro de la estructura común. En nuestro ejemplo, la figura 7.4 muestra la Estructura compromiso del Tetra-STATIS mediante el círculo de correlaciones.

En este gráfico se puede observar que las variables se agrupan en dos conjuntos bien diferenciados, uno asociado a cada eje. Si observamos las variables de cada uno de los grupos podemos apreciar que las variables asociadas al eje 2 pertenecen a la pregunta P10 de la encuesta (I) y (V) y P3 de la encuesta (VI) (se puede encontrar la encuesta

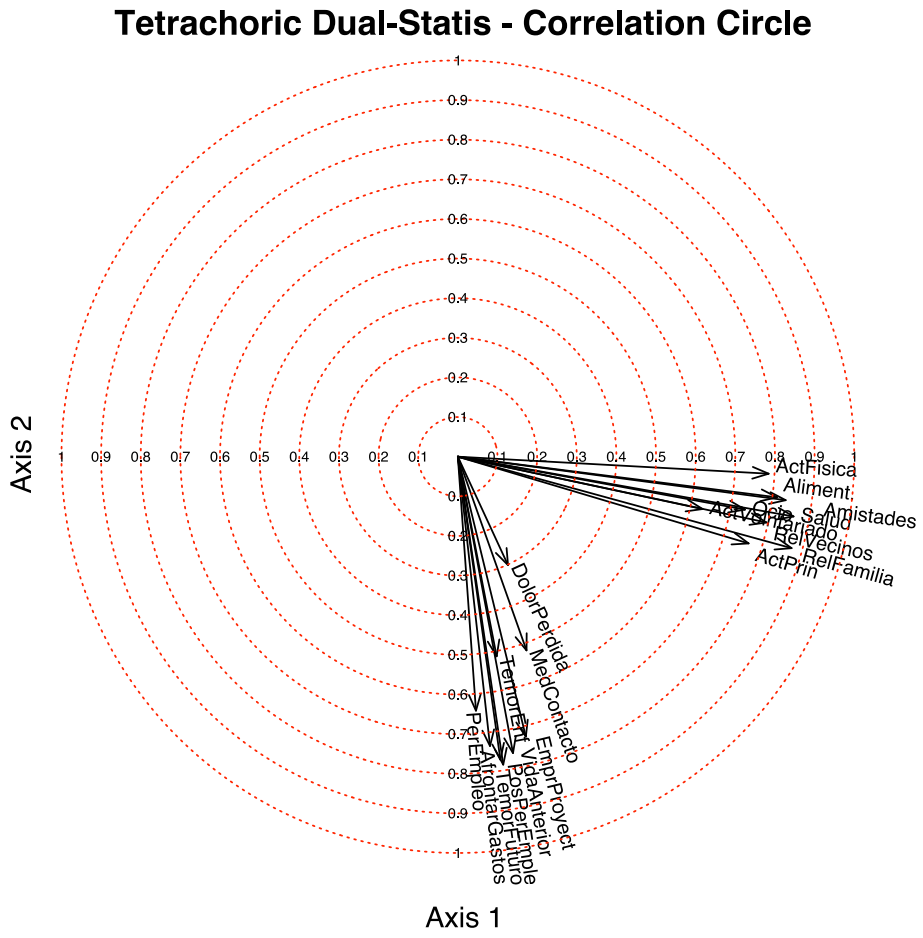

Figura 7.4: Representación de las *Estructura*: Correlaciones compromiso

completa en el Anexo I), que están relacionadas con propositos de mejora que se han planteado los encuestados durante la pandemia. Las variables asociadas al eje 1 pertenecen a la pregunta P5 de la encuesta (I), P3 de la encuesta (V) y P2 de la encuesta (VI), que están relacionadas con los sentimientos de los encuestados desde que comenzó la pandemia. Dentro de cada uno de los grupos las variables están muy relacionadas entre sí.

Las contribuciones de cada una de las variables a la Estructura compromiso del STATIS tetracórico Dual se encuentra en la figura 7.5.

La variable que menos contribuye a la Estructura es el sentimiento de dolor por la

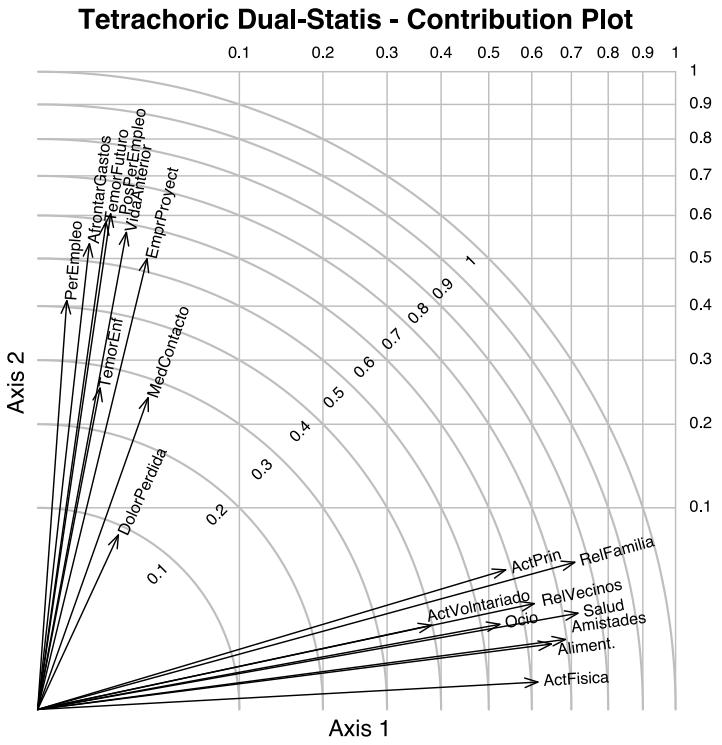

Figura 7.5: Representación de la *Estructura*: Contribuciones del compromiso

pérdida de un familiar, seguida del temor a enfermar y la inquietud por medidas que pudieran limitar el contacto y las relaciones cara a cara, las tres pertenecientes al grupo de preguntas asociado al eje 2. Por el contrario, las variables que más contribuyen a la Estructuras, todas pertenecientes al eje 1, serán los propósitos de mejora de la relación con la familia, la salud y las amistades.

Por último, realizaremos la representación de cada una de las observaciones sobre la Estructura compromiso a través del biplot para el STATIS tetracórico Dual.

Observamos que, sobre el plano de la Estructura consenso que hemos analizado a través de los gráficos anteriores, se han proyectado las observaciones. Cada una de las

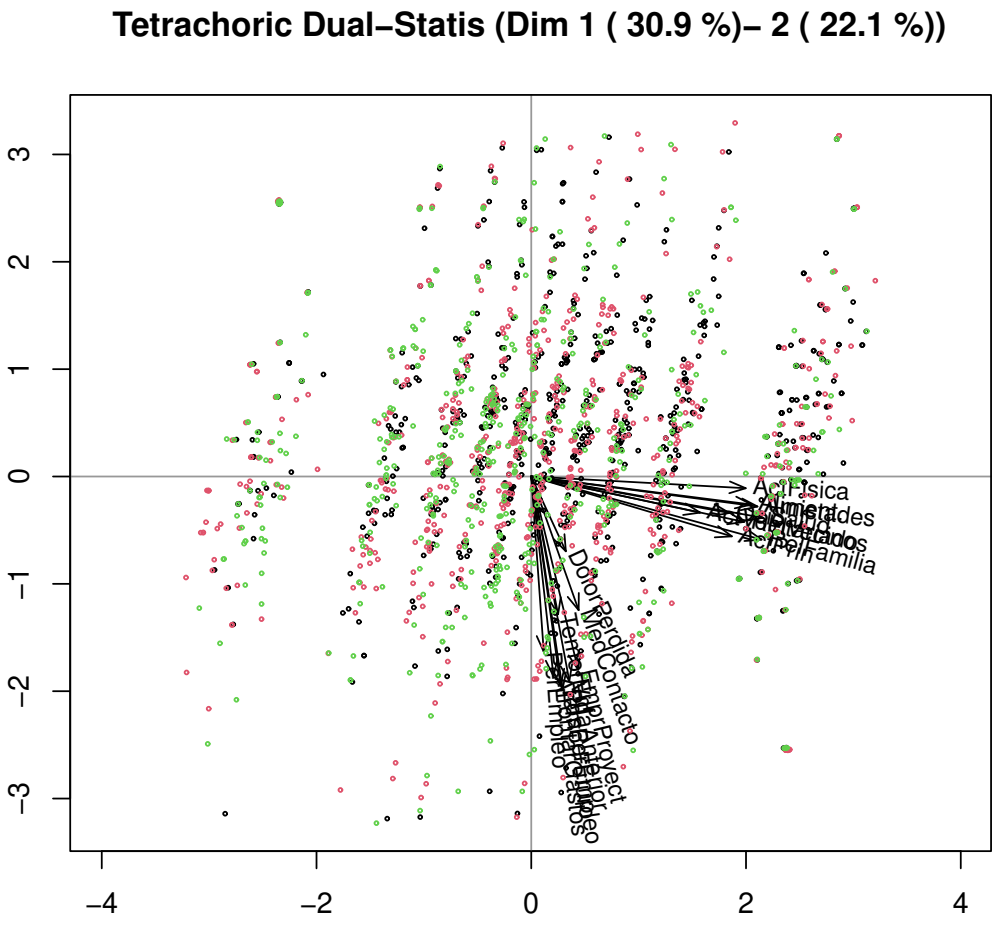

Figura 7.6: Representación de las *Estructura*: Biplot STATIS del compromiso

encuestas realizadas se encuentran representadas en un color diferente. Los puntos de todos los momentos analizados se encuentran mezclados y no se pueden identificar cada una de las encuestas por separado.

Conclusiones del ejemplo

- Las correlaciones entre las matrices de datos obtenidas de las encuestas (I), (V) y (VI) son fuertes en todos los casos, aunque son mayores entre las encuestas (I) y (V).

- Existe una estructura común muy representativa de todas las las encuestas de "Efectos y consecuencias del Coronavirus" en los tres momentos de estudio.
- Las variables dentro de la Estructura compromiso se dividen en dos grupos, cada uno asociado a una de las preguntas incluidas en este análisis. Dentro de cada uno de los grupos las correlaciones entre las variables son muy altas. Por regla general, las respuestas asociadas a los propósitos de mejora tienen contribuciones más altas a la estructura compromiso que los sentimientos presentes desde que comenzó la pandemia.
- La representación biplot muestra que los puntos de todos los momentos analizados se encuentran mezclados y no se pueden identificar cada una de las encuestas por separado al representarlas en la estructura común.

Conclusiones

1 Se ha realizado una revisión de los fundamentos de los métodos biplot necesarios para el desarrollo de este trabajo, tanto para datos cuantitativos, como para datos binarios. La revisión se ha hecho tanto desde el punto de vista teórico como del de su interpretabilidad e utilidad en el análisis de datos. Se han estudiado métodos iterativos para realizar el cálculo de las Componentes Principales de una única matriz de datos continuos a través del algoritmo NIPALS. Basándonos en la aplicación recursiva de este algoritmo, hemos propuesto una generalización para datos binarios.

2 Se han estudiado y descrito los fundamentos teóricos de los Modelos Lineales Generales Multivariantes y el Análisis Multivariante de la Varianza, con el objeto de mostrar algunas de sus deficiencias para el análisis de matrices de datos, por ejemplo, cuando se dispone de variables colineales o de un número muy elevado. Se revisa la utilización de los test de permutaciones para obtener el denominado PERMANOVA y se completa la propuesta de utilizar Bootstraping iniciada en trabajos anteriores, que hemos denominado BOOTMANOVA. Como método de representación gráfica de las técnicas estudiadas se ha utilizado un Análisis de Coordenadas Principales sobre los centroides que extiende el Análisis Canónico tradicionalmente asociado al MANOVA.

3 Se han estudiado los antecedentes y los fundamentos teóricos del Análisis de la Redundancia para datos continuos con sus representaciones biplot asociadas, y se ha propuesto una generalización cuando las respuestas son binarias y los predictores

continuos. Se obtiene un conjunto de variables latentes para el grupo de variables respuesta que son combinación lineal de las predictoras. La relación entre las latentes y las respuestas es logística. Se ha denominado Análisis de la Redundancia para datos de respuesta binaria.

4 Se ha revisado la Regresión de Mínimos Cuadrado Parciales como alternativa a la Regresión Lineal Multivariante como método para estudiar las relaciones entre dos conjuntos de datos cuando las variables predictoras son muy numerosas o colineales. Cuando las respuestas son binarias, se utiliza el PLS con variables dummy que no es óptimo, de la misma manera que una Regresión Lineal no es óptima cuando la respuesta es binaria. Proponemos una extensión de la técnica PLS cuando disponemos de un conjunto de respuestas basada en la generalización del algoritmo NIPALS para este tipo de datos. Hemos denominado a esta técnica Regresión Logística Binaria de Mínimos Cuadrados Parciales.

5 Revisamos los conceptos fundamentales de la técnica denominada STATIS-Dual, utilizada cuando buscamos obtener una estructura común de varias tablas en las que se han medido un conjunto de variables continuas comunes y proponemos el método STATIS tetracórico Dual para datos binarios, que emplea las correlaciones tetracóricas en lugar de la correlaciones de Pearson. Esta propuesta extiende, no solamente el STATIS-Dual al caso de datos binarios sino también el denominado DISTATIS cuando los individuos son comunes.

6 Se ha realizado una revisión de los software con los que se pueden realizar las representaciones biplot, el PERMANOVA, el BOOTMANOVA, el Análisis de Coordenadas Principales sobre los centroides y sus regiones de confianza asociadas, el Análisis de la Redundancia para datos continuos, la Regresión de Mínimos Cuadrados Parciales y el método STATIS-Dual. Se ha detallado si se trata de paquetes comerciales o no, y si disponen de interfaz gráfica. Se ha puesto particular énfasis en los paquetes del software estadístico R.

7 Se han realizado y se han recogido en un único paquete las funciones propias, en R, para el cálculo de las distancias, la realización del MANOVA, el PERMANOVA, el BOOTMANOVA, el Análisis de Coordenadas Principales sobre los centroides y sus regiones de confianza asociadas. Se contempla la opción de realizar los análisis con modelos simples y modelos más complejos que incluyan matrices de contrastes, por ejemplo. Dicho paquete se ha denominado PERMANOVA y está alojado en el repositorio CRAN.

8 Se han elaborado funciones, en R, que permiten llevar a cabo los métodos propuestos para el Análisis de la Redundancia, el PLS y el STATIS-Dual para datos binarios, y se han colocado como parte del paquete MultBiplotR que ha quedado alojado en el repositorio CRAN.

9 Todas las propuestas desarrolladas se han aplicado a conjuntos de datos reales mediante la utilización de los procedimientos de software detallados. Las aplicaciones muestran la clara utilidad práctica de las propuestas teóricas. Entre los datos utilizados podemos mencionar:

- Encuesta sobre "Efectos y Consecuencias del Coronavirus" realizada por el CIS en varios momentos desde noviembre de 2020 hasta diciembre de 2021.
- Datos genómicos de distintas cepas de cepas de *Colletotrichum graminicola*.
- Datos del proyecto Hapmap.
- Datos de presencia/ausencia de varias especies de arañas relacionados con las condiciones ambientales de los lugares de muestreo.
- Datos de caracterización de vinos jóvenes de Ribera de Duero y Toro.

Bibliografía

Aart, P. J. V. D. y Smeenk-Enserink, N. (1974). Correlations between distributions of hunting spiders (lycosidae, ctenidae) and environmental characteristics in a dune area. *Netherlands Journal of Zoology*, 25:1–45.

Abdi, H., O'Toole, A. J., Valentin, D., y Edelman, B. (2005). Distatis: The analysis of multiple distance matrices. En *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)-Workshops*, pp. 42–42. IEEE.

Abdi, H., Williams, L. J., Valentin, D., y Bennani-Dosse, M. (2012). Statis and distatis: optimum multitable principal component analysis and three way metric multidimensional scaling. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(2):124–167.

Ackerman, T. (1996). Graphical representation of multidimensional item response theory analyses. *Applied psychological measurement*, 20(4):311–329.

Addinsoft (2022). Xlstat.

Aelst, S. V. y Willems, G. (2011). Robust and efficient one-way manova tests on jstor. *Journal of American Statistical Association*, 106:706–718.

Afanador, N. L., Tran, T. N., y Buydens, L. M. (2013). Use of the bootstrap and permutation methods for a more robust variable importance in the projection metric for partial least squares regression. *Analytica Chimica Acta*, 768:49–56.

Albert, A. y Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71:1–10.

Amaro, R., Vicente-Villardon, J. L., y Galindo Villardón, M. P. (2008). Contribuciones al manova-biplot: regiones de confianza alternativas. *Revista de Investigación Operacional*, 29:231–241.

Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, 26(1):32–46.

Anderson, M. J. (2005). Permutational multivariate analysis of variance. *Department of Statistics, University of Auckland, Auckland*, 26:32–46.

Anderson, M. J. y Willis, T. J. (2003). Canonical analysis of principal coordinates: a useful method of constrained ordination for ecology. *Ecology*, 84(2):511–525.

Anderson, P. K., Cunningham, A. A., Patel, N. G., Morales, F. J., Epstein, P. R., y Daszak, P. (2004). Emerging infectious diseases of plants: pathogen pollution, climate change and agrotechnology drivers. *Trends in ecology & evolution*, 19:535–544.

Andrecut, M. (2009). Parallel gpu implementation of iterative pca algorithms. *Journal of Computational Biology*, 16(11):1593–1599.

Anzanello, M. J. y Fogliatto, F. S. (2014). A review of recent variable selection methods in industrial and chemometrics applications. *European Journal of Industrial Engineering*, 8:619–645.

Arnold, S. F. (1981). *The theory of linear models and multivariate analysis*. Wiley.

Babativa-Márquez, J. G. (2020). *BiplotML: Biplots Estimation with Machine Learning Algorithms*.

Babativa-Márquez, J. G. y Vicente-Villardón, J. L. (2021). Logistic biplot by conjugate gradient algorithms and iterated svd. *Mathematics 2021, Vol. 9, Page 2015*, 9:2015.

Baker, F. B. y Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques*. CRC press.

Barker, M. y Rayens, W. (2003). Partial least squares for discrimination. *Journal of Chemometrics*, 17:166–173.

Bartholomew, D. J., Knott, M., y Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach*. John Wiley & Sons.

Bastien, P., Vinzi, V. E., y Tenenhaus, M. (2005). Pls generalised linear regression. *Computational Statistics & Data Analysis*, 48:17–46.

Bell, J. T., Pai, A. A., Pickrell, J. K., Gaffney, D. J., Pique-Regi, R., Degner, J. F., Gilad, Y., y Pritchard, J. K. (2011). DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biology*, 12:R10.

Benzécri, J.-P. (1973). *L'Analyse des Données*, volumen 2. Dunod Paris.

Bergner, Y., Halpin, P., y Vie, J.-J. (2022). Multidimensional item response theory in the style of collaborative filtering. *Psychometrika*, 87(1):266–288.

Blazquez-Zaballos, A. (1998). Análisis biplot basado en modelos lineales generalizados.

Bock, R. D. y Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an em algorithm. *Psychometrika*, 46(4):443–459.

Bonifay, W. (2019). *Multidimensional item response theory*. Sage Publications.

