

UNIVERSIDAD DE SALAMANCA  
DEPARTAMENTO DE ESTADÍSTICA

---



TESIS DOCTORAL

**VERSIÓN INFERENCIAL DE LOS  
MÉTODOS BIPLLOT BASADA EN  
REMUESTREO BOOTSTRAP Y SU  
APLICACIÓN A TABLAS DE TRES VÍAS**

ANA BELÉN NIETO LIBRERO

2015







VERSIÓN INFERENCIAL DE LOS MÉTODOS BILOT BASADA  
EN REMUESTREO BOOTSTRAP Y SU APLICACIÓN A TABLAS  
DE TRES VÍAS

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

*Ana Belén Nieto Librero*

Salamanca

2015





**VNiVERSiDAD  
D SALAMANCA**

CAMPUS DE EXCELENCIA INTERNACIONAL

DEPARTAMENTO DE ESTADÍSTICA

---

**M. PURIFICACIÓN GALINDO VILLARDÓN**

**M. PURIFICACIÓN VICENTE GALINDO**

*Profesoras del Área de Estadística e Investigación Operativa de la Universidad  
de Salamanca*

CERTIFICAN:

Que **Doña Ana Belén Nieto Librero**, licenciada en Matemáticas, ha realizado en el Departamento de Estadística de la Universidad de Salamanca, bajo su dirección, el trabajo que para optar al Grado de Doctor, presenta con el título: *“Versión Inferencial de los Métodos Biplot basada en Remuestreo Bootstrap y su Aplicación a Tablas de Tres Vías”*; y para que conste, firma el presente certificado en Salamanca, en julio de 2015.



# Agradecimientos

A las directoras Dra. Purificación Galindo y Dra. Purificación Vicente por sus conocimientos, sus recomendaciones, su manera de trabajar y sobretodo por su confianza en mí para llevar a cabo este trabajo.

Al Dr. José Luis Vicente Villardón por compartir sus conocimientos cuando fue necesario.

A la Dra. Adelaide Freitas por la aportación que hizo posible que se pudiera terminar la última parte de este trabajo.

A Elisa Frutos Bernal por las innumerables veces que me prestó su ayuda y por haberme acompañado a lo largo de esta travesía.

A Luis por haber hecho de padre y madre de nuestros hijos para que este trabajo se pudiera finalizar. A Diego y Arturo que me inspiran la motivación necesaria para cumplir mis metas.

A todos los que, sin saberlo, han contribuído a que este proyecto viera su fin.  
Gracias.



*“El pensamiento estadístico  
será algún día tan necesario  
para el ciudadano competente  
como la habilidad de leer  
y escribir”*

*H. G. Wells*





*A mis padres*

*A Luis,  
Diego y  
Arturo*



---

# Índice general

<b>Introducción</b>	<b>27</b>
<b>1. APROXIMACIÓN DE MATRICES E INTERPRETACIONES GEOMÉTRICAS</b>	<b>35</b>
1.1. Valores y Vectores Propios . . . . .	37
1.1.1. Valores y Vectores Propios de una Matriz y su Traspuesta	41
1.1.2. Valores y Vectores Propios de una Matriz Simétrica . . . . .	50
1.2. Aproximación de Matrices . . . . .	55
1.2.1. Subespacio de Mejor Ajuste . . . . .	56
1.2.2. Descomposición en Valores Singulares . . . . .	62
<b>2. BIPLLOT</b>	<b>67</b>
2.1. Introducción y Revisión Bibliográfica . . . . .	69
2.2. Formulación Teórica . . . . .	75
2.2.1. Representaciones Biplot . . . . .	75
2.2.2. Interpretaciones Geométricas . . . . .	77
2.2.3. Propiedades de los marcadores . . . . .	81
2.2.4. Bondad de ajuste . . . . .	87
2.2.5. Contribuciones . . . . .	90
2.2.6. Software para Biplot . . . . .	92

2.3. Bootstrap sobre Biplot . . . . .	98
2.3.1. Bootstrap . . . . .	101
2.3.2. Intervalos de Confianza Bootstrap . . . . .	104
2.4. Paquete <i>biplotbootGUI</i> . . . . .	112
2.5. Aplicación a Datos . . . . .	121
2.5.1. Datos Simulados . . . . .	121
2.5.2. Datos Iris . . . . .	136
<b>3. ANÁLISIS DE PARES DE MATRICES CON UNA DIMEN-</b>	
<b>SIÓN COMÚN</b>	<b>151</b>
3.1. Introducción . . . . .	153
3.1.1. Software . . . . .	157
3.2. Análisis Canónico de Correspondencias . . . . .	159
3.2.1. Análisis de las Contribuciones de los elementos y los factores en el CCA . . . . .	162
3.2.2. Análisis Canónico de Correspondencias para variables cualitativas . . . . .	164
3.2.3. Interpretación . . . . .	167
3.3. Análisis Canónico No Simétrico de Correspondencias . . . . .	169
3.3.1. Análisis de las Contribuciones de los elementos y los factores en el CNCA . . . . .	173
3.3.2. Análisis Canónico No Simétrico de Correspondencias para variables cualitativas . . . . .	176
3.3.3. Interpretación . . . . .	178
3.4. Análisis de Coinercia . . . . .	180
3.5. Bootstrap sobre Análisis Canónico de Correspondencias Simétrico y no Simétrico . . . . .	184

3.6. Programa <i>cncaGUI</i> . . . . .	185
3.7. Aplicación a Datos . . . . .	195
<b>4. BILOT MÚLTIPLES</b>	<b>209</b>
4.1. Introducción . . . . .	211
4.2. Biplot Múltiples . . . . .	215
4.3. Varios conjuntos de individuos sobre un mismo conjunto de variables	216
4.3.1. Primera Etapa: Estandarización por columnas de la tabla $\mathbf{X}$ .	217
4.3.2. Segunda Etapa: Análisis Individuales . . . . .	218
4.3.3. Tercera Etapa: Biplot ponderado de la matriz $\tilde{\mathbf{X}}$ . . . . .	219
4.3.4. Cuarta Etapa: Medidas de la calidad de representación Biplot para $\