Boriah, S., Chandola, V., y Kumar, V. (2008). Similarity measures for categorical data: A comparative evaluation. En *Proceedings of the 2008 SIAM International Conference on Data Mining*, pp. 243–254. SIAM.

Braak, C. J. T. (1986). Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology*, 67:1167–1179.

Cai, L., Choi, K., Hansen, M., y Harrell, L. (2016). Item response theory. *Annual Review of Statistics and Its Application*, 3:297–321.

Castellan, N. J. (1966). On the estimation of the tetrachoric correlation coefficient. *Psychometrika*, 31(1):67–73.

Chen, Y., Li, X., y Zhang, S. (2019). Joint maximum likelihood estimation for high-dimensional exploratory item factor analysis. *Psychometrika*, 84(1):124–146.

Cheng, R. y Palmer, A. A. (2013). A simulation study of permutation, bootstrap, and gene dropping for assessing statistical significance in the case of unequal relatedness. *Genetics*, 193(3):1015–8.

Choi, S.-S., Cha, S.-H., y Tappert, C. C. (2010). A survey of binary similarity and distance measures. *Journal of Systemics, Cybernetics and Informatics*, 8(1):43–48.

Clarke, K. R. (1993). Non-parametric multivariate analyses of changes in community structure. *Australian journal of ecology*, 18(1):117–143.

Clarke, K. R., Gorley, R. N., Somerfield, P. J., Anderson, M. J., Afeitado, L., Punnet, A., Euinton, D., Brien, S., Goldie, A., Adegoke, N., Smith, A. N. H., Cruz-Motta, J. J., Castro, E. G., Quintino, V., Mascaró, M., Martins, P. M., y Miller, A. (2017). Primer-e.

Cuadras Avellanas, C. (1988). Distancias estadísticas. *Estadística Española*, (119):295–357.

de Falguerolles, A. (1998). Chapter 35 - log-bilinear biplots in action. En Blasius, J. y Greenacre, M. J., editores, *Visualization of Categorical Data*, pp. 527–539. Academic Press, San Diego.

de Leeuw, J. (2006). Principal component analysis of binary data by iterated singular value decomposition. *Computational Statistics and Data Analysis*, 50(1):21–39.

de Leeuw, J. y Meulman, J. (1986). A special jackknife for multidimensional scaling. *Journal of Classification*, 3(1):97–112.

Deloukas, P. y Bentley, D. (2004). The HapMap project and its application to genetic studies of drug response. *The Pharmacogenomics Journal*, 4(2):88–90.

Demey, J. R., Vicente-Villardon, J. L., Galindo-Villardón, M. P., y Zambrano, A. Y. (2008). Identifying molecular markers associated with classification of genotypes by external logistic biplots. *Bioinformatics*, 24:2832–2838.

des Plantes, H. L. (1976). *Structuration des tableaux à trois indices de la statistique: théorie et application d'une méthode d'analyse conjointe*. Tesis doctoral, Université des sciences et techniques du Languedoc.

Divgi, D. R. (1979). Calculation of the tetrachoric correlation coefficient. *Psychometrika*, 44(2):169–172.

Dray, S., Dufour, A.-B., Thioulouse, J., Jombart, T., Pavoine, S., Lobry, J. R., Ollier, S., Borcard, D., Legendre, P., Bougeard, S., y Siberchicot, A. (2022). *Analysis of Ecological Data: Exploratory and Euclidean Methods in Environmental Sciences*. R package version 1.7-19.

Duarte, L. C., Von Zuben, F. J. A., y Reis, S. A. F. d. (1998). Orthogonal projections and bootstrap resampling procedures in the study of intraspecific variation. *Genetics and Molecular Biology*, 21.

Dumble, S. (2022). *GGEBiplots: GGE Biplots with 'ggplot2'*. R package version 0.1.3.

Eeuwijk, F. A. V. (1995). Linear and bilinear models for the analysis of multi-environment trials: I. an inventory of models. *Euphytica*, 84:1–7.

Efron, B. (1979). Bootstrap methods: another look at the jacknife. *Annals of Statistics*, 7:1–26.

Efron, B. y Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy on jstor. *Statistical Science*, 1:54–75.

Efron, B. y Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Eigenvector Research, I. (2022). Pls_toolbox.

Elizondo, W. C. y Varela, J. G. (1998). Statis dual: Software y análisis de datos reales. *Revista de Matemática: Teoría y Aplicaciones*, 5:149–162.

Eslami, A., Qannari, E. M., Bougeard, S., Questions, G. S., comments go to Aida Eslami, y Bougeard, S. (2020). *multigroup: Multigroup Data Analysis*. R package version 0.4.5.

Figueiredo, A. (2017). Bootstrap and permutation tests in anova for directional data. *Computational Statistics*, 32:1213–1240.

Finney, D. J. (1952). Probit analysis. 2nd ed. cambridge university press. *Journal of the American Pharmaceutical Association*, 41:627–627.

Firinguetti, L., Kibria, G., y Araya, R. (2017). Study of partial least squares and ridge regression methods. *http://dx.doi.org/10.1080/03610918.2016.1210168*, 46:6631–6644.

Gabriel, K. (1981). Biplot display of multivariate matrices for inspection of data and diagnosis. En Barnet, V., editor, *Interpreting Multivariate Data*, pp. 147–173. John Wiley and Sons.

Gabriel, K. (1995). Biplot display of multivariate categorical data, with comments on multiple correspondence analysis. *Recent advances in descriptive multivariate analysis*, 190:226.

Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58:453.

Gabriel, K. R. (1972). Analysis of meteorological data by means of canonical decomposition and biplots on jstor. *Journal of Applied Meteorology*, 11:1071–1077.

Gabriel, K. R. (1998). Generalised bilinear regression. *Biometrika*, 85:689–700.

Gabriel, K. R. y Odoroff, C. L. (1990). Biplots in biomedical research. *Statistics in medicine*, 9(5):469–485.

Galindo-Villardón, P. (1986). Una alternativa de representación simultánea: Hj-biplot. *Questiió: Quaderns d'Estadística, Sistemes, Informatica i Investigació Operativa*, 10:13–23.

Gardner-Lubbe, S., Roux, N. J. L., y Gower, J. C. (2008). Measures of fit in principal component and canonical variate analyses. *Journal of Applied Statistics*, 35:947–965.

Gifi, A. (1990). *Nonlinear multivariate analysis*. Wiley-Blackwell.

Gitschier, J. (2009). Inferential Genotyping of Y Chromosomes in Latter-Day Saints Founders and Comparison to Utah Samples in the HapMap Project. *The American Journal of Human Genetics*, 84(2):251–258.

Golub, G. y Van Loan, C. (2013). *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press.

Gonzalez, I. y Voillet, V. (2022). *missRows: Handling Missing Individuals in Multi-Omics Data Integration*. R package version 1.18.0.

Goodnight, C. J. y Schwartz, J. M. (1997). A bootstrap comparison of genetic covariance matrices. *Biometrics*, 53:1026.

Goshtasby, A. A. (2012). Similarity and Dissimilarity Measures. En *Image Registration*, pp. 7–66. Springer London, London.

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53(3-4):325–338.

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

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, 27(4):857–871.

Gower, J. C. y Hand, D. J. (1995). *Biplots*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

Gower, J. C. y Hand, D. J. (1996). *Biplots*, volumen 54. Chapman & Hall.

Gower, J. C. y Krzanowski, W. J. (1999). Analysis of distance for structured multivariate data and extensions to multivariate analysis of variance. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48(4):505–519.

Gower, J. C., Lubbe, S. G., y Le Roux, N. J. (2011). *Understanding biplots*. John Wiley and Sons.

Gray, A. y Markel, J. (1976). Distance measures for speech processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(5):380–391.

Greenacre, M. J. (1984). *Theory and Applications of Correspondence Analysis*. Academic Press Inc. (London) LTD., 3 edición.

Greenacre, M. J. (2018). *Compositional Data Analysis in Practice*. Chapman & Hall / CRC Press.

Hammer, O., Harper, D., y Ryan, P. (2001). Past: Paleontological statistics software package for education and data analysis. *Palaeontologia Electronica*, 4:1–9.

Han, D., Dezert, J., Han, C., y Yang, Y. (2011). New Dissimilarity Measures in Evidence Theory. En *Information Fusion (FUSION), 2011 Proceedings of the 14th International Conference on*, pp. 1–7. IEEE.

Hand, D. J., Daly, F., McConway, K., Lunn, D., y Ostrowski, E. (1993). *A Handbook of Small Data Sets*. CRC Press.

Heinze, G. y Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine*, 21:2409–2419.

Hernández-Sánchez, J. C. y Vicente-Villardón, J. L. (2017). Logistic biplot for nominal data. *Advances in Data Analysis and Classification*, 11:307–326.

Highland Statistics Ltd. (2017). Brodgar.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, 28(3/4):321–377.

IBM Corp. (2021). Ibm spss statistics.

Inc., W. R. (2022). Mathematica, Version 13.1. Champaign, IL, 2022.

Inc. Statgraphics Technologies (2020). Statgraphics.

Instituto Nacional de Estadística (INE) (2019). Gráficos de ayer y hoy.

International HapMap Consortium (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164):851–861.

International HapMap Consortium y others (2004). Integrating ethics and science in the International HapMap Project. *Nature reviews. Genetics*, 5(6):467–475.

International HapMap Consortium y others (2005). A haplotype map of the human genome. *Nature*, 437(7063):1299–1320.

Israels, A. Z. (1984). Redundancy analysis for qualitative variables. *Psychometrika 1984 49:3*, 49:331–346.

James Grace, Martin Hutten, B. M. M. M. J. P. y Hatch, W. (2018). Pc-ord.

Janssen, A. y Pauls, T. (2005). A monte carlo comparison of studentized bootstrap and permutation tests for heteroscedastic two-sample problems. *Computational Statistics*, 20:369–383.

Jöreskog, K. G. y Moustaki, I. (2001). Factor analysis of ordinal variables: A comparison of three approaches. *Multivariate Behavioral Research,*, 36(3):347–387.

Kirk, D. B. (1973). On the numerical approximation of the bivariate normal (tetrachoric) correlation coefficient. *Psychometrika*, 38(2):259–268.

Konietschke, F., Bathke, A. C., Harrar, S. W., y Pauly, M. (2015). Parametric and nonparametric bootstrap methods for general manova. *Journal of Multivariate Analysis*, 140:291–301.

Kovach, W. (1999). Multivariate statistical package (mvsp).

Krishnamoorthy, K. y Lu, F. (2010). A parametric bootstrap solution to the manova under heteroscedasticity. *Journal of Statistical Computation and Simulation*, 80:873–887.

Kuhn, M. (2022). *caret: Classification and Regression Training*. R package version 6.0-93.

Lavit, C., Escoufier, Y., Sabatier, R., y Traissac, P. (1994). The act (statis method). *Computational Statistics & Data Analysis*, 18(1):97–119.

le Cessie, S. y van Houwelingen, J. C. (1992). Ridge estimators in logistic regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 41:191–201.

Legendre, P. y Anderson, M. J. (1999). Distance-based redundancy analysis: Testing multispecies responses in multifactorial ecological experiments. *Ecological Monographs*, 69:1–24.

Legendre, P. y Legendre, L. F. (2012). *Numerical Ecology*. Elsevier.

Librero, A. B. N., Baccala, N., Galindo, P. V., y Villardon, P. G. (2022). *multibiplotGUI: Multibiplot Analysis in R*. R package version 1.1.

Librero, A. B. N., Villardon, P. G., y Freitas, A. (2019). *biplotbootGUI: Bootstrap on Classical Biplots and Clustering Disjoint Biplot*. R package version 1.2.

Liland, K. H., Mevik, B.-H., y Wehrens, R. (2022). *pls: Partial Least Squares and Principal Component Regression*. R package version 2.8-1.

Lin, Z., Lopes, M. E., y Müller, H. G. (2021). High-dimensional manova via bootstrapping and its application to functional and sparse count data. *Journal of the American Statistical Association*.

LLC (2022). Minitab 17.

Llobell, F., Vigneau, E., Cariou, V., y Qannari, E. M. (2022). *ClustBlock: Clustering of Datasets*. R package version 3.0.0.

Ltd. Analyse-it Software (2022). Analyse-it.

Manolio, T. A., Brooks, L. D., y Collins, F. S. (2008). A HapMap harvest of insights into the genetics of common disease. *The Journal of Clinical Investigation*, 118(5):1590–1605.

Mardia, K., Kent, J., y Bibby, J. (2009). *Multivariate Analysis*. Academic Press.

Martín-Rodríguez, J., Galindo-Villardón, M. P., y Vicente-Villardón, J. L. (2002). Comparison and integration of subspaces from a biplot perspective. *Journal of Statistical Planning and Inference*, 102(2):411–423.

McArdle, B. H. y Anderson, M. J. (2001). Fitting Multivariate Models to Community Data: A Comment on Distance-Based Redundancy Analysis. *Ecology*, 82:290–297.

McVean, G., Spencer, C. C. A., y Chaix, R. (2005). Perspectives on Human Genetic Variation from the HapMap Project. *PLOS Genetics*, 1(4):e54.

Michailidis, G. y De Leeuw, J. (1998). The gifi system of descriptive multivariate analysis. *Statistical Science*, pp. 307–336.

Milan, L. y Whittaker, J. (1995). Application of the parametric bootstrap to models that incorporate a singular value decomposition. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 44(1):31–49.

Morrison, D. F. (2005). Multivariate analysis of variance. *Encyclopedia of biostatistics*, 5.

Myers, R. H. y Montgomery, D. C. (2018). A tutorial on generalized linear models. *https://doi.org/10.1080/00224065.1997.11979769*, 29:274–291.

Nash, J. C. (2019). *optimr: A Replacement and Extension of the 'optim' Function*. R package version 12.6.

Nelder, J. A. (1966). Inverse polynomials, a useful group of multi-factor response functions. *Biometrics*, 22:128.

Nelder, J. A. y Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135:370.

Neyman y S., J. (1923). On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. (Tlanslated and edited by D. M. Dabrowska and T. P. Speed, Statistical Science (1990), 5, 465-480). *Annals of Agricultural Sciences*, 10:1–51.

O'Connell, R. J., Thon, M. R., Hacquard, S., Amyotte, S. G., Kleemann, J., Torres, M. F., Damm, U., Buiate, E. A., Epstein, L., Alkan, N., Altmüller, J., Alvarado-Balderrama, L., Bauser, C. A., Becker, C., Birren, B. W., Chen, Z., Choi, J., Crouch, J. A., Duvick, J. P., Farman, M. A., Gan, P., Heiman, D., Henrissat, B., Howard, R. J., Kabbage, M., Koch, C., Kracher, B., Kubo, Y., Law, A. D., Lebrun, M. H., Lee, Y. H., Miyara, I., Moore, N., Neumann, U., Nordström, K., Panaccione, D. G., Panstruga, R., Place, M., Proctor, R. H., Prusky, D., Rech, G., Reinhardt, R., Rollins, J. A., Rounsley, S., Schardl, C. L., Schwartz, D. C., Shenoy, N., Shirasu, K., Sikhakolli, U. R., Stüber, K., Sukno, S. A., Sweigard, J. A., Takano, Y., Takahara, H., Trail, F., Does, H. C. V. D., Voll, L. M., Will, I., Young, S., Zeng, Q., Zhang, J., Zhou, S., Dickman, M. B., Schulze-Lefert, P., Themaat, E. V. L. V., Ma, L. J., y Vaillancourt, L. J. (2012). Lifestyle transitions in plant pathogenic colletotrichum fungi deciphered by genome and transcriptome analyses. *Nature Genetics 2012 44:9*, 44:1060–1065.

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E., y Wagner, H. (2017). *vegan: Community Ecology Package*. R package version 2.4-5.

OriginLab Corporation (2022). Originpro.

Oyedele, O. F. y Lubbe, S. (2015). The construction of a partial least-squares biplot. *Journal of Applied Statistics*, 42:2449–2460.

Pearson, K. (1901). One lines and planes of closest fit to systems of points in space. *Philosophical magazine*, 2:559–72.

Præstgaard, J. T. (1995). Permutation and bootstrap kolmogorov-smirnov tests for the equality of two distributions. *Scandinavian Journal of Statistics*, pp. 305–322.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rao, C. R. (1964). The use and interpretation of principal component analysis in applied research. *Sankhya: The Indian Journal of Statistics, Series A*, 26:329–358.

Rivas-Gonzalo, J., Gutiérrez, Y., Polanco, A., Hebrero, E., Vicente, J., Galindo, P., y Santos-Buelga, C. (1993). Biplot analysis applied to enological parameters in the geographical classification of young red wines. *American journal of enology and viticulture*, 44(3):302–308.

Rotimi, C., Leppert, M., Matsuda, I., Zeng, C., Zhang, H., Adebamowo, C., Ajayi, I., Aniagwu, T., Dixon, M., Fukushima, Y., Macer, D., Marshall, P., Nkwodimmah, C., Peiffer, A., Royal, C., Suda, E., Zhao, H., Wang, V. O., y McEwen, J. (2007). Community Engagement and Informed Consent in the International HapMap Project. *Public Health Genomics*, 10(3):186–198.

Roy, S. (1953). On a heuristic method of test construction and its use in multivariate analysis. *The Annals of Mathematical Statistics*, 24(2):220–238.

SAS Institute Inc. (2022). Statistical analysis software (sas).

Schein, A. I., Saul, L. K., y Ungar, L. H. (2003). A generalized linear model for principal component analysis of binary data. En Bishop, C. M. y Frey, B. J., editores, *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, volumen R4 de *Proceedings of Machine Learning Research*, pp. 240–247. PMLR. Reissued by PMLR on 01 April 2021.

Seber, G. A. F. (2009). *Multivariate observations*. John Wiley & Sons.

Silva, A., Dimas, I. D., Lourenço, P. R., Rebelo, T., y Freitas, A. (2020). Pls visualization using biplots: An application to team effectiveness. volumen 12251 LNTCS, pp. 214–230. Springer Science and Business Media Deutschland GmbH.

Skipper, M. (2007). Genomics: HapMap Phase II unveiled. *Nature Reviews Genetics*, 8(11):827–827.

Smyth, D. J., Cooper, J. D., Bailey, R., Field, S., Burren, O., Smink, L. J., Guja, C., Ionescu-Tirgoviste, C., Widmer, B., Dunger, D. B., Savage, D. A., Walker, N. M., Clayton, D. G., y Todd, J. A. (2006). A genome-wide association study of nonsynonymous

SNPs identifies a type 1 diabetes locus in the interferon-induced helicase (*IFIH1*) region. *Nature Genetics*, 38(6):617–619.

Splawa-Neyman, J., Dabrowska, D. M., y Speed, T. P. (1990). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5:465–472.

StataCorp (2021). Stata statistical software.

StatSoft Europe (2022). Statistica.

Takane, Y. y De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, 52(3):393–408.

Takane, Y. y Shibayama, T. (1991). Principal component analysis with external information on both subjects and variables. *Psychometrika 1991 56:1*, 56:97–120.

ter Braak, C. J. F. (1992). Permutation versus bootstrap significance tests in multiple regression and anova. pp. 79–85.

Terwilliger, J. D. y Hiekkalinna, T. (2006). An utter refutation of the 'Fundamental Theorem of the HapMap'. *European Journal of Human Genetics*, 14(4):426–437.

Thioulouse, J., Chessel, D., Dolédec, S., y Olivier, J. M. (1997). Ade-4: a multivariate analysis and graphical display software. *Statistics and Computing*, 7:75–83.

Thorgeirsson, T. E., Geller, F., Sulem, P., Rafnar, T., Wiste, A., Magnusson, K. P., Manolescu, A., Thorleifsson, G., Stefansson, H., Ingason, A., Stacey, S. N., Bergthorsson, J. T., Thorlacius, S., Gudmundsson, J., Jonsson, T., Jakobsdottir, M., Saemundsdottir, J., Olafsdottir, O., Gudmundsson, L. J., Bjornsdottir, G., Kristjansson, K., Skuladottir, H., Isaksson, H. J., Gudbjartsson, T., Jones, G. T., Mueller, T., Gottsäter, A., Flex, A., Aben, K. K. H., Vegt, F. d., Mulders, P. F. A., Isla, D., Vidal, M. J., Asin, L., Saez, B., Murillo, L., Blondal, T., Kolbeinsson, H., Stefansson, J. G., Hansdottir, I., Runarsdottir, V., Pola, R., Lindblad, B., Rij, A. M. v., Dieplinger, B., Haltmayer, M., Mayordomo,

J. I., Kiemeney, L. A., Matthiasson, S. E., Oskarsson, H., Tyrfingsson, T., Gudbjartsson, D. F., Gulcher, J. R., Jonsson, S., Thorsteinsdottir, U., Kong, A., y Stefansson, K. (2008). A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature*, 452(7187):638–642.

Udina, F. (2004). Interactive biplot construction. *Journal of Statistical Software*, 13.

Udina, F. (2005). Xls-biplot.

van den Wollenberg, A. L. (1977). Redundancy analysis an alternative for canonical correlation analysis. *Psychometrika*, 42:207–219.

Vargas, M., Crossa, J., Eeuwijk, F. A. V., Ramírez, M. E., y Sayre, K. (1999). Using partial least squares regression, factorial regression, and ammi models for interpreting genotype by environment interaction. *Crop Science, Genetics & Cytology*, 39:955–967.

Vicente-Gonzalez, L. y Vicente-Villardon, J. L. (2019). Manova bootstrap basado en distancias.

Vicente-Gonzalez, L. y Vicente-Villardon, J. L. (2021). *PERMANOVA: Multivariate Analysis of Variance Based on Distances andPermutations*. R package version 0.2.0.

Vicente-Gonzalez, L. y Vicente-Villardon, J. L. (2022). Partial least squares regression for binary responses and its associated biplot representation. *Mathematics 2022, Vol. 10, Page 2580*, 10:2580.

Vicente-Villardon, J. L. (2020). Multbiplot.

Vicente-Villardon, J. L. (2021). *MultBiplotR: Multivariate Analysis Using Biplots in R*. R package version 1.6.14.

Vicente-Villardon, J. L., Galindo-Villardón, P., y Blazquez-Zaballos, A. (2006). Logistic biplots. En Greenacre, M. J. y Blasius, J., editores, *Multiple Correspondence Analysis and Related Methods*, Statistics in the Social and Behavioral Sciences, pp. 503–521. Chapman & Hall/CRC.

Vicente-Villardon, J. L. y Vicente-Gonzalez, L. (2021). Redundancy analysis for binary data based on logistic responses. En Chadjipadelis, T., Lausen, B., Markos, A., Lee, T., Montanari, A., y Nugent, R., editores, *Data Analysis and Rationality in a Complex World*, Studies in Classification, Data Analysis, and Knowledge Organization, pp. 331–339. Springer International Publishing.

Vicente-Villardón, J. L. y Hernández Sánchez, J. C. (2014). Logistic biplots for ordinal data with an application to job satisfaction of doctorate degree holders in spain. *undefined*.

Vicente-Villardón, J. L. y Hernández-Sánchez, J. C. (2020). External logistic biplots for mixed types of data. En Imaizumi, T., Okada, A., Miyamoto, S., Sakaori, F., Yamamoto, Y., y Vichi, M., editores, *Advanced Studies in Classification and Data Science*, Studies in Classification, Data Analysis, and Knowledge Organization, pp. 169–183. Springer.

Wold, H. (1975). Soft modelling by latent variables: the non-linear iterative partial least squares (nipals) approach. *Journal of Applied Probability*, 12:117–142.

Wold, S., Esbensen, K., y Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1-3):37–52.

Wold, S., Ruhe, A., Wold, H., y Dunn, I. W. J. (2006). The collinearity problem in linear regression. the partial least squares (pls) approach to generalized inverses. *http://dx.doi.org/10.1137/0905052*, 5:735–743.

Xu, J. y Cui, X. (2008). Robustified manova with applications in detecting differentially expressed genes from oligonucleotide arrays. *Bioinformatics*, 24:1056–1062.

Xu, L. W. (2015). Parametric bootstrap approaches for two-way manova with unequal cell sizes and unequal cell covariance matrices. *Journal of Multivariate Analysis*, 133:291–303.

Yan, W. y Kang, M. (2006). Gge-biplot.

Young, F. W. (1990). Vista the visual statistics system.

Zhang, B. y Srihari, S. N. (2003). Properties of binary vector dissimilarity measures. En *Proc. JCIS Int'l Conf. Computer Vision, Pattern Recognition, and Image Processing*, volumen 1.

Šmilauer, P. (2012). Canoco.

Anexos

Anexo I: Encuestas CIS

En las siguientes páginas se incluye la encuesta realizada por el CIS para obtener la base de datos del ejemplo presentado en la sección 3.5 y en la sección 7.5.

Efectos y consecuencias del Coronavirus (IV)

Encuesta realizada entre el 14 y el 29 de mayo de 2021. En el ejemplo 3.5 se han seleccionado las preguntas P.20A, P.20B, P.21B y P.21C.

Estudio: Efectos y consec. del Coronavirus (IV)
Clave: ECIS3324

«Información sujeta a secreto estadístico (Ley 12/89, de 9 de mayo, de la Función Estadística Pública) y al Reglamento General de Protección de Datos y la Ley Orgánica de Protección de Datos Personales y garantía de los derechos digitales.»

«Buenos días/tardes, soy (nombre propio) y estoy realizando una encuesta telefónica para el Centro de Investigaciones Sociológicas (CIS) sobre temas de interés general. Por este motivo solicitamos su colaboración y se la agradecemos anticipadamente. Este teléfono ha sido obtenido al azar. Esta conversación será grabada para supervisar la calidad y después se borrará en un plazo inferior a un mes, le garantizamos el absoluto anonimato y secreto de sus respuestas en el más estricto cumplimiento de las leyes sobre secreto estadístico y protección de datos personales. Tras la realización de la encuesta su número de teléfono será disociado de las respuestas que pueda dar, que a su vez serán anonimizadas para que en ningún caso puedan ser asociadas a usted. Si desea conocer sus derechos de protección de datos y ampliar esta información puede consultar la página web www.cis.es ¿Ha comprendido la información leída? ¿Sería tan amable de contestar a unas preguntas, algunas de ellas sobre datos de carácter sensible como la intención de voto? No está obligado a contestar todas las preguntas. Muchas gracias.»

PC1. Pregunta contacto 1. ¿Me puede decir a qué provincia y municipio estoy llamando...?

[TIPO_TEL]
FIJO ***1***
MÓVIL ***2***

[PROVINCIA]
_____

[MUNICIPIO]
_____

[CAPITAL]
_____

[TAMUNI]
_____

[ENTREV]

ENTREVISTADOR: SI LA PERSONA QUE CONTESTA ES DIFERENTE DE LA QUE COGIÓ EL TELÉFONO PRESENTARSE:

Buenos días/tardes, mi nombre es… y le llamo del Centro de Investigaciones Sociológicas porque estamos realizando una encuesta de opinión sobre temas de interés general. Dura unos 5 minutos. ¿Sería tan amable de colaborar con nosotros?

[SEXO]
Hombre ***1***
Mujer ***2***

[EDAD]
de 18 a 24 ..... ***1***
de 25 a 34 ..... ***2***
de 35 a 44 ..... ***3***
de 45 a 54 ..... ***4***
de 55 a 64 ..... ***5***
65 y más ***6***

P.0 En primer lugar quisiera preguntarle si tiene Ud....

[P0]
La nacionalidad española ***1***
La nacionalidad española y otra ***2***
Otra nacionalidad ***3***

Saltos:
Si P0=3 ir a fin cuestionario.

P.1 La situación del coronavirus que se está viviendo en España y en otros lugares ¿le preocupa a Ud. mucho, bastante, algo, nada o casi nada?

[P1]
Mucho ***1***
Bastante ***2***
(NO LEER) Regular ***3***
Algo ***4***
Nada o casi nada ***5***
N.S. ***8***
N.C. ***9***

P.2 En general, ¿diría Ud. que su familia directa (padres, hijos/as, nietos/as, abuelos/as, hermanos/as) se está viendo muy afectada por la crisis del coronavirus, bastante, poco, nada o casi nada afectada?

[P2]
Muy afectada ***1***
Bastante afectada ***2***
(NO LEER) Regular ***3***
Poco afectada ***4***
Nada o casi nada afectada ***5***
N.S. ***8***
N.C. ***9***

P.2a ¿Y en qué aspectos se ha visto afectada su familia? RESPUESTA MÚLTIPLE.

[P2A]
En aspectos económicos ***1***
En aspectos laborales ***2***
En aspectos de salud ***3***
En sus relaciones y formas de vivir ***4***
En aspectos emocionales ***5***
En aspectos escolares, de estudios ***6***
En otros aspectos. Especificar ***96***
N.S. ***98***
N.C. ***99***

Filtros:
Si NO P2A=(96) ir a la siguiente.

[P2A_COD]
_____

Estudio: Efectos y consec. del Coronavirus (IV)
Clave: ECIS3324

P3. ¿Ha tenido Ud. el coronavirus?
¿Y algún familiar?
¿Y algún/a amigo/a?
¿Y algún/a conocido/a?
[P3]

	Sí	*No*	*N.S.*	*N.C.*
Ud.	1	2	8	9
Algún familiar	1	2	8	9
Algún/a amigo/a	1	2	8	9
Algún conocido/a	1	2	8	9

Saltos:
Si P3_1=(2;8;9) Y P3_2=(2;8;9)Y P3_3=(2;8;9)Y P3_4=(2;8;9) ir a [P5_1]
Si P3_1=(1) Y P3_2=(2;8;9)Y P3_3=(2;8;9)Y P3_4=(2;8;9) ir a [P5_1]

P4. Y alguna o algunas de estas personas (familiar no conviviente, y/o amigo/a, y/o conocido/a) que lo ha/n tenido, no ha/n podido superar la enfermedad y ha/n fallecido?
[P4]

	Sí	*No*	*N.S.*	*N.C.*
Algún familiar	1	2	8	9
Algún/a amigo/a	1	2	8	9
Algún conocido/a	1	2	8	9

P5. ¿Diría Ud. que esta pandemia está cambiando mucho, bastante, algo, poco o no le está cambiando nada o casi nada su forma de vivir? ¿Y su forma de pensar? ¿Y la forma de cuidar de su salud? ¿Y sus hábitos sociales y comportamiento social?
[P5]

	Mucho	*Bastante*	*(NO LEER) Regular*	*Algo*	*Poco*	*Nada o casi nada*	*Está en duda, no lo sabría decir*	*N.C.*
Su forma de vivir	1	2	3	4	5	6	8	9
Su forma de pensar	1	2	3	4	5	6	8	9
La forma de cuidar de su salud	1	2	3	4	5	6	8	9
Sus hábitos sociales y de comportamiento social	1	2	3	4	5	6	8	9

P.5a ¿Y en qué aspectos principales está cambiando su forma de vivir?
P.5b ¿Y en qué aspectos su forma de pensar?
P.5c ¿Y la forma de cuidar de su salud?
P.5d ¿Y sus hábitos sociales y de comportamiento en la sociedad?
(ENTREVISTADOR/A: DOS RESPUESTAS. EN CADA CASILLA ANOTE TODO LO QUE MENCIONE LA PERSONA ENTREVISTADA)

Filtros:
Si P5_1=(5;6;8;9) ir a [P5B01] - Principales aspectos que está cambiando su forma de pensar (1º)
[P5A01]

Saltos:
Si P5A01="98" O P5A01="99" ir a [P5B01] - Principales aspectos que está cambiando su forma de pensar (1º)

Su forma de vivir
N.S. = 98
N.C. = 99
[P5A02]

N.S. = 98
N.C. = 99
Su forma de pensar

Filtros:
Si P5_2=(5;6;8;9) ir a [P5C01] - Principales aspectos que está cambiando su forma de cuidar de su salud (1º)
[P5B01]

Saltos:
Si P5B01="98" O P5B01="99" ir a [P5C01] - Principales aspectos que está cambiando su forma de cuidar de su salud (1º)

N.S. = 98
N.C. = 99
[P5B02]

N.S. = 98
N.C. = 99
La forma de cuidar de su salud

Filtros:
Si P5_3=(5;6;8;9) ir a [P5D01] - Principales aspectos que está cambiando su forma de cuidar de su salud (1º)
[P5C01]

Estudio: Efectos y consec. del Coronavirus (IV)
Clave: ECIS3324

N.S. = 98
N.C. = 99

[P5C02]

_____

N.S. = 98
N.C. = 99

Sus hábitos sociales y comportamiento en la sociedad

Filtros:
Si P5_4=(5;6;8;9) ir a [P6] - Adecuación de las medidas adoptadas por el Gobierno y las comunidades autónomas

[P5D01]

_____

N.S. = 98
N.C. = 99

[P5D02]

_____

N.S. = 98
N.C. = 99

P.6 En general, ¿cree Ud. que ante los riesgos de la pandemia habría que haber tomado medidas de control más estrictas que las que han tomado el Gobierno español y los gobiernos de las comunidades autónomas, o bien que son (eran) adecuadas y necesarias las medidas adoptadas y no hace falta tomar más medidas, o bien que no hay que tomar medidas que limiten las libertades?

[P6]

- *Habría que haber tomado medidas más estrictas que las que han tomado el Gobierno español y los gobiernos de las comunidades autónomas 1*
- *Eran (son) adecuadas y necesarias las medidas adoptadas 2*
- *No había (hay) que tomar medidas que limiten las libertades 3*
- *No tiene información suficiente 4*
- *N.S. 8*
- *N.C. 9*

P.7 ¿Usa Ud. mascarillas habitualmente como medida de protección?

[P7]

- *Sí 1*
- *No 2*
- *N.C. 9*

Saltos:
Si P7=(2;9) ir a [P8]

P.7A ¿Me podría decir cuántas mascarillas suele Ud. usar a la semana?

[P7A]

- *Una 1*
- *Dos 2*
- *Tres 3*
- *Cuatro 4*
- *Cinco 5*
- *Seis 6*
- *Siete 7*
- *Ocho 8*
- *Nueve o más 9*
- *Utiliza siempre la misma (no es desechable) 97*
- *No sabe/Duda 98*
- *N.C. 99*

P.7B Y piense ahora en un día de actividad normal. ¿Cuántas horas suele Ud. usar la mascarilla a lo largo del día, en lugares públicos (calle, comercio, centro de estudios,transporte, etc.) o de trabajo?

[P7B]

- *Menos de una hora 0*
- *Entre 1 y 2 horas 1*
- *Entre 2 y 3 horas 2*
- *Entre 3 y 4 horas 3*
- *Entre 4 y 5 horas 4*
- *Entre 5 y 6 horas 5*
- *Entre 6 y 7 horas 6*
- *Entre 7 y 8 horas 7*
- *Entre 8 y 9 horas 8*
- *Entre 9 y 10 horas 9*
- *Más de 10 horas 10*
- *N.S. 98*
- *N.C. 99*

P.8 ¿Usa Ud. gel hidroalcohólico para desinfectarse las manos?

[P8]

- *Sí 1*
- *No 2*
- *N.C. 9*

Saltos:
Si P8=(2;9) ir a [P9_1] -

P.8A ¿Cuándo usa Ud. gel hidroalcohólico para desinfectarse las manos? (RESPUESTA MÚLTIPLE)

[P8A]

- *Siempre o casi siempre 1*
- *Cada vez que toca algún objeto (botón, pulsador, picaporte, pasamanos...) 2*
- *Al llegar al trabajo 3*
- *Al llegar a casa 4*
- *Al entrar o salir de un comercio 5*
- *Al entrar o salir de un bar o restaurante 6*
- *En otras situaciones (ESPECIFICAR) 96*
- *N.C. 99*

Filtros:
Si NO P8A=(96) ir a la siguiente.

[P8A_COD]

_____

Estudio: Efectos y consec. del Coronavirus (IV)
Clave: ECIS3324

P.9 Aparte de las mascarillas y el gel hidroalcohólico, de las siguientes medidas de protección frente al virus COVID- 19. ¿cuáles practica Ud. normalmente?
[P9]

	Sí	*No*	*N.C.*
Guarda la distancia de seguridad entre personas recomendada por las autoridades sanitarias	1	2	9
Se lava las manos con frecuencia en casa u otros lugares	1	2	9
Desinfecta productos alimenticios antes de comerlos	1	2	9

P.10 ¿Utiliza Ud. alguna otra medida de protección frente a la COVID-19?
[P10]
***Sí*.....1**
***No*.....2**
***N.C.*.....9**

Saltos:
Si P10=(2;9) ir a [P11_1]

P.10A ¿Podría decirme cuál o cuáles? (ENTREVISTADOR/A: MÁXIMO TRES REPUESTAS. EN CADA CASILLA ANOTE TODO LO QUE MENCIONE LA PERSONA ENTREVISTADA).
N.C = 99
[P10A01]

Saltos:
Si P10A01="99" ir a [P11_1]

Ninguna más = 97
[P10A02]

Saltos:
Si P10A02="97" ir a [P11_1]

MEDIDA 1
MEDIDA 2
[P10A03]

MEDIDA 3
Ninguna más = 97

P.11 En relación con el coronavirus, ¿podría decirme cuántas veces durante los últimos meses (muchas, bastantes, algunas, pocas o ninguna o casi ninguna), Ud. ha tenido...
[P11]

	Muchas veces	*Bastantes veces*	*Algunas veces*	*Pocas veces*	*Nunca o casi nunca*	*N.S.*	*N.C.*
... imágenes, pensamientos o recuerdos desagradables sobre el coronavirus?	1	2	3	4	5	8	9
...pesadillas o no ha podido dormir por esas imágenes, pensamientos o recuerdos sobre el coronavirus?	1	2	3	4	5	8	9
...pensamientos, imágenes o recuerdos que le han provocado que se sienta abrumado/a o agobiado/a?	1	2	3	4	5	8	9
... reacciones físicas como sudoración, taquicardia o de otro tipo producidas por esos pensamientos, recuerdos e imágenes?	1	2	3	4	5	8	9
...pensamientos, recuerdos o imágenes que han alterado sus relaciones familiares o con amigos/as?	1	2	3	4	5	8	9
... pensamientos, recuerdos o imágenes que han alterado su trabajo o actividades de la vida diaria?	1	2	3	4	5	8	9

P.12 En los últimos meses, el coronavirus y las situaciones de confinamiento, ¿le han producido a Ud. (muchos/as, bastantes, algunos/as, pocos/as o ninguno/a), problemas, dificultades o malestar por …
[P12]

	Muchos/as	*Bastantes*	*Algunos/as*	*Pocos/as*	*Ninguno/a o casi ninguno/a*	*N.S.*	*N.C.*
...tener discusiones o conflictos con familiares?	1	2	3	4	5	8	9
...no poder ver a algunos de sus familiares que veía habitualmente?	1	2	3	4	5	8	9
...no poder ver a sus amigos/as?	1	2	3	4	5	8	9
...no poder realizar actividades de ocio fuera de casa (viajar, salir, fiestas, etc.)?	1	2	3	4	5	8	9
...tener la sensación de no saber lo que le puede pasar a su familia?	1	2	3	4	5	8	9

Estudio: Efectos y consec. del Coronavirus (IV)
Clave: ECIS3324

P.13 Actualmente, ¿está Ud. en activo laboralmente, es decir, está trabajando, o en paro buscando empleo?
[P13]

Sí ***1***
No ***2***
N.C. ***9***

Saltos:
Si P13=(2;9) ir a [P14_1]

P.13A En los últimos meses, el coronavirus y las situaciones de confinamiento, ¿le han producido a Ud. muchos/as, bastantes, algunos/as, pocos/as o ninguno/a...
[P13A]

	Muchos/as	*Bastantes*	*Algunos/as*	*Pocos/as*	*Ninguno/a o casi ninguno/a*	*N.C.*
…problemas laborales graves (despido, ERTE, etc.)?	1	2	3	4	5	9
…dificultades en el trabajo (relacionadas con el desplazamiento, cambios de horarios o en las funciones a desempeñar, problemas con compañeros o superiores, problemas con clientes, etc.)?	1	2	3	4	5	9
…sensación de no poder realizar bien el trabajo debido a las inseguridades producidas por la pandemia?	1	2	3	4	5	9

P.14 Durante la pandemia algunas personas han cambiado costumbres o formas de pensar. ¿Diría Ud. que...?
[P14]

	Sí	*No*	*(NO LEER) Igual, lo mismo que antes*	*N.S.*	*N.C.*
...ha aprendido a organizar mejor su tiempo para no aburrirse?	1	2	3	8	9
...ha descubierto aficiones nuevas o actividades que nunca antes había realizado y que le gustan?	1	2	3	8	9
...se ha hecho más religioso/a o espiritual?	1	2	3	8	9
...ha cambiado sus valores y ahora valora y aprecia cosas que antes no?	1	2	3	8	9
...se ha interesado más por la gente que le importa, por si se encuentran bien física y emocionalmente?	1	2	3	8	9
...se ha interesado por el futuro más que antes?	1	2	3	8	9
...ha aprendido a valorar más las relaciones personales?	1	2	3	8	9
...ha aprendido a valorar más los beneficios de las actividades al aire libre?	1	2	3	8	9
...ha disfrutado más de actividades lúdicas con sus familiares (juegos, cocina, etc.)?	1	2	3	8	9

P.15 En general, ¿está durmiendo bien y se siente Ud. suficientemente descansado/a al despertarse, o está durmiendo mal?
[P15]

Está durmiendo bien y se siente suficientemente descansado/a al despertarse ***1***
(NO LEER) Ni bien ni mal, regular ***2***
Está durmiendo mal ***3***
N.S., duda ***8***
N.C. ***9***

P.16 Y en general, ¿está Ud. durmiendo ahora más o menos tiempo que antes de la pandemia?
[P16]

Está durmiendo más ***1***
Está durmiendo menos ***2***
(NO LEER) Prácticamente igual ***3***
N.S., duda ***8***
N.C. ***9***

P.17 ¿Y ahora suele Ud. tener más o menos pesadillas al dormir que antes de la pandemia?
[P17]

Está teniendo más ***1***
Está teniendo menos ***2***
(NO LEER) Prácticamente igual ***3***
(NO LEER) No tiene ***4***
N.S., duda ***8***
N.C. ***9***

P.18 Y en general, desde la pandemia, ¿se suele Ud. acostar más tarde, más pronto, o prácticamente a la misma hora que antes?
[P18]

Más tarde ***1***
Más pronto ***2***
Prácticamente a la misma hora ***3***
N.S. ***8***
N.C. ***9***

Estudio: Efectos y consec. del Coronavirus (IV)
Clave: ECIS3324

P.19 Y desde que estamos en la pandemia, ¿ha mantenido Ud. habitualmente las mismas costumbres y horarios de sus cuidados personales (aseo, vestirse, arreglarse,...), o ha cambiado sus costumbres y horarios?

[P19]

Mantiene las mismas costumbres y horarios 1
Ha cambiado sus costumbres y horarios 2
N.C. 9

Saltos:
Si P19=(1;9) ir a [P20_P20A_1]

P.19A ¿En qué los ha cambiado?

[P19A]

No se arregla todos los días 1
Lo hace más tarde 2
Pone menos esmero en los cuidados personales 3
No atiende prácticamente nada a sus cuidados personales 4
Otra respuesta ¿cuál? (ESPECIFICAR) 96
No sabe, duda 98
N.C. 99

Filtros:
Si NO P19A=(96) ir a la siguiente.

[P19A_COD]

P.20A. Antes de la pandemia, ¿habitualmente solía Ud. ...?
P.20B. Y desde que tenemos la pandemia, ¿suele Ud.,...?

[P20]

	Antes de la pandemia				*Desde la pandemia*			
	Sí	*No*	*N.S.*	*N.C.*	*Sí*	*No*	*N.S.*	*N.C.*
...comer o cenar en días festivos con familiares?	1	2	8	9	1	2	8	9
...asistir a cumpleaños o "santos" de familiares?	1	2	8	9	1	2	8	9
...participar en otras celebraciones familiares como comuniones, bodas o similares?	1	2	8	9	1	2	8	9
...asistir a actividades culturales, deportivas y de ocio?	1	2	8	9	1	2	8	9
...comunicarse por videollamada con más frecuencia que antes?	1	2	8	9	1	2	8	9

P.21 ¿Tiene Ud. parientes mayores de 65 años?

[P21]

Sí 1
No 2
N.C. 9

Saltos:
Si P21=(2;9) ir a [P22] -

P.21A ¿Ha convivido Ud. con parientes mayores de 65 años durante la pandemia?

[P21A]

Sí 1
No 2
N.C. 9

Saltos:
Si P21A=(1;9) ir a [P22] -

P.21B. Antes de la pandemia, con estos parientes mayores con los que no convive (con al menos uno de ellos) ¿habitualmente solía Ud. ...?
P.21C. Y desde que tenemos la pandemia, ¿suele Ud....?

[P21B]

	Antes de la pandemia			*Desde la pandemia*		
	Sí	*No*	*N.C.*	*Sí*	*No*	*N.C.*
...ir a visitarles?	1	2	9	1	2	9
...hacerles la compra?	1	2	9	1	2	9
...acompañarles a hacer gestiones?	1	2	9	1	2	9
...hacerles tareas domésticas (limpiar, hacer la comida)?	1	2	9	1	2	9
...salir con ellos a actividades de entretenimiento, ocio o similares?	1	2	9	1	2	9
...acompañarles al médico, pruebas o análisis?	1	2	9	1	2	9
...hablar regularmente por teléfono?	1	2	9	1	2	9
...comunicarse con ellos habitualmente por videollamada?	1	2	9	1	2	9

Estudio: Efectos y consec. del Coronavirus (IV)
Clave: ECIS3324

P.22 ¿Tiene Ud. hijos/as menores de 18 años con los/as que convive?
[P22]
Sí ***1***
No ***2***
N.C. ***9***

Saltos:
Si P22=(2;9) ir a [P24A01]

P.22A. Antes de la pandemia, ¿las abuelas o abuelos de sus hijos/as (que no viven juntos) solían... ?
P.22B. Y desde que tenemos la pandemia, ¿suelen...?
[P22A]

	Antes de la pandemia			*Desde la pandemia*		
	Sí	***No***	***N.C.***	***Sí***	***No***	***N.C.***
...llevar o traer a sus hijos/as del colegio o escuela infantil?	1	2	9	1	2	9
...cuidar de sus hijos/as mientras Ud. y/o su pareja trabajan?	1	2	9	1	2	9
...cuidar de sus hijos después del colegio?	1	2	9	1	2	9
...quedarse con sus hijos/as durante algún fin de semana?	1	2	9	1	2	9
...cuidar de sus hijos/as si se ponían/ponen enfermos y no podían/pueden ir al colegio o escuela infantil?	1	2	9	1	2	9
...llevar a sus hijos/as a actividades culturales, deportivas o de ocio?	1	2	9	1	2	9
...comunicarse con ellos regularmente por teléfono o videollamada?	1	2	9	1	2	9

P.24A Durante la última semana, ¿podría decirme cuántas horas y/o minutos diarios en promedio ha estado Ud. viendo la televisión...?
P.24B ¿Y escuchando la radio?
P.24C ¿Y conectado a las redes sociales?
(ENTREVISTADOR/A: si la persona entrevistada contesta un intervalo de horas, recoger el número de horas más alto. Si la persona entrevistada no recuerda el tiempo que pasó, anotar "No recuerda" en horas y "No recuerda" en minutos. Si la respuesta es "Nada, cero" (no lo hace o no tiene) anotar "0" en horas y "0" en minutos. Si contesta solo en minutos, poner 0 en horas, y si contesta solo en horas, poner 0 en minutos. Si, por ejemplo, dice 'ocho horas y media', anotar: 8 horas, 30 minutos; si dice 8 horas, anotar: 8 horas, 0 minutos; si dice media hora, anotar 0 horas, 30 minutos. A partir de 59 minutos anotar en horas).
HORAS
Televisión
MINUTOS
[P24A01]

No recuerda = 98
N.C = 99
[P24A02]

No recuerda = 98
N.C = 99
Radio
[P24B01]

No recuerda = 98
N.C = 99
[P24B02]

No recuerda = 98
N.C = 99
Redes sociales
[P24C01]

No recuerda = 98
N.C = 99
[P24C02]

No recuerda = 98
N.C = 99
HORAS
HORAS
MINUTOS
MINUTOS

P.25A Y antes de que comenzara la pandemia, ¿cuántas horas y/o minutos diarios como promedio solía Ud. estar viendo la televisión... ?
P.25B ¿Y escuchando la radio?
P.25C ¿Y conectado a las redes sociales?
(ENTREVISTADOR/A: si la persona entrevistada contesta un intervalo de horas, recoger el número de horas más alto. Si la persona entrevistada no recuerda el tiempo que pasó, anotar "No recuerda" en horas y "No recuerda" en minutos. Si la respuesta es "Nada, cero" (no lo hace o no tiene) anotar "0" en horas y "0" en minutos. Si contesta solo en minutos, poner 0 en horas, y si contesta solo en horas, poner 0 en minutos. Si, por ejemplo, dice 'ocho horas y media', anotar: 8 horas, 30 minutos; si dice 8 horas, anotar: 8 horas, 0 minutos; si dice media hora, anotar 0 horas, 30 minutos. A partir de 59 minutos anotar en horas).
HORAS
Televisión
MINUTOS
[P25A01]

Estudio: Efectos y consec. del Coronavirus (IV)
Clave: ECIS3324

No recuerda = 98
N.C = 99
[P25A02]

No recuerda = 98
N.C = 99
Radio
[P25B01]

No recuerda = 98
N.C = 99
[P25B02]

No recuerda = 98
N.C = 99
Redes sociales
[P25C01]

No recuerda = 98
N.C = 99
[P25C02]

No recuerda = 98
N.C = 99
HORAS
HORAS
MINUTOS
MINUTOS

P.26 Me dijo Ud. que durante la última semana vio algo de tiempo la televisión. ¿Qué canal o canales de televisión ha visto? (ENTREVISTADOR/A: RESPUESTA ESPONTÁNEA. SELECCIONAR EN LA LISTA DESPLEGABLE COMO MÁXIMO TRES, DEL MÁS AL MENOS VISTO. TAMBIÉN PUEDE SELECCIONAR EN LA LISTA "NO RECUERDA" Y "NO CONTESTA" O "NINGUNO" Y "NINGUNO MÁS")

Filtros:
Si (P24A01=0 Y P24A02=0) ir a [P2701] -
Si (P24A01=99 Y P24A02=99) ir a [P2701]

[P2601]

- ***TVE1*** ***1***
- ***La 2*** ***2***
- ***Antena 3*** ***3***
- ***La Cuatro*** ***4***
- ***Telecinco*** ***5***
- ***La Sexta*** ***6***
- ***Canal 24 horas*** ***7***
- ***TVE (Sin especificar canal)*** ***8***
- ***Canal Sur*** ***9***
- ***TV3*** ***10***
- ***8TV*** ***11***
- ***3/24*** ***12***
- ***Barcelona TV (BTV)*** ***13***
- ***TeleMadrid*** ***14***
- ***13TV*** ***15***
- ***CANAL 33*** ***16***
- ***TVG (Galicia)*** ***17***
- ***ETB 1*** ***18***
- ***ETB 2*** ***19***
- ***ETB (sin especificar canal)*** ***20***
- ***Televisión de Castilla-La Mancha (TVCM)*** ***21***
- ***TVC (Canarias)*** ***22***
- ***Aragón TV*** ***23***
- ***IB3 (Baleares)*** ***24***
- ***7 Televisión Región Murcia*** ***25***
- ***Canal Extremadura*** ***26***
- ***TPA (Asturias)*** ***27***
- ***TVCYL (Castilla-León)*** ***28***
- ***Intereconomia TV*** ***29***
- ***NAFAR TELEBISTA (NTB)*** ***30***
- ***Á PUNT*** ***31***
- ***Otras televisiones autonómicas*** ***60***
- ***Televisiones insulares*** ***61***
- ***Televisiones locales*** ***62***
- ***Otras cadenas de televisión*** ***94***
- ***Cualquiera sin especificar*** ***95***
- ***Otras (ESPECIFICAR)*** ***96***
- ***Ninguno*** ***97***
- ***No recuerda*** ***98***
- ***N.C*** ***99***

Saltos:
Si P2601=95 O P2601=97 O P2601=98 O P2601=99 ir a [P2701]

Filtros:
Si NO P2601=(96) ir a la siguiente.

[P2601_COD_]

Saltos:
Si P2601=95 O P2601=97 O P2601=98 O P2601=99 ir a [P2701] -

Estudio: Efectos y consec. del Coronavirus (IV)
Clave: ECIS3324

[P2602]

TVE1 ..... ***1***
La 2 ..... ***2***
Antena 3 ..... ***3***
La Cuatro ***4***
Telecinco ***5***
La Sexta ***6***
Canal 24 horas ***7***
TVE (Sin especificar canal) ***8***
Canal Sur ***9***
TV3 ..... ***10***
8TV ***11***
3/24 ..... ***12***
Barcelona TV (BTV) ***13***
TeleMadrid ***14***
13TV ***15***
CANAL 33 ..... ***16***
TVG (Galicia) ***17***
ETB 1 ..... ***18***
ETB 2 ..... ***19***
ETB (sin especificar canal) ***20***
Televisión de Castilla-La Mancha (TVCM) ***21***
TVC (Canarias) ***22***
Aragón TV ***23***
IB3 (Baleares) ***24***
7 Televisión Región Murcia ***25***
Canal Extremadura ***26***
TPA (Asturias) ***27***
TVCYL (Castilla-León) ***28***
Intereconomia TV ***29***
NAFAR TELEBISTA (NTB) ***30***
Á PUNT ***31***
Otras televisiones autonómicas ***60***
Televisiones insulares ***61***
Televisiones locales ***62***
Otras cadenas de televisión ***94***
Cualquiera sin especificar ***95***
Otra (ESPECIFICAR) ***96***
Ninguno más ***97***
No recuerda ***98***
N.C. ***99***

Saltos:
Si P2602=95 O P2602=97 O P2602=98 O P2602=99 ir a [P2701]
Filtros:
Si NO P2602=(96) ir a la siguiente.

[P2602_COD]

Saltos:
Si P2602=95 O P2602=97 O P2602=98 O P2602=99 ir a [P2701] -

[P2603]

TVE1 ..... ***1***
La 2 ..... ***2***
Antena 3 ..... ***3***
La Cuatro ***4***
Telecinco ***5***
La Sexta ***6***
Canal 24 horas ***7***
TVE (Sin especificar canal) ***8***
Canal Sur ***9***
TV3 ..... ***10***
8TV ***11***
3/24 ..... ***12***
Barcelona TV (BTV) ***13***
TeleMadrid ***14***
13TV ***15***
CANAL 33 ..... ***16***
TVG (Galicia) ***17***
ETB 1 ..... ***18***
ETB 2 ..... ***19***
ETB (sin especificar canal) ***20***
Televisión de Castilla-La Mancha (TVCM) ***21***
TVC (Canarias) ***22***
Aragón TV ***23***
IB3 (Baleares) ***24***
7 Televisión Región Murcia ***25***
Canal Extremadura ***26***
TPA (Asturias) ***27***
TVCYL (Castilla-León) ***28***
Intereconomia TV ***29***
NAFAR TELEBISTA (NTB) ***30***
Á PUNT ***31***
Otras televisiones autonómicas ***60***
Televisiones insulares ***61***
Televisiones locales ***62***
Otras cadenas de televisión ***94***
Cualquiera sin especificar ***95***
Otra (ESPECIFICAR) ***96***
Ninguno más ***97***
No recuerda ***98***
N.C. ***99***

Filtros:
Si NO P2603=(96) ir a la siguiente.

[P2603_COD]

Si la respuesta es "Otra" (ESPECIFICAR)
Si la respuesta es "Otra" (ESPECIFICAR)
Si la respuesta es "Otra" (ESPECIFICAR)

Estudio: Efectos y consec. del Coronavirus (IV)
Clave: ECIS3324

P.27 ¿Y en el último mes? (ENTREVISTADOR/A: RESPUESTA ESPONTÁNEA. SELECCIONAR EN LA LISTA DESPLEGABLE COMO MÁXIMO TRES, DEL MÁS AL MENOS VISTO. TAMBIÉN PUEDE SELECCIONAR EN LA LISTA "NO RECUERDA" Y "NO CONTESTA" O "NINGUNO" Y "NINGUNO MÁS")

[P2701]

- *TVE1 ..... 1*
- *La 2 ..... 2*
- *Antena 3 ..... 3*
- *La Cuatro 4*
- *Telecinco 5*
- *La Sexta 6*
- *Canal 24 horas 7*
- *TVE (Sin especificar canal) 8*
- *Canal Sur 9*
- *TV3 ..... 10*
- *8TV 11*
- *3/24 ..... 12*
- *Barcelona TV (BTV) 13*
- *TeleMadrid 14*
- *13TV 15*
- *CANAL 33 ..... 16*
- *TVG (Galicia) 17*
- *ETB 1 ..... 18*
- *ETB 2 ..... 19*
- *ETB (sin especificar canal) 20*
- *Televisión de Castilla-La Mancha (TVCM) 21*
- *TVC (Canarias) 22*
- *Aragón TV 23*
- *IB3 (Baleares) 24*
- *7 Televisión Región Murcia 25*
- *Canal Extremadura 26*
- *TPA (Asturias) 27*
- *TVCYL (Castilla-León) 28*
- *Intereconomia TV 29*
- *NAFAR TELEBISTA (NTB) 30*
- *Á PUNT 31*
- *Otras televisiones autonómicas 60*
- *Televisiones insulares 61*
- *Televisiones locales 62*
- *Otras cadenas de televisión 94*
- *Cualquiera sin especificar 95*
- *Otra (ESPECIFICAR) 96*
- *Ninguno 97*
- *No recuerda 98*
- *N.C. 99*

Saltos:
Si P2701=95 O P2701=97 O P2701=98 O P2701=99 ir a [P28] - Lectura de periódico en papel durante el último mes

Filtros:
Si NO P2701=(96) ir a la siguiente.

[P2701_COD]

Saltos:
Si P2701=95 O P2701=97 O P2701=98 O P2701=99 ir a [P28] - Lectura de periódico en papel durante el último mes

[P2702]

- *TVE1 ..... 1*
- *La 2 ..... 2*
- *Antena 3 ..... 3*
- *La Cuatro 4*
- *Telecinco 5*
- *La Sexta 6*
- *Canal 24 horas 7*
- *TVE (Sin especificar canal) 8*
- *Canal Sur 9*
- *TV3 ..... 10*
- *8TV 11*
- *3/24 ..... 12*
- *Barcelona TV (BTV) 13*
- *TeleMadrid 14*
- *13TV 15*
- *CANAL 33 ..... 16*
- *TVG (Galicia) 17*
- *ETB 1 ..... 18*
- *ETB 2 ..... 19*
- *ETB (sin especificar canal) 20*
- *Televisión de Castilla-La Mancha (TVCM) 21*
- *TVC (Canarias) 22*
- *Aragón TV 23*
- *IB3 (Baleares) 24*
- *7 Televisión Región Murcia 25*
- *Canal Extremadura 26*
- *TPA (Asturias) 27*
- *TVCYL (Castilla-León) 28*
- *Intereconomia TV 29*
- *NAFAR TELEBISTA (NTB) 30*
- *Á PUNT 31*
- *Otras televisiones autonómicas 60*
- *Televisiones insulares 61*
- *Televisiones locales 62*
- *Otras cadenas de televisión 94*
- *Cualquiera sin especificar 95*
- *Otra (ESPECIFICAR) 96*
- *Ninguno más 97*
- *No recuerda 98*
- *N.C. 99*

Saltos:
Si P2702=95 O P2702=97 O P2702=98 O P2702=99 ir a [P28] -

Filtros:
Si NO P2702=(96) ir a la siguiente.

[P2702_COD]

Saltos:
Si P2702=95 O P2702=97 O P2702=98 O P2702=99 ir a [P28] - Lectura de periódico en papel durante el último mes

Estudio: Efectos y consec. del Coronavirus (IV)
Clave: ECIS3324

[P2703]

- *TVE1* 1
- *La 2* 2
- *Antena 3* 3
- *La Cuatro* 4
- *Telecinco* 5
- *La Sexta* 6
- *Canal 24 horas* 7
- *TVE (Sin especificar canal)* 8
- *Canal Sur* 9
- *TV3* 10
- *8TV* 11
- *3/24* 12
- *Barcelona TV (BTV)* 13
- *TeleMadrid* 14
- *13TV* 15
- *CANAL 33* 16
- *TVG (Galicia)* 17
- *ETB 1* 18
- *ETB 2* 19
- *ETB (sin especificar canal)* 20
- *Televisión de Castilla-La Mancha (TVCM)* 21
- *TVC (Canarias)* 22
- *Aragón TV* 23
- *IB3 (Baleares)* 24
- *7 Televisión Región Murcia* 25
- *Canal Extremadura* 26
- *TPA (Asturias)* 27
- *TVCYL (Castilla-León)* 28
- *Intereconomia TV* 29
- *NAFAR TELEBISTA (NTB)* 30
- *Á PUNT* 31
- *Otras televisiones autonómicas* 60
- *Televisiones insulares* 61
- *Televisiones locales* 62
- *Otras cadenas de televisión* 94
- *Cualquiera sin especificar* 95
- *Otra (ESPECIFICAR)* 96
- *Ninguno más* 97
- *No recuerda* 98
- *N.C.* 99

Filtros:
Si NO P2703=(96) ir a la siguiente.

[P2703_COD]

Si la respuesta es "Otra" (ESPECIFICAR)
Si la respuesta es "Otra" (ESPECIFICAR)
Si la respuesta es "Otra" (ESPECIFICAR)

P.28 ¿Ha leído Ud. algún periódico en papel durante el último mes?

[P28]

- *Sí* 1
- *No* 2
- *N.C.* 9

P.28A¿Y concretamente durante la última semana?

[P28A]

- *Sí* 1
- *No* 2
- *N.C.* 9

Saltos:
Si (P28=2 O P28=9) Y (P28A=2 O P28A=9) ir a [P29] - Lectura de periódico en formato digital en el último mes

P. 28B ¿Cuál o cuáles? (ENTREVISTADOR/A: RESPUESTA ESPONTÁNEA. SELECCIONAR EN LA LISTA DESPLEGABLE, COMO MÁXIMO TRES PERIÓDICOS, DEL MÁS AL MENOS LEÍDO. TAMBIÉN PUEDE SELECCIONAR EN LA LISTA "NO RECUERDA" Y "NO CONTESTA" O "NINGUNO MÁS")

[P28B01]

- *EL PAÍS* 1
- *EL MUNDO* 2
- *ABC* 3
- *LA RAZÓN* 4
- *LA GACETA* 5
- *IDEAL (GRANADA, JAEN, ALMERIA)* 6
- *DIARIO SUR (MALAGA)* 7
- *DIARIO DE CÁDIZ* 8
- *DIARIO DE SEVILLA* 9
- *CÓRDOBA* 10
- *EL CORREO DE ANDALUCÍA* 11
- *LA VOZ DE CÁDIZ* 12
- *LA VOZ DE ALMERÍA* 13
- *DIARIO DE JEREZ* 14
- *LA OPINIÓN DE MÁLAGA* 15
- *DIARIO DE JAÉN* 16
- *HUELVA INFORMACIÓN* 17
- *GRANADA HOY* 18
- *EL DÍA DE CÓRDOBA* 19
- *HERALDO DE ARAGÓN* 20
- *EL PERIÓDICO DE ARAGÓN* 21
- *DIARIO DEL ALTO ARAGÓN* 22
- *LA NUEVA ESPAÑA* 23
- *EL COMERCIO* 24
- *LA VOZ DE ASTURIAS* 25
- *DIARIO DE MALLORCA* 26
- *ÚLTIMA HORA* 27
- *DIARIO BALEARES* 28
- *EL DIA DE CANARIAS* 29
- *CANARIAS 7* 30
- *LA PROVINCIA* 31
- *EL DIARIO MONTAÑÉS* 32
- *EL ALERTA (Cantabria)* 33
- *EL NORTE DE CASTILLA* 34
- *DIARIO DE LEÓN* 35
- *LA GACETA REGIONAL DE SALAMANCA* 36
- *DIARIO DE BURGOS* 37
- *LA OPINIÓN-EL CORREO DE ZAMORA* 38
- *EL ADELANTO DE SALAMANCA* 39
- *DIARIO PALENTINO* 40
- *DIARIO DE ÁVILA* 41
- *EL ADELANTADO DE SEGOVIA* 42
- *HERALDO DE SORIA* 43
- *LA TRIBUNA DE CIUDAD REAL, TOLEDO Y TALAVERA* 44
- *EL DÍA (DE CASTILLA-LA MANCHA)* 45
- *LA TRIBUNA DE ALBACETE* 46
- *EL DIA DE CUENCA* 47
- *LA REGIÓN (ORENSE)* 48
- *ARA* 49
- *EL PUNT-AVUI* 50
- *LA VANGUARDIA* 51
- *EL PERIÓDICO de CATALUÑA* 52
- *DIARI DE TARRAGONA* 53
- *DIARI DE GIRONA* 54
- *DIARI DE TERRASSA* 55
- *DIARI DE SABADELL* 56
- *HOY-DIARIO DE EXTREMADURA* 57
- *PERIÓDICO DE EXTREMADURA* 58
- *LEVANTE* 59
- *LAS PROVINCIAS* 60
- *INFORMACIÓN DE ALICANTE* 61
- *MEDITERRANEO* 62
- *LA VOZ DE GALICIA* 63
- *FARO DE VIGO* 64
- *EL CORREO GALLEGO* 65
- *FARO DE OURENSE* 66

Estudio: Efectos y consec. del Coronavirus (IV)
Clave: ECIS3324

DIARIO DE PONTEVEDRA *67*
EL PROGRESO (LUGO) *68*
DIARIO DE NAVARRA *69*
LA VERDAD *70*
LA OPINIÓN DE MURCIA *71*
BERRIA *72*
EL CORREO *73*
EL DIARIO VASCO *74*
GARA *76*
DEIA *77*
DIARIO DE NOTICIAS (ALAVA, NAVARRA) *78*
LA RIOJA *79*
EL PUEBLO DE CEUTA *80*
EL FARO- CEUTA Y MELILLA *81*
EL DIARIO AVISOS *82*
EXPANSIÓN *89*
CINCO DÍAS *90*
20 MINUTOS *91*
OTROS LOCALES *94*
CUALQUIERA *95*
OTROS *96*
NO CONTESTA *99*

Saltos:
Si P28B01=97 O P28B01=99 ir a [P29]

Filtros:
Si NO P28B01=(96) ir a la siguiente.

[P28B01_COD]

Saltos:
Si P28B01=97 O P28B01=99 ir a [P29]

[P28B02]

EL PAÍS *1*
EL MUNDO *2*
ABC *3*
LA RAZÓN *4*
LA GACETA *5*
IDEAL (GRANADA, JAEN, ALMERIA) *6*
DIARIO SUR (MALAGA) *7*
DIARIO DE CÁDIZ *8*
DIARIO DE SEVILLA *9*
CÓRDOBA *10*
EL CORREO DE ANDALUCÍA *11*
LA VOZ DE CÁDIZ *12*
LA VOZ DE ALMERÍA *13*
DIARIO DE JEREZ *14*
LA OPINIÓN DE MÁLAGA *15*
DIARIO DE JAÉN *16*
HUELVA INFORMACIÓN *17*
GRANADA HOY *18*
EL DÍA DE CÓRDOBA *19*
HERALDO DE ARAGÓN *20*
EL PERIÓDICO DE ARAGÓN *21*
DIARIO DEL ALTO ARAGÓN *22*
LA NUEVA ESPAÑA *23*
EL COMERCIO *24*
LA VOZ DE ASTURIAS *25*
DIARIO DE MALLORCA *26*
ÚLTIMA HORA *27*
DIARIO BALEARES *28*
EL DIA DE CANARIAS *29*
CANARIAS 7 ..... *30*
LA PROVINCIA *31*
EL DIARIO MONTAÑÉS *32*
EL ALERTA (Cantabria) *33*
EL NORTE DE CASTILLA *34*
DIARIO DE LEÓN *35*
LA GACETA REGIONAL DE SALAMANCA *36*
DIARIO DE BURGOS *37*
LA OPINIÓN-EL CORREO DE ZAMORA *38*
EL ADELANTO DE SALAMANCA *39*
DIARIO PALENTINO *40*
DIARIO DE ÁVILA *41*
EL ADELANTADO DE SEGOVIA *42*
HERALDO DE SORIA *43*
LA TRIBUNA DE CIUDAD REAL, TOLEDO Y TALAVERA *44*
EL DÍA (DE CASTILLA-LA MANCHA) *45*
LA TRIBUNA DE ALBACETE *46*
EL DIA DE CUENCA *47*
LA REGIÓN (ORENSE) *48*
ARA *49*
EL PUNT-AVUI *50*
LA VANGUARDIA *51*
EL PERIÓDICO de CATALUÑA *52*
DIARI DE TARRAGONA *53*
DIARI DE GIRONA *54*
DIARI DE TERRASSA *55*
DIARI DE SABADELL *56*
HOY-DIARIO DE EXTREMADURA *57*
PERIÓDICO DE EXTREMADURA *58*
LEVANTE *59*
LAS PROVINCIAS *60*
INFORMACIÓN DE ALICANTE *61*
MEDITERRANEO *62*
LA VOZ DE GALICIA *63*
FARO DE VIGO *64*
EL CORREO GALLEGO *65*
FARO DE OURENSE *66*
DIARIO DE PONTEVEDRA *67*
EL PROGRESO (LUGO) *68*
DIARIO DE NAVARRA *69*
LA VERDAD *70*
LA OPINIÓN DE MURCIA *71*
BERRIA *72*
EL CORREO *73*

Estudio: Efectos y consec. del Coronavirus (IV)
Clave: ECIS3324

EL DIARIO VASCO *74*
GARA *76*
DEIA *77*
DIARIO DE NOTICIAS (ALAVA, NAVARRA) *78*
LA RIOJA *79*
EL PUEBLO DE CEUTA *80*
EL FARO- CEUTA Y MELILLA *81*
EL DIARIO AVISOS *82*
EXPANSIÓN *89*
CINCO DÍAS *90*
20 MINUTOS *91*
OTROS LOCALES *94*
CUALQUIERA *95*
OTRO *96*
NINGUNO MÁS *97*
NO CONTESTA *99*

Saltos:
Si P28B02=95 O P28B02=97 O P28B02=99 ir a [P29] - Lectura de periódico en formato digital en el último mes

Filtros:
Si NO P28B02=(96) ir a la siguiente.

[P28B02_COD]

Saltos:
Si P28B02=95 O P28B02=97 O P28B02=99 ir a [P29]

[P28B03]

EL PAÍS *1*
EL MUNDO *2*
ABC *3*
LA RAZÓN *4*
LA GACETA *5*
IDEAL (GRANADA, JAEN, ALMERIA) *6*
DIARIO SUR (MALAGA) *7*
DIARIO DE CÁDIZ *8*
DIARIO DE SEVILLA *9*
CÓRDOBA *10*
EL CORREO DE ANDALUCÍA *11*
LA VOZ DE CÁDIZ *12*
LA VOZ DE ALMERÍA *13*
DIARIO DE JEREZ *14*
LA OPINIÓN DE MÁLAGA *15*
DIARIO DE JAÉN *16*
HUELVA INFORMACIÓN *17*
GRANADA HOY *18*
EL DÍA DE CÓRDOBA *19*
HERALDO DE ARAGÓN *20*
EL PERIÓDICO DE ARAGÓN *21*
DIARIO DEL ALTO ARAGÓN *22*
LA NUEVA ESPAÑA *23*
EL COMERCIO *24*
LA VOZ DE ASTURIAS *25*
DIARIO DE MALLORCA *26*
ÚLTIMA HORA *27*
DIARIO BALEARES *28*
EL DIA DE CANARIAS *29*
CANARIAS 7 ..... *30*
LA PROVINCIA *31*
EL DIARIO MONTAÑÉS *32*
EL ALERTA (Cantabria) *33*
EL NORTE DE CASTILLA *34*
DIARIO DE LEÓN *35*
LA GACETA REGIONAL DE SALAMANCA *36*
DIARIO DE BURGOS *37*
LA OPINIÓN-EL CORREO DE ZAMORA *38*
EL ADELANTO DE SALAMANCA *39*
DIARIO PALENTINO *40*
DIARIO DE ÁVILA *41*
EL ADELANTADO DE SEGOVIA *42*
HERALDO DE SORIA *43*
LA TRIBUNA DE CIUDAD REAL, TOLEDO Y TALAVERA *44*
EL DÍA (DE CASTILLA-LA MANCHA) *45*
LA TRIBUNA DE ALBACETE *46*
EL DIA DE CUENCA *47*
LA REGIÓN (ORENSE) *48*
ARA *49*
EL PUNT-AVUI *50*
LA VANGUARDIA *51*
EL PERIÓDICO de CATALUÑA *52*
DIARI DE TARRAGONA *53*
DIARI DE GIRONA *54*
DIARI DE TERRASSA *55*
DIARI DE SABADELL *56*
HOY-DIARIO DE EXTREMADURA *57*
PERIÓDICO DE EXTREMADURA *58*
LEVANTE *59*
LAS PROVINCIAS *60*
INFORMACIÓN DE ALICANTE *61*
MEDITERRANEO *62*
LA VOZ DE GALICIA *63*
FARO DE VIGO *64*
EL CORREO GALLEGO *65*
FARO DE OURENSE *66*
DIARIO DE PONTEVEDRA *67*
EL PROGRESO (LUGO) *68*
DIARIO DE NAVARRA *69*
LA VERDAD *70*
LA OPINIÓN DE MURCIA *71*
BERRIA *72*
EL CORREO *73*

Estudio: Efectos y consec. del Coronavirus (IV)
Clave: ECIS3324

EL DIARIO VASCO 74
GARA 76
DEIA 77
DIARIO DE NOTICIAS (ALAVA, NAVARRA) 78
LA RIOJA 79
EL PUEBLO DE CEUTA 80
EL FARO- CEUTA Y MELILLA 81
EL DIARIO AVISOS 82
EXPANSIÓN 89
CINCO DÍAS 90
20 MINUTOS 91
OTROS LOCALES 94
CUALQUIERA 95
OTRO 96
NINGUNO MÁS 97
NO CONTESTA 99

Filtros:
Si NO P28B03=(96) ir a la siguiente.
[P28B03_COD]

Si la respuesta es "Otro" ESPECIFICAR
Si la respuesta es "Otro" ESPECIFICAR
Si la respuesta es "Otro" ESPECIFICAR

P.29 ¿Ha leído Ud. algún periódico en formato digital durante el último mes?
[P29]

Sí 1
No 2
N.C. 9

P.29A ¿Y concretamente durante la última semana?
[P29A]

Sí 1
No 2
N.C. 9

Saltos:
Si (P29=2 O P29=9) Y (P29A=2 O P29A=9) ir a [P30_1] -

P.29B ¿Cuál o cuáles? (ENTREVISTADOR/A: RESPUESTA ESPONTÁNEA. SELECCIONAR EN LA LISTA DESPLEGABLE COMO MÁXIMO TRES PERIÓDICOS, DEL MÁS AL MENOS LEÍDO. TAMBIÉN PUEDE SELECCIONAR EN LA LISTA "NO RECUERDA" Y "NO CONTESTA" O "NINGUNO MÁS")
[P29B01]

EL PAÍS 1
EL MUNDO 2
ABC 3
LA RAZÓN 4
LA GACETA 5
IDEAL (GRANADA, JAEN, ALMERIA) 6
DIARIO SUR (MALAGA) 7
DIARIO DE CÁDIZ 8
DIARIO DE SEVILLA 9
CÓRDOBA 10
EL CORREO DE ANDALUCÍA 11
LA VOZ DE CÁDIZ 12
LA VOZ DE ALMERÍA 13
DIARIO DE JEREZ 14
LA OPINIÓN DE MÁLAGA 15
DIARIO DE JAÉN 16
HUELVA INFORMACIÓN 17
GRANADA HOY 18
EL DÍA DE CÓRDOBA 19
HERALDO DE ARAGÓN 20
EL PERIÓDICO DE ARAGÓN 21
DIARIO DEL ALTO ARAGÓN 22
LA NUEVA ESPAÑA 23
EL COMERCIO 24
LA VOZ DE ASTURIAS 25
DIARIO DE MALLORCA 26
ÚLTIMA HORA 27
DIARIO BALEARES 28
EL DIA DE CANARIAS 29
CANARIAS 7 30
LA PROVINCIA 31
EL DIARIO MONTAÑÉS 32
EL ALERTA (Cantabria) 33
EL NORTE DE CASTILLA 34
DIARIO DE LEÓN 35
LA GACETA REGIONAL DE SALAMANCA 36
DIARIO DE BURGOS 37
LA OPINIÓN-EL CORREO DE ZAMORA 38
EL ADELANTO DE SALAMANCA 39
DIARIO PALENTINO 40
DIARIO DE ÁVILA 41
EL ADELANTADO DE SEGOVIA 42
HERALDO DE SORIA 43
LA TRIBUNA DE CIUDAD REAL, TOLEDO Y TALAVERA 44
EL DÍA (DE CASTILLA-LA MANCHA) 45
LA TRIBUNA DE ALBACETE 46
EL DIA DE CUENCA 47
LA REGIÓN (ORENSE) 48
ARA 49
EL PUNT-AVUI 50
LA VANGUARDIA 51
EL PERIÓDICO de CATALUÑA 52
DIARI DE TARRAGONA 53
DIARI DE GIRONA 54
DIARI DE TERRASSA 55
DIARI DE SABADELL 56
HOY-DIARIO DE EXTREMADURA 57
PERIÓDICO DE EXTREMADURA 58
LEVANTE 59
LAS PROVINCIAS 60
INFORMACIÓN DE ALICANTE 61
MEDITERRANEO 62
LA VOZ DE GALICIA 63
FARO DE VIGO 64
EL CORREO GALLEGO 65
FARO DE OURENSE 66

Estudio: Efectos y consec. del Coronavirus (IV)
Clave: ECIS3324

DIARIO DE PONTEVEDRA....67
EL PROGRESO (LUGO)....68
DIARIO DE NAVARRA....69
LA VERDAD....70
LA OPINIÓN DE MURCIA....71
BERRIA....72
EL CORREO....73
EL DIARIO VASCO....74
EL IMPARCIAL.ES....75
GARA....76
DEIA....77
DIARIO DE NOTICIAS (ALAVA, NAVARRA)....78
LA RIOJA....79
EL PUEBLO DE CEUTA....80
EL FARO- CEUTA Y MELILLA....81
EL DIARIO AVISOS....82
ELNACIONAL.CAT....83
PÚBLICO.ES....84
EL DIARIO.ES....85
OK DIARIO.ES....86
PERIÓDICOS DIGITALES....87
INFOLIBRE....88
EXPANSIÓN....89
CINCO DÍAS....90
20 MINUTOS....91
EL CONFIDENCIAL.COM....92
ELESPAÑOL.COM....93
OTROS LOCALES....94
CUALQUIERA....95
OTROS....96
NO RECUERDA....98
NO CONTESTA....99

Saltos:
Si P29B01=97 O P29B01=98 O P29B01=99 ir a [P30_1]

Filtros:
Si NO P29B01=(96) ir a la siguiente.

[P29B01_COD]

Saltos:
Si P29B01=97 O P29B01=98 O P29B01=99 ir a [P30_1]

[P29B02]

EL PAÍS.... 1
EL MUNDO.... 2
ABC.... 3
LA RAZÓN.... 4
LA GACETA.... 5
IDEAL (GRANADA, JAEN, ALMERIA).... 6
DIARIO SUR (MALAGA).... 7
DIARIO DE CÁDIZ.... 8
DIARIO DE SEVILLA.... 9
CÓRDOBA....10
EL CORREO DE ANDALUCÍA....11
LA VOZ DE CÁDIZ....12
LA VOZ DE ALMERÍA....13
DIARIO DE JEREZ....14
LA OPINIÓN DE MÁLAGA....15
DIARIO DE JAÉN....16
HUELVA INFORMACIÓN....17
GRANADA HOY....18
EL DÍA DE CÓRDOBA....19
HERALDO DE ARAGÓN....20
EL PERIÓDICO DE ARAGÓN....21
DIARIO DEL ALTO ARAGÓN....22
LA NUEVA ESPAÑA....23
EL COMERCIO....24
LA VOZ DE ASTURIAS....25
DIARIO DE MALLORCA....26
ÚLTIMA HORA....27
DIARIO BALEARES....28
EL DIA DE CANARIAS....29
CANARIAS 7....30
LA PROVINCIA....31
EL DIARIO MONTAÑÉS....32
EL ALERTA (Cantabria)....33
EL NORTE DE CASTILLA....34
DIARIO DE LEÓN....35
LA GACETA REGIONAL DE SALAMANCA....36
DIARIO DE BURGOS....37
LA OPINIÓN-EL CORREO DE ZAMORA....38
EL ADELANTO DE SALAMANCA....39
DIARIO PALENTINO....40
DIARIO DE ÁVILA....41
EL ADELANTADO DE SEGOVIA....42
HERALDO DE SORIA....43
LA TRIBUNA DE CIUDAD REAL, TOLEDO Y TALAVERA....44
EL DÍA (DE CASTILLA-LA MANCHA)....45
LA TRIBUNA DE ALBACETE....46
EL DIA DE CUENCA....47
LA REGIÓN (ORENSE)....48
ARA....49
EL PUNT-AVUI....50
LA VANGUARDIA....51
EL PERIÓDICO de CATALUÑA....52
DIARI DE TARRAGONA....53
DIARI DE GIRONA....54
DIARI DE TERRASSA....55
DIARI DE SABADELL....56
HOY-DIARIO DE EXTREMADURA....57
PERIÓDICO DE EXTREMADURA....58
LEVANTE....59
LAS PROVINCIAS....60
INFORMACIÓN DE ALICANTE....61
MEDITERRANEO....62
LA VOZ DE GALICIA....63
FARO DE VIGO....64
EL CORREO GALLEGO....65
FARO DE OURENSE....66
DIARIO DE PONTEVEDRA....67
EL PROGRESO (LUGO)....68
DIARIO DE NAVARRA....69
LA VERDAD....70
LA OPINIÓN DE MURCIA....71
BERRIA....72
EL CORREO....73

Estudio: Efectos y consec. del Coronavirus (IV)
Clave: ECIS3324

EL DIARIO VASCO 74
EL IMPARCIAL.ES 75
GARA 76
DEIA 77
DIARIO DE NOTICIAS (ALAVA, NAVARRA) 78
LA RIOJA 79
EL PUEBLO DE CEUTA 80
EL FARO- CEUTA Y MELILLA 81
EL DIARIO AVISOS 82
ELNACIONAL.CAT 83
PÚBLICO.ES 84
EL DIARIO.ES 85
OK DIARIO.ES 86
PERIÓDICOS DIGITALES 87
INFOLIBRE 88
EXPANSIÓN 89
CINCO DÍAS 90
20 MINUTOS 91
EL CONFIDENCIAL.COM 92
ELESPAÑOL.COM 93
OTROS LOCALES 94
CUALQUIERA 95
OTROS 96
NINGUNO MÁS 97
NO RECUERDA 98
NO CONTESTA 99

Saltos:
Si P29B02=95 O P29B02=97 O P29B02=98 O P29B02=99 ir a [P30_1] -

Filtros:
Si NO P29B02=(96) ir a la siguiente.

[P29B02_COD]

Saltos:
Si P29B02=95 O P29B02=97 O P29B02=98 O P29B02=99 ir a [P30_1] -

[P29B03]

EL PAÍS 1
EL MUNDO 2
ABC 3
LA RAZÓN 4
LA GACETA 5
IDEAL (GRANADA, JAEN, ALMERIA) 6
DIARIO SUR (MALAGA) 7
DIARIO DE CÁDIZ 8
DIARIO DE SEVILLA 9
CÓRDOBA 10
EL CORREO DE ANDALUCÍA 11
LA VOZ DE CÁDIZ 12
LA VOZ DE ALMERÍA 13
DIARIO DE JEREZ 14
LA OPINIÓN DE MÁLAGA 15
DIARIO DE JAÉN 16
HUELVA INFORMACIÓN 17
GRANADA HOY 18
EL DÍA DE CÓRDOBA 19
HERALDO DE ARAGÓN 20
EL PERIÓDICO DE ARAGÓN 21
DIARIO DEL ALTO ARAGÓN 22
LA NUEVA ESPAÑA 23
EL COMERCIO 24
LA VOZ DE ASTURIAS 25
DIARIO DE MALLORCA 26
ÚLTIMA HORA 27
DIARIO BALEARES 28
EL DIA DE CANARIAS 29
CANARIAS 7 30
LA PROVINCIA 31
EL DIARIO MONTAÑÉS 32
EL ALERTA (Cantabria) 33
EL NORTE DE CASTILLA 34
DIARIO DE LEÓN 35
LA GACETA REGIONAL DE SALAMANCA 36
DIARIO DE BURGOS 37
LA OPINIÓN-EL CORREO DE ZAMORA 38
EL ADELANTO DE SALAMANCA 39
DIARIO PALENTINO 40
DIARIO DE ÁVILA 41
EL ADELANTADO DE SEGOVIA 42
HERALDO DE SORIA 43
LA TRIBUNA DE CIUDAD REAL, TOLEDO Y TALAVERA 44
EL DÍA (DE CASTILLA-LA MANCHA) 45
LA TRIBUNA DE ALBACETE 46
EL DIA DE CUENCA 47
LA REGIÓN (ORENSE) 48
ARA 49
EL PUNT-AVUI 50
LA VANGUARDIA 51
EL PERIÓDICO de CATALUÑA 52
DIARI DE TARRAGONA 53
DIARI DE GIRONA 54
DIARI DE TERRASSA 55
DIARI DE SABADELL 56
HOY-DIARIO DE EXTREMADURA 57
PERIÓDICO DE EXTREMADURA 58
LEVANTE 59
LAS PROVINCIAS 60
INFORMACIÓN DE ALICANTE 61
MEDITERRANEO 62
LA VOZ DE GALICIA 63
FARO DE VIGO 64
EL CORREO GALLEGO 65
FARO DE OURENSE 66
DIARIO DE PONTEVEDRA 67
EL PROGRESO (LUGO) 68
DIARIO DE NAVARRA 69
LA VERDAD 70
LA OPINIÓN DE MURCIA 71
BERRIA 72
EL CORREO 73

Estudio: Efectos y consec. del Coronavirus (IV)
Clave: ECIS3324

EL DIARIO VASCO 74
EL IMPARCIAL.ES 75
GARA 76
DEIA 77
DIARIO DE NOTICIAS (ALAVA, NAVARRA) 78
LA RIOJA 79
EL PUEBLO DE CEUTA 80
EL FARO- CEUTA Y MELILLA 81
EL DIARIO AVISOS 82
ELNACIONAL.CAT 83
PÚBLICO.ES 84
EL DIARIO.ES 85
OK DIARIO.ES 86
PERIÓDICOS DIGITALES 87
INFOLIBRE 88
EXPANSIÓN 89
CINCO DÍAS 90
20 MINUTOS 91
EL CONFIDENCIAL.COM 92
ELESPAÑOL.COM 93
OTROS LOCALES 94
CUALQUIERA 95
OTROS 96
NINGUNO MÁS 97
NO RECUERDA 98
NO CONTESTA 99

Filtros:
Si NO P29B03=(96) ir a la siguiente.
[P29B03_COD]

Si la respuesta es "Otro" (ESPECIFICAR)
Si la respuesta es "Otro" (ESPECIFICAR)
Si la respuesta es "Otro" (ESPECIFICAR)

P.30 ¿Ha sentido Ud. durante las dos últimas semanas miedo a…
[P30]

	Sí	*No*	*N.S.*	*N.C.*
…que Ud. pueda enfermar o a que se agrave alguna enfermedad que ya tenía o tiene?	1	2	8	9
…que escaseen los alimentos u otros productos de primera necesidad?	1	2	8	9
…que le ocurra algo grave (un accidente, una enfermedad, etc.) y tenga que ir a urgencias?	1	2	8	9
…no poder celebrar eventos importantes (un bautizo, una comunión, una boda, etc.)?	1	2	8	9
…estar aislado/a socialmente?	1	2	8	9
…no poder celebrar las Navidades?	1	2	8	9

P.31 ¿Cree Ud. que cuando alcancemos la inmunidad de grupo contra la COVID-19 a través de las vacunas, volverá Ud. a poder hacer todo lo que hacía antes de la pandemia?
[P31]

Sí 1
Al principio no 2
Definitivamente no 3
No lo sabe, duda 8
N.C. 9

Saltos:
Si P31=(1;8;9) ir a [ESCIDEOL]

P.31A ¿Podría decirme por qué piensa Ud. así?
[P31A]

Cree que tiene que pasar tiempo para la normalidad y para ver los efectos 1
No confía en la vacuna, en la rapidez con que se ha fabricado, en su eficacia 2
Cree que ha habido muchos cambios en todos los ámbitos y la vida no va a ser igual 3
Cree que tendría que estar toda la población vacunada, que sería necesaria la inmunidad total 4
Cree que tendremos que seguir tomando medidas 5
Otras respuestas 6
N.S. 8
N.C. 9

P.32 Cambiando de tema, cuando se habla de política se utilizan normalmente las expresiones izquierda y derecha. Situándonos en una escala de 10 casillas, como un termómetro, que van del 1 al 10, en la que 1 significa "lo más a la izquierda" y 10 "lo más a la derecha", ¿en qué casilla se colocaría Ud.?
[ESCIDEOL]

1 Izda. 1
2 2
3 3
4 4
5 5
6 6
7 7
8 8
9 9
10 Dcha. 10
N.S. 98
N.C. 99

P.33 ¿Me podría decir si en las elecciones generales del 10 de noviembre de 2019...? (LEER RESPUESTAS).
[PARTICIPACIONG]

Fue a votar y votó 1
No tenía edad para votar 2
Fue a votar pero no pudo hacerlo 3
No fue a votar porque no pudo 4
Prefirió no votar 5
No tenía derecho a voto 6
Votó por correo 7
No recuerda 8

Estudio: Efectos y consec. del Coronavirus (IV)
Clave: ECIS3324

P.33a ¿Y podría decirme a qué partido o coalición votó? (RESPUESTA ESPONTÁNEA).

Filtros:
Si NO PARTICIPACIONG=(1;7) ir a la siguiente.

[RECUVOTOG]

PSOE	*2*
PP	*1*
VOX	*18*
Unidas Podemos	*21*
En Comú Podem	*6*
En Común - Unidas Podemos	*67*
Ciudadanos	*4*
Más País	*50*
ERC	*8*
JxCat	*9*
CUP	*19*
EAJ-PNV	*11*
EH Bildu	*12*
CCa-PNC-NC	*13*
Navarra Suma (UPN)	*14*
Més Compromís	*7*
BNG (Bloque Nacionalista Galego)	*24*
PRC (Partido Regionalista de Cantabria)	*43*
Teruel Existe	*68*
PACMA (Partido Animalista)	*17*
Otros partidos	*95*
En blanco	*96*
Voto nulo	*77*
No recuerda	*98*
N.C.	*99*

P.34 ¿En qué situación laboral se encuentra Ud. actualmente?

[SITLAB]

Trabaja	*1*
Jubilado/a o pensionista (anteriormente ha trabajado)	*2*
Pensionista (anteriormente no ha trabajado)	*3*
En paro y ha trabajado antes	*4*
En paro y busca su primer empleo	*5*
Estudiante	*6*
Trabajo doméstico no remunerado	*7*
Otra situación	*8*
N.C.	*9*

P.35 ¿Me puede decir cuál es su ocupación actual?

Filtros:
Si NO SITLAB=1 ir a [ESCUELA] - Escolarización de la persona entrevistada

[CNO11]

Directores/as y gerentes	*1*
Profesionales y científicos/as e intelectuales	*2*
Técnicos/as y profesionales de nivel medio	*3*
Personal de apoyo administrativo	*4*
Trabajadores/as de los servicios y vendedores/as de comercios y mercados	*5*
Agricultores/as y trabajadores/as cualificados/as agropecuarios/as, forestales y pesqueros/as	*6*
Oficiales/as, operarios/as y artesanos/as de artes mecánicas y de otros oficios	*7*
Operadores/as de instalaciones y máquinas y ensambladores/as	*8*
Ocupaciones elementales	*9*
Ocupaciones militares y cuerpos policiales	*10*
Otra/o	*11*
N.C.	*99*

P.36 ¿Cuál es su situación de convivencia, es decir está Ud. viviendo…?

[SITCONVIVEN]

Solo/a	*1*
Solo/a con su/s hijo/a/s (con o sin otros/as parientes)	*2*
Con su marido o mujer o pareja con hijos/as (con o sin otros/as parientes o familiares)	*3*
Con su marido o mujer o pareja sin hijos/as (con o sin otros/as parientes o familiares)	*4*
Con su padre y/o madre con o sin hermanos/as (con o sin otros/as parientes o familiares)	*5*
Otra situación	*6*
N.C.	*9*

P.37 ¿Ha ido Ud. a la escuela o cursado algún tipo de estudios? (ENTREVISTADOR/A: en caso negativo, preguntar si sabe leer y escribir).

[ESCUELA]

No, es analfabeto/a	*1*
No, pero sabe leer y escribir	*2*
Sí, ha ido a la escuela	*3*
N.C.	*9*

Saltos:
Si NO ESCUELA=3 ir a [RELIGION]

P.37a ¿Cuáles son los estudios de más alto nivel oficial que Ud. ha cursado (con independencia de que los haya terminado o no)? Por favor, especifique lo más posible, diciéndome el curso en que estaba cuando los terminó (o los interrumpió) y también el nombre que tenían entonces esos estudios (ej: 3 años de estudios primarios, primaria, 5º de bachillerato, Maestría Industrial, preuniversitario, 4º de EGB, licenciatura, doctorado, FP1, etc.). (ENTREVISTADOR/A: si aún está estudiando, anotar el último curso que haya completado y el ciclo correcto en las opciones de respuesta. Si no ha completado la primaria, anotar nº de años que asistió a la escuela , diferenciando entre menos de 5 y más de 5).

[CURSOENTREV]

CURSO _____ *N.S. - N.R. = 98 N.C. = 99*

[NOMBREESTENTREV]

NOMBRE DE ESTUDIOS _____ *N.S. - N.R. = 98 N.C. = 99*

Estudio: Efectos y consec. del Coronavirus (IV)
Clave: ECIS3324

[NIVELESTENTREV]

- *01. Menos de 5 años de escolarización ... 1*
- *02. Educación primaria (Educación primaria de LOGSE, 5º Curso de EGB, Enseñanza primaria antigua) ... 2*
- *03. Cualificación profesional grado inicial (FP grado inicial). PCPI (Programas de Cualificación Profesional Inicial, que no precisan de titulación académica de la primera etapa de secundaria para su realización). Programas de garantía social ... 3*
- *04. Educación secundaria (ESO, EGB. Graduado Escolar. Certificado de Escolaridad, Bachillerato Elemental) ... 4*
- *05. FP de grado medio (Ciclo/módulo formativo de FP (grado medio), de Artes Plásticas y Diseño, Música y danza, Enseñanzas deportivas, FP I, Bachiller laboral elemental. Oficialía Industrial; Bachillerato Comercial). ... 5*
- *06. Bachillerato (Bachillerato LOGSE, BUP, Bachillerato superior (6º), Bachillerato universitario (7º), Incluidos COU y PREU) ... 6*
- *07. FP de grado superior (Ciclo/módulo formativo de FP (grado superior) de Artes Plásticas, Diseño, Música y danza, Deporte, FP II, Bach. Laboral Sup., Maestría industrial, Perito Mercantil; Secretariado de 2º grado; Grado Medio conservatorio) ... 7*
- *08. Arquitectura-ingeniería técnica (Arquitectura/ingeniería técnica, Aparejador; Peritos) ... 8*
- *09. Diplomatura (ATENCIÓN: solo Diplomaturas oficiales, no codificar aquí los tres primeros años de una licenciatura o grado con mayor duración). ... 9*
- *10. Grado (Estudios de grado, Enseñanzas Artísticas equivalentes (desde 2006)) ... 10*
- *11. Licenciatura (Titulaciones con equivalencia oficial: 2º ciclo INEF; Danza y arte dramático (desde 1992); Grado superior de música) ... 11*
- *12. Arquitectura/ingeniería ... 12*
- *13. Máster oficial universitario (Especialidades médicas o equivalente) ... 13*
- *14. Doctorado ... 14*
- *15. Títulos propios de posgrado (máster no oficial, etc.) ... 15*
- *16. Otros estudios ... 16*
- *N.S./No recuerda ... 98*
- *N.C. ... 99*

P.38 ¿Cómo se define Ud. en materia religiosa: católico/a practicante, católico/a no practicante, creyente de otra religión, agnóstico/a, indiferente o no creyente, o ateo/a?

[RELIGION]

- *Católico/a practicante ... 1*
- *Católico/a no practicante ... 2*
- *Creyente de otra religión ... 3*
- *Agnóstico/a (no niegan la existencia de Dios pero tampoco la descartan) ... 4*
- *Indiferente, no creyente ... 5*
- *Ateo/a (niegan la existencia de Dios) ... 6*
- *N.C. ... 9*

Saltos:
Si RELIGION=(4;5;6;9) ir a [CLASESOCIAL

P.38a ¿Con qué frecuencia asiste Ud. a misa u otros oficios religiosos, sin contar las ocasiones relacionadas con ceremonias de tipo social, por ejemplo, bodas, comuniones o funerales?

Filtros:
Si NO RELIGION=(1;2;3) ir a [CLASESOCIAL

[FRECUENCIARELIGION]

- *Nunca ... 1*
- *Casi nunca ... 2*
- *Varias veces al año ... 3*
- *Dos o tres veces al mes ... 4*
- *Todos los domingos y festivos ... 5*
- *Varias veces a la semana ... 6*
- *N.C. ... 9*

P.39 ¿A qué clase social diría Ud. que pertenece? (RESPUESTA ESPONTÁNEA).

[CLASESOCIAL]

- *Clase alta ... 1*
- *Clase media-alta ... 2*
- *Clase media-media ... 3*
- *Clase media-baja ... 4*
- *Clase trabajadora/obrera ... 5*
- *Clase baja ... 12*
- *Clase pobre ... 6*
- *Infraclase ... 7*
- *Proletariado ... 8*
- *A los/as de abajo ... 9*
- *Excluidos/as ... 10*
- *A la gente común ... 11*
- *Otra (especificar) ... 96*
- *No cree en las clases ... 97*
- *No sabe, duda ... 98*
- *N.C. ... 99*

Filtros:
Si NO CLASESOCIAL=(96) ir a la siguiente.

[CLASESOCIAL_COD]

_____

FIN DE LA ENTREVISTA.

MUCHAS GRACIAS POR SU AMABILIDAD Y POR EL TIEMPO QUE NOS HA DEDICADO.

Efectos y consecuencias del Coronavirus (V)

Encuesta realizada entre el 11 y el 30 de septiembre de 2021. En el ejemplo 7.5 se han seleccionado las preguntas P.5 y P.10.

Estas preguntas corresponden con la P.3 y P.10 de la encuesta Efectos y consecuencias del Coronavirus (I) y con las preguntas P.2 y P.3 de la encuesta Efectos y consecuencias del Coronavirus (VI).

Estudio: Efectos y consec. del Coronavirus Septiembre (V)
Clave: ECIS3336

«Información sujeta a secreto estadístico (Ley 12/89, de 9 de mayo, de la Función Estadística Pública) y al Reglamento General de Protección de Datos y la Ley Orgánica de Protección de Datos Personales y garantía de los derechos digitales.» Plan Estadístico Nacional 2021-2024. RD 1110/2020, de 15 de diciembre.

«Buenos días/tardes, soy (nombre propio) y estoy realizando una encuesta telefónica especial para el Centro de Investigaciones Sociológicas (CIS) sobre las consecuencias y efectos de la pandemia de la COVID-19. Por este motivo solicitamos su colaboración y se la agradecemos anticipadamente. Este teléfono ha sido obtenido al azar. Esta conversación será grabada para supervisar la calidad y después se borrará en un plazo inferior a un mes, le garantizamos el absoluto anonimato y secreto de sus respuestas en el más estricto cumplimiento de las leyes sobre secreto estadístico y protección de datos personales. Tras la realización de la encuesta su número de teléfono será disociado de las respuestas que pueda dar, que a su vez serán anonimizadas para que en ningún caso puedan ser asociadas a usted. Si desea conocer sus derechos de protección de datos y ampliar esta información puede consultar la página web www.cis.es ¿Ha comprendido la información leída? ¿Sería tan amable de contestar a unas preguntas? La encuesta dura unos 15 minutos. No está obligado a contestar todas las preguntas. Muchas gracias.»

PC1. Pregunta contacto 1. ¿Me puede decir a qué provincia y municipio estoy llamando...?

[TIPO_TEL]
FIJO ***1***
MÓVIL ***2***

[CCAA]
_____

[PROVINCIA]
_____

[MUNICIPIO]
_____

[CAPITAL]
_____

[TAMUNI]
_____

[ENTREV]

ENTREVISTADOR/A: SI LA PERSONA QUE CONTESTA ES DIFERENTE DE LA QUE COGIÓ EL TELÉFONO PRESENTARSE:

Buenos días/tardes, mi nombre es… y le llamo del Centro de Investigaciones Sociológicas porque estamos realizando una encuesta de opinión sobre temas de interés general. Dura unos 15 minutos. ¿Sería tan amable de colaborar con nosotros?

[SEXO]
Hombre ***1***
Mujer ***2***

[EDADEXACTA]
_____

[EDAD]
de 18 a 24 ..... ***1***
de 25 a 34 ..... ***2***
de 35 a 44 ..... ***3***
de 45 a 54 ..... ***4***
de 55 a 64 ..... ***5***
65 y más ***6***

SEXO Y EDAD - @1 @2

P.0 En primer lugar quisiera preguntarle si tiene Ud....

[P0]
La nacionalidad española ***1***
La nacionalidad española y otra ***2***
Otra nacionalidad ***3***

P.1 La situación del coronavirus que se está viviendo en España y en otros lugares ¿le preocupa a Ud. mucho, bastante, algo, nada o casi nada?

[P1]
Mucho ***1***
Bastante ***2***
(NO LEER) Regular ***3***
Algo ***4***
Nada o casi nada ***5***
N.S. ***8***
N.C. ***9***

P.2 ¿Ha tenido Ud. el coronavirus?
¿Y algún/a familiar?
¿Y algún/a amigo/a?
¿Y algún/a conocido/a?

[P2]

	Sí	*No*	*N.S.*	*N.C.*
Ud.	1	2	8	9
Algún/a familiar	1	2	8	9
Algún/a amigo/a	1	2	8	9
Algún/a conocido/a	1	2	8	9

P.3 Y alguna o algunas de estas personas (familiar no conviviente, y/o amigo/a, y/o conocido/a) que lo ha/n tenido, no ha/n podido superar la enfermedad y ha/n fallecido?

[P3]

Filtros:
Si Filtro por condiciones en filas ir a la siguiente.

	Sí	*No*	*N.S.*	*N.C.*
Algún/a familiar	1	2	8	9
Algún/a amigo/a	1	2	8	9
Algún/a conocido/a	1	2	8	9

Estudio: Efectos y consec. del Coronavirus Septiembre (V)
Clave: ECIS3336

P.4 ¿Diría Ud. que esta pandemia está cambiando mucho, bastante, algo, poco o no le está cambiando nada o casi nada su forma de vivir? ¿Y su forma de pensar? ¿Y la forma de cuidar de su salud? ¿Y sus hábitos sociales y comportamiento social?
[P4]

	Mucho	*Bastante*	*(NO LEER) Regular*	*Algo*	*Poco*	*Nada o casi nada*	*Está en duda, no lo sabría decir*	*N.C.*
Su forma de vivir	1	2	3	4	5	6	8	9
Su forma de pensar	1	2	3	4	5	6	8	9
La forma de cuidar de su salud	1	2	3	4	5	6	8	9
Sus hábitos sociales y de comportamiento social	1	2	3	4	5	6	8	9

P.4a ¿Y en qué aspectos principales está cambiando su forma de vivir?
P.4b ¿Y en qué aspectos su forma de pensar?
P.4c ¿Y la forma de cuidar de su salud?
P.4d ¿Y sus hábitos sociales y de comportamiento en la sociedad?
(ENTREVISTADOR/A: DOS RESPUESTAS. EN CADA CASILLA ANOTE TODO LO QUE MENCIONE LA PERSONA ENTREVISTADA)

Filtros:
Si P4_1=(5;6;8;9) ir a [P4B01] - Principales aspectos en que está cambiando su forma de pensar (1º)
[P4A01]

Su forma de vivir
N.S. = 98
N.C. = 99
[P4A02]

N.S. = 98
N.C. = 99
Su forma de pensar

Filtros:
Si P4_2=(5;6;8;9) ir a [P4C01] - Principales aspectos en que está cambiando su forma de cuidar de su salud (1º)
[P4B01]

N.S. = 98
N.C. = 99
[P4B02]

N.S. = 98
N.C. = 99
La forma de cuidar de su salud

Filtros:
Si P4_3=(5;6;8;9) ir a [P4D01] - Principales aspectos en que está cambiando sus hábitos y comportamientos (1º)
[P4C01]

N.S. = 98
N.C. = 99
[P4C02]

N.S. = 98
N.C. = 99
Sus hábitos sociales y comportamiento en la sociedad

Filtros:
Si P4_4=(5;6;8;9) ir a [P5_1] - Temores emocionales causados por la pandemia del covid-19 | Temor a enfermar?
[P4D01]

N.S. = 98
N.C. = 99
[P4D02]

N.S. = 98
N.C. = 99

P.5 La pandemia del covid-19 ha dado lugar a diversas situaciones que pueden afectar a la salud de las personas, ¿podría decirme si desde que se declaró la pandemia del coronavirus ha sentido usted…
[P5]

	Sí	*No*	*N.S.*	*N.C.*
Temor a enfermar?	1	2	8	9
Dolor por la pérdida de algún/a familiar, amigo/a o conocido/a?	1	2	8	9
Preocupación por haber perdido su empleo personal o el de algún/a familiar?	1	2	8	9
Inquietud por las medidas que pueden limitar los contactos y relaciones cara a cara con sus familiares, amigos/as y vecinos/as?	1	2	8	9
Miedo por la posibilidad de poder perder su empleo personal o el de algún/a familiar?	1	2	8	9
Intranquilidad por no poder afrontar sus gastos (hipotecas, alquileres, préstamos, suministros, telefonía, etc.)?	1	2	8	9
Miedo por no recuperar su vida tal como era antes de la pandemia?	1	2	8	9
Miedo por no poder emprender ya proyectos vitales como emanciparse, o abrir un negocio, o hacer algún viaje?	1	2	8	9
Inquietud y temor ante el futuro?	1	2	8	9

Estudio: Efectos y consec. del Coronavirus Septiembre (V)
Clave: ECIS3336

P.6 ¿Con qué frecuencia: siempre, la mayor parte del tiempo, a veces, o nunca, se ha sentido Ud. últimamente...
[P6]

	Siempre	*La mayor parte del tiempo*	*A veces*	*Nunca*	*N.S.*	*N.C.*
Especialmente tenso/a o ansioso/a?	1	2	3	4	8	9
Solo/a?	1	2	3	4	8	9
Deprimido/a?	1	2	3	4	8	9
Preocupado/a?	1	2	3	4	8	9
Enfadado/a?	1	2	3	4	8	9
Triste?	1	2	3	4	8	9

P.7 ¿Cree que la crisis del coronavirus ha tenido efectos en la salud emocional de las personas que la han sufrido directamente, como los/las sanitarios/as, ancianos/as, etc. o cree que ha tenido efectos en la salud emocional de todo tipo de personas?
[P7]

Ha tenido efectos en la salud emocional de las personas que la han sufrido directamente 1
Ha tenido efectos en la salud emocional de todo tipo de personas 2
N.S. 8
N.C. 9

P.8 Durante la pandemia algunas personas han cambiado costumbres o formas de pensar. ¿Diría Ud. que...?
[P8]

	Sí	*No*	*(NO LEER) Igual, lo mismo que antes*	*N.S.*	*N.C.*
...ha aprendido a organizar mejor su tiempo para no aburrirse?	1	2	3	8	9
...ha descubierto aficiones nuevas o actividades que nunca antes había realizado y que le gustan?	1	2	3	8	9
...se ha hecho más religioso/a o espiritual?	1	2	3	8	9
...ha cambiado sus valores y ahora valora y aprecia cosas que antes no?	1	2	3	8	9
...se ha interesado más por la gente que le importa, por si se encuentran bien física y emocionalmente?	1	2	3	8	9
...se ha interesado por el futuro más que antes?	1	2	3	8	9
...ha aprendido a valorar más las relaciones personales?	1	2	3	8	9
...ha aprendido a valorar más los beneficios de las actividades al aire libre?	1	2	3	8	9
...ha disfrutado más de actividades lúdicas con sus familiares (juegos, cocina, etc.)?	1	2	3	8	9

P.9 Me gustaría que valorase la relación con su familia, pareja, amigos/as, vecinos/as y personas de su entorno laboral y profesional en estos últimos meses. Diría Ud. que en estos últimos meses la relación con su familia,¿mejoró, empeoró o no cambió, sigue igual?
Y la relación con su pareja, ¿mejoró, empeoró, o no cambió, sigue igual?
Y la relación con sus vecinos/as, ¿mejoró, empeoró, o no cambió, sigue igual?
Y la relación con sus amigos/as, ¿mejoró, empeoró, o no cambió, sigue igual?
Y la relación con su entorno profesional, académico o laboral, ¿mejoró, empeoró,o no cambió, sigue igual?
[P9]

	Mejoró	*Empeoró*	*No cambió/Sigue igual*	*(NO LEER) No procede (No tiene familia, pareja, vecinos/as, amigos/as, etc.)*	*N.S./Duda*	*N.C.*
Relación con la familia	1	2	3	4	8	9
Relación con la pareja	1	2	3	4	8	9
Relación con los/as vecinos/as	1	2	3	4	8	9
Relación con los/as amigos/as	1	2	3	4	8	9
Relación con el entorno profesional, académico o laboral	1	2	3	4	8	9

Estudio: Efectos y consec. del Coronavirus Septiembre (V)
Clave: ECIS3336

P.10 A lo largo de estos meses de pandemia hay personas que han estado reflexionando o pensando sobre distintos aspectos de su vida. Me gustaría saber si Ud. personalmente en estos meses ha tomado decisiones o ha hecho propósitos para mejorar…

[P10]

	Sí	*No*	*(NO LEER) No Procede (No tiene)*	*N.C.*
Sus hábitos de alimentación	1	2	8	9
Su actividad física	1	2	8	9
Su salud	1	2	8	9
Su relación con la familia	1	2	8	9
Su relación con los/as vecinos/as	1	2	8	9
Su implicación en actividades de voluntariado y de ayuda comunitaria	1	2	8	9
Sus actividades de ocio	1	2	8	9
Sus amistades, sus relaciones sociales	1	2	8	9
Su trabajo, sus estudios, su actividad principal	1	2	8	9

P.11 ¿Ha hecho durante estos meses algún otro, u otros, propósitos de cambio?

[P11]

Sí ***1***
No ***2***
N.C. ***9***

Saltos:
Si NO P11=1 ir a [P12_1] -

P.11a ¿Cuál o cuáles? (ESPECIFICAR). MÁXIMO DOS RESPUESTAS. (RESPUESTA ESPONTÁNEA).

[P11A01]

Saltos:
Si P11A01="98" O P11A01="99" ir a [P12_1] -

N.S. = 98
N.C. = 99

[P11A02]

N.S. = 98
N.C. = 99

P.12 Durante estos meses de pandemia, dígame si ha realizado las siguientes actividades en su casa con más frecuencia de lo que lo hacía habitualmente antes de la pandemia.

[P12]

	Sí	*No*	*(NO LEER) No procede (No tiene)*	*N.S.*	*N.C.*
Ha utilizado más juegos de mesa con su familia, pareja o compañeros/as de piso	1	2	3	8	9
Se ha conectado más por videollamada con sus familiares o amigos/as	1	2	3	8	9
Se ha conectado por videoconferencia o webex más con sus compañeros/as de trabajo o de estudios, profesores/as, jefes/as, clientes/as, proveedores/as, etc.	1	2	3	8	9
Ha intercambiado más mensajes, fotos, vídeos, chistes en sus grupos de chats	1	2	3	8	9
Ha visto más series, películas, documentales o eventos deportivos, etc. (en TV, tablet, PC, móvil, etc.)	1	2	3	8	9
Ha leído más libros y revistas	1	2	3	8	9
Ha estado más pendiente de actividades vecinales y/o de voluntariado	1	2	3	8	9
Ha seguido más las noticias de los distintos medios de comunicación social	1	2	3	8	9
Ha estado más pendiente de sus redes sociales	1	2	3	8	9
Ha hecho más compras online	1	2	3	8	9
Ha realizado más actividades deportivas en casa	1	2	3	8	9
Ha dedicado más tiempo a las tareas del hogar, cocinar, ordenar armarios, etc.	1	2	3	8	9
Ha hecho más reparaciones y tareas de mantenimiento en su casa, como pintar, lijar, cuidar el jardín, etc.	1	2	3	8	9
Ha estado más pendiente de los miembros de su familia (contactando más por teléfono con su/s padre/madre, supervisando las tareas escolares de su/s hijos/as, etc.)	1	2	3	8	9
Ha descansado y dormido más	1	2	3	8	9
Ha hecho más teletrabajo	1	2	3	8	9

Estudio: Efectos y consec. del Coronavirus Septiembre (V)
Clave: ECIS3336

P.13 Y haciendo un balance general, la realización de estas tareas, ¿le ha hecho sentirse: mucho mejor, algo mejor, algo peor, mucho peor, en unas cosas mejor y en otras peor?

Filtros:
Si P13=6 ir a la siguiente.

[P13]

- *Mucho mejor 1*
- *Algo mejor 2*
- *(NO LEER) Regular 3*
- *Algo peor 4*
- *Mucho peor 5*
- *En unas cosas mejor y en otras peor 6*
- *(NO LEER) Igual 7*
- *N.S. 8*
- *N.C. 9*

P.13a ¿Por qué le han hecho sentirse mejor? (ENTREVISTADOR/A: MARCA EL PRINCIPAL MOTIVO. UNA RESPUESTA).

[P13A]

- *Por aprovechar el tiempo para realizar cosas y tareas pendientes 1*
- *Por sentirme útil, activo/a, realizado/a 2*
- *Porque me distraigo, entretengo, tengo la mente ocupada 3*
- *Por pasar así más tiempo con la familia y amigos/as 4*
- *Por valorar más otras cosas (la vida, el desarrollo personal, etc.) 5*
- *Por colaborar, ayudar, acompañar a gente que lo necesita 6*
- *Otras razones (Especificar) 96*
- *N.S. 98*
- *N.C. 99*

Saltos:
Si P13=(1;2) ir a [P14_1] - Sentimientos/comportamientos en la última semana | Se ha sentido feliz

Filtros:
Si NO P13A=(96) ir a la siguiente.

[P13A_COD]

P.13b ¿Por qué le han hecho sentirse peor? (ENTREVISTADOR/A: MARCA EL PRINCIPAL MOTIVO. UNA RESPUESTA).

[P13B]

- *Porque pensaba que estaba haciendo cosas impropias de mí, que era una pérdida de tiempo 1*
- *Por el malestar emocional causado por la situación 2*
- *Por estar encerrado/a, limitado/a, no poder salir libremente 3*
- *Por la falta de alguna actividad social 4*
- *Por el cansancio en general 5*
- *Por problemas de conciliación con el teletrabajo 6*
- *Otras razones (Especificar) 96*
- *N.S. 98*
- *N.C. 99*

Filtros:
Si NO P13B=(96) ir a la siguiente.

[P13B_COD]

P.14 A continuación le voy a leer una lista de sentimientos o comportamientos que quizá Ud. haya tenido durante la última semana. Por favor, dígame con qué frecuencia: todo o casi todo el tiempo, buena parte del tiempo, en algún momento, en ningún momento o casi en ningún momento, durante la última semana…

[P14]

	Todo o casi todo el tiempo	*Buena parte del tiempo*	*En algún momento*	*En ningún momento o casi en ningún momento*	*N.S.*	*N.C.*
Se ha sentido feliz	1	2	3	4	8	9
Se ha sentido deprimido/a	1	2	3	4	8	9
Se ha sentido tranquilo/a y relajado/a	1	2	3	4	8	9
Ha tenido la sensación de disfrutar de la vida	1	2	3	4	8	9
Se ha sentido lleno/a de energía y vitalidad	1	2	3	4	8	9
Se ha sentido solo/a	1	2	3	4	8	9
Se ha sentido realmente descansado/a al levantarse por las mañanas	1	2	3	4	8	9
Se ha sentido triste	1	2	3	4	8	9
Se ha sentido preocupado/a	1	2	3	4	8	9
Se ha sentido estresado/a	1	2	3	4	8	9

P.15 ¿Cuál es su estado civil?

[ECIVIL]

- *Casado/a 1*
- *Soltero/a 2*
- *Viudo/a 3*
- *Separado/a 4*
- *Divorciado/a 5*
- *N.C. 9*

Estudio: Efectos y consec. del Coronavirus Septiembre (V)
Clave: ECIS3336

P.16 Actualmente, ¿cuál es su situación de convivencia, es decir está Ud. viviendo…?
[SITCONVIVEN]
- *Solo/a* 1
- *Solo/a con su/s hijo/a/s (con o sin otros/as parientes)* 2
- *Con su marido o mujer o pareja con hijos/as (con o sin otros/as parientes o familiares)* 3
- *Con su marido o mujer o pareja sin hijos/as (con o sin otros/as parientes o familiares)* 4
- *Con su padre y/o madre con o sin hermanos/as (con o sin otros/as parientes o familiares)* 5
- *Otra situación* 6
- *N.C.* 9

P.17 ¿Ha cambiado Ud. su situación de convivencia o residencia debido a la pandemia?
[CAMBIOSITCONVIVEN]
- *Sí* 1
- *No* 2
- *N.C.* 9

P.17a ¿En qué sentido?

Filtros:
Si NO CAMBIOSITCONVIVEN=1 ir a la siguiente.

[P17A]
- *Cambio de casa* 1
- *Cambio de las personas con las que convive* 2
- *Otros (Especificar)* 96
- *N.C.* 99

Filtros:
Si NO P17A=(96) ir a la siguiente.

[P17A_COD]
_____

P.18 Durante el período de pandemia, desde que se declaró hasta ahora, ¿ha tenido que recurrir Ud. a alguna ayuda profesional debido a su estado de ánimo o situación emocional?
[P18]
- *Sí* 1
- *No* 2
- *N.S./Duda* 8
- *N.C.* 9

Saltos:
Si NO P18=1 ir a [P18B] -

P.18a ¿A qué tipo de profesional ha recurrido Ud.? (RESPUESTA MÚLTIPLE: ANOTAR TODAS LAS RESPUESTAS)
[P18A]
- *Al/la médico/a de cabecera* 1
- *Al/la psiquiatra* 2
- *Al/la psicólogo/a* 3
- *A un/a terapeuta* 4
- *A un/a masajista o fisioterapeuta* 5
- *A grupos de apoyo en la red* 6
- *A otros/as (Especificar)* 96
- *N.C.* 99

Filtros:
Si NO P18A=(96) ir a la siguiente.

[P18A_COD]
_____

P.18b Y durante los doce meses anteriores al período de pandemia, es decir antes de marzo de 2020, ¿recurrió en algún momento a alguna ayuda profesional debido a su estado de ánimo o situación emocional?
[P18B]
- *Sí* 1
- *No* 2
- *N.S./Duda* 8
- *N.C.* 9

Saltos:
Si NO P18B=1 ir a [P19] -

P.18c Y en concreto, durante los doce meses anteriores al período de pandemia, es decir antes de marzo de 2020, ¿acudió a un/a psicólogo/a y/o a un/a psiquiatra? (PUEDE MARCAR UNO O LOS DOS)
[P18C]
- *A un/a psicólogo/a* 1
- *A un/a psiquiatra* 2
- *Ninguno de los anteriores* 3
- *N.C.* 9

P.19 ¿Y alguna persona de su familia, sabe Ud. si ha tenido que recurrir a alguna ayuda profesional debido a su estado de ánimo o situación emocional?
[P19]
- *Sí* 1
- *No* 2
- *N.S./Duda* 8
- *N.C.* 9

Saltos:
Si NO P19=1 ir a [P19B]

P.19a ¿A qué tipo de profesional ha recurrido su familiar?
[P19A]
- *Al/la médico/a de cabecera* 1
- *Al/la psiquiatra* 2
- *Al/la psicólogo/a* 3
- *A un/a terapeuta* 4
- *A un/a masajista o fisioterapeuta* 5
- *A grupos de apoyo en la red* 6
- *A otros (Especificar)* 96
- *N.C.* 99

Filtros:
Si NO P19A=(96) ir a la siguiente.

[P19A_COD]
_____

P.19b ¿Y alguna persona de su familia sabe Ud. si recurrió en los doce meses anteriores a la pandemia, antes de marzo de 2020, a alguna ayuda profesional debido a su estado de ánimo o situación emocional?
[P19B]
- *Sí* 1
- *No* 2
- *N.S./Duda* 8
- *N.C.* 9

Saltos:
Si NO P19B=1 ir a [P20]

Estudio: Efectos y consec. del Coronavirus Septiembre (V)
Clave: ECIS3336

P.19c ¿Y alguna persona de su familia sabe Ud. si acudió en concreto, en los doce meses anteriores a la pandemia, a un/a psicólogo/a y/o a un/a psiquiatra? (PUEDE MARCAR UNO O LOS DOS)

[P19C]

- *A un/a psicólogo/a 1*
- *A un/a psiquiatra 2*
- *Ninguno de los anteriores 7*
- *No sabe, duda 8*
- *N.C. 9*

P.20 En general, ¿se siente Ud. muy optimista, algo optimista, poco optimista o nada o casi nada optimista sobre el futuro de España?

[P20]

- *Muy optimista 1*
- *Algo optimista 2*
- *Poco optimista 3*
- *Nada o casi nada optimista 4*
- *N.S./Duda 8*
- *N.C. 9*

P.21 Durante estos meses de la pandemia, ¿ha pensado Ud. en algún momento que ya no podrá realizar algunos de los proyectos, viajes o actividades que le hubiera gustado realizar?

[P21]

- *Sí, lo ha pensado 1*
- *No, no lo ha pensado 2*
- *N.S./Duda 8*
- *N.C. 9*

P.22 ¿Y ha pensado en algún momento que Ud. podría ser una de las víctimas mortales de esta pandemia?

[P22]

- *Sí, lo ha pensado 1*
- *No, no lo ha pensado 2*
- *N.S./Duda 8*
- *N.C. 9*

P.23 ¿Cree Ud. que cuando alcancemos la inmunidad de grupo contra la COVID-19 a través de las vacunas, volverá Ud. a poder hacer todo lo que hacía antes de la pandemia?

[P23]

- *Sí 1*
- *Al principio no 2*
- *Definitivamente no 3*
- *No lo sabe, duda 8*
- *N.C. 9*

P.23a ¿Podría decirme por qué piensa Ud. así?

Filtros:
Si P23=1 ir a la siguiente.

[P23A]

- *Cree que tiene que pasar tiempo para la normalidad y para ver los efectos 1*
- *No confía en la vacuna, en la rapidez con que se ha fabricado, en su eficacia 2*
- *Cree que ha habido muchos cambios en todos los ámbitos y la vida no va a ser igual 3*
- *Cree que tendría que estar toda la población vacunada, que sería necesaria la inmunidad total 4*
- *Cree que tendremos que seguir tomando medidas 5*
- *Otras respuestas 6*
- *N.S. 8*
- *N.C. 9*

P.24 En general, ¿cree Ud. que ante los riesgos de la pandemia habría que haber tomado medidas de control más estrictas que las que han tomado el Gobierno español y los gobiernos de las comunidades autónomas, o bien que son (eran) adecuadas y necesarias las medidas adoptadas y no hace falta tomar más medidas, o bien que no hay que tomar medidas que limiten las libertades?

[MEDIDAS]

- *Habría que haber tomado medidas más estrictas que las que han tomado el Gobierno español y los gobiernos de las comunidades autónomas 1*
- *Eran (son) adecuadas y necesarias las medidas adoptadas 2*
- *No había (hay) que tomar medidas que limiten las libertades 3*
- *No tiene información suficiente 4*
- *N.S. 8*
- *N.C. 9*

P.25 Cuando se habla de política se utilizan normalmente las expresiones izquierda y derecha. Situándonos en una escala de 10 casillas, como un termómetro, que van del 1 al 10, en la que 1 significa "lo más a la izquierda" y 10 "lo más a la derecha",¿en qué casilla se colocaría Ud.?

[ESCIDEOL]

- *1 Izda. 1*
- *2 ..... 2*
- *3 ..... 3*
- *4 ..... 4*
- *5 ..... 5*
- *6 ..... 6*
- *7 ..... 7*
- *8 ..... 8*
- *9 ..... 9*
- *10 Dcha. 10*
- *N.S. 98*
- *N.C. 99*

P.26 ¿Me podría decir si en las elecciones generales del 10 de noviembre de 2019...? (LEER RESPUESTAS).

[PARTICIPACIONG]

- *Fue a votar y votó 1*
- *Votó por correo 7*
- *No tenía edad para votar 2*
- *Fue a votar pero no pudo hacerlo 3*
- *No fue a votar porque no pudo 4*
- *Prefirió no votar 5*
- *No tenía derecho a voto 6*
- *No recuerda 8*
- *N.C. 9*

Estudio: Efectos y consec. del Coronavirus Septiembre (V)
Clave: ECIS3336

P.26a ¿Y podría decirme a qué partido o coalición votó? (RESPUESTA ESPONTÁNEA).

Filtros:
Si NO PARTICIPACIONG=(1;7) ir a la siguiente.

[RECUVOTOG]

PSOE	*2*
PP	*1*
VOX	*18*
Unidas Podemos	*21*
En Comú Podem	*6*
En Común - Unidas Podemos	*67*
Ciudadanos	*4*
Más País	*50*
ERC	*8*
JxCat	*9*
CUP	*19*
EAJ-PNV	*11*
EH Bildu	*12*
CCa-PNC-NC	*13*
Navarra Suma (UPN)	*14*
Més Compromís	*7*
BNG (Bloque Nacionalista Galego)	*24*
PRC (Partido Regionalista de Cantabria)	*43*
Teruel Existe	*68*
PACMA (Partido Animalista)	*17*
Otros partidos	*95*
En blanco	*96*
Voto nulo	*77*
No recuerda	*98*
N.C.	*99*

P.27 ¿En qué situación laboral se encuentra Ud. actualmente?

[SITLAB]

Trabaja	*1*
Jubilado/a o pensionista (anteriormente ha trabajado)	*2*
Pensionista (anteriormente no ha trabajado)	*3*
En paro y ha trabajado antes	*4*
En paro y busca su primer empleo	*5*
Estudiante	*6*
Trabajo doméstico no remunerado	*7*
Otra situación	*8*
N.C.	*9*

P.28 ¿Me puede decir cuál es su ocupación actual?

Filtros:
Si NO SITLAB=1 ir a [ESCUELA] - Escolarización de la persona entrevistada

[CNO11]

Directores/as y gerentes	*1*
Profesionales y científicos/as e intelectuales	*2*
Técnicos/as y profesionales de nivel medio	*3*
Personal de apoyo administrativo	*4*
Trabajadores/as de los servicios y vendedores/as de comercios y mercados	*5*
Agricultores/as y trabajadores/as cualificados/as agropecuarios/as, forestales y pesqueros/as	*6*
Oficiales/as, operarios/as y artesanos/as de artes mecánicas y de otros oficios	*7*
Operadores/as de instalaciones y máquinas y ensambladores/as	*8*
Ocupaciones elementales	*9*
Ocupaciones militares y cuerpos policiales	*10*
Otra/o	*11*
N.C.	*99*

P.29 ¿Ha ido Ud. a la escuela o cursado algún tipo de estudios? (ENTREVISTADOR/A: en caso negativo, preguntar si sabe leer y escribir).

[ESCUELA]

No, es analfabeto/a	*1*
No, pero sabe leer y escribir	*2*
Sí, ha ido a la escuela	*3*
N.C.	*9*

Saltos:
Si NO ESCUELA=3 ir a [RELIGION]

P.29a ¿Cuáles son los estudios de más alto nivel oficial que Ud. ha cursado (con independencia de que los haya terminado o no)? Por favor, especifique lo más posible, diciéndome el curso en que estaba cuando los terminó (o los interrumpió) y también el nombre que tenían entonces esos estudios (ej: 3 años de estudios primarios, primaria, 5º de bachillerato, Maestría Industrial, preuniversitario, 4º de EGB, licenciatura, doctorado, FP1, etc.). (ENTREVISTADOR/A: si aún está estudiando, anotar el último curso que haya completado y el ciclo correcto en las opciones de respuesta. Si no ha completado la primaria, anotar nº de años que asistió a la escuela , diferenciando entre menos de 5 y más de 5).

[CURSOENTREV]

CURSO _____ *N.S. - N.R. = 98* *N.C. = 99*

[NOMBREESTENTREV]

NOMBRE DE ESTUDIOS _____ *N.S. - N.R. = 98* *N.C. = 99*

Estudio: Efectos y consec. del Coronavirus Septiembre (V)
Clave: ECIS3336

[NIVELESTENTREV]

- *01. Menos de 5 años de escolarización ... 1*
- *02. Educación primaria (Educación primaria de LOGSE, 5º Curso de EGB, Enseñanza primaria antigua) ... 2*
- *03. Cualificación profesional grado inicial (FP grado inicial). PCPI (Programas de Cualificación Profesional Inicial, que no precisan de titulación académica de la primera etapa de secundaria para su realización). Programas de garantía social ... 3*
- *04. Educación secundaria (ESO, EGB. Graduado Escolar. Certificado de Escolaridad, Bachillerato Elemental) ... 4*
- *05. FP de grado medio (Ciclo/módulo formativo de FP (grado medio), de Artes Plásticas y Diseño, Música y danza, Enseñanzas deportivas, FP I, Bachiller laboral elemental. Oficialía Industrial; Bachillerato Comercial). ... 5*
- *06. Bachillerato (Bachillerato LOGSE, BUP, Bachillerato superior (6º), Bachillerato universitario (7º), Incluidos COU y PREU) ... 6*
- *07. FP de grado superior (Ciclo/módulo formativo de FP (grado superior) de Artes Plásticas, Diseño, Música y danza, Deporte, FP II, Bach. Laboral Sup., Maestría industrial, Perito Mercantil; Secretariado de 2º grado; Grado Medio conservatorio) ... 7*
- *08. Arquitectura-ingeniería técnica (Arquitectura/ingeniería técnica, Aparejador; Peritos) ... 8*
- *09. Diplomatura (ATENCIÓN: solo Diplomaturas oficiales, no codificar aquí los tres primeros años de una licenciatura o grado con mayor duración). ... 9*
- *10. Grado (Estudios de grado, Enseñanzas Artísticas equivalentes (desde 2006)) ... 10*
- *11. Licenciatura (Titulaciones con equivalencia oficial: 2º ciclo INEF; Danza y arte dramático (desde 1992); Grado superior de música) ... 11*
- *12. Arquitectura/ingeniería ... 12*
- *13. Máster oficial universitario (Especialidades médicas o equivalente) ... 13*
- *14. Doctorado ... 14*
- *15. Títulos propios de posgrado (máster no oficial, etc.) ... 15*
- *16. Otros estudios ... 16*
- *N.S./No recuerda ... 98*
- *N.C. ... 99*

P.30 ¿Cómo se define Ud. en materia religiosa: católico/a practicante, católico/a no practicante, creyente de otra religión, agnóstico/a, indiferente o no creyente, o ateo/a?

[RELIGION]

- *Católico/a practicante ... 1*
- *Católico/a no practicante ... 2*
- *Creyente de otra religión ... 3*
- *Agnóstico/a (no niegan la existencia de Dios pero tampoco la descartan) ... 4*
- *Indiferente, no creyente ... 5*
- *Ateo/a (niegan la existencia de Dios) ... 6*
- *N.C. ... 9*

Saltos:
Si RELIGION=(4;5;6;9) ir a [CLASESOCIAL] - Clase social subjetiva de la persona entrevistada

P.30a ¿Con qué frecuencia asiste Ud. a misa u otros oficios religiosos, sin contar las ocasiones relacionadas con ceremonias de tipo social, por ejemplo, bodas, comuniones o funerales?

Filtros:
Si NO RELIGION=(1;2;3) ir a [CLASESOCIAL] - Clase social subjetiva de la persona entrevistada

[FRECUENCIARELIGION]

- *Nunca ... 1*
- *Casi nunca ... 2*
- *Varias veces al año ... 3*
- *Dos o tres veces al mes ... 4*
- *Todos los domingos y festivos ... 5*
- *Varias veces a la semana ... 6*
- *N.C. ... 9*

P.31 ¿A qué clase social diría Ud. que pertenece? (RESPUESTA ESPONTÁNEA).

[CLASESOCIAL]

- *Clase alta ... 1*
- *Clase media-alta ... 2*
- *Clase media-media ... 3*
- *Clase media-baja ... 4*
- *Clase trabajadora/obrera ... 5*
- *Clase baja ... 12*
- *Clase pobre ... 6*
- *Infraclase ... 7*
- *Proletariado ... 8*
- *A los/as de abajo ... 9*
- *Excluidos/as ... 10*
- *A la gente común ... 11*
- *Otra (especificar) ... 96*
- *No cree en las clases ... 97*
- *No sabe, duda ... 98*
- *N.C. ... 99*

Filtros:
Si NO CLASESOCIAL=(96) ir a la siguiente.

[CLASESOCIAL_COD]

_____

FIN DE LA ENTREVISTA.

MUCHAS GRACIAS POR SU AMABILIDAD Y POR EL TIEMPO QUE NOS HA DEDICADO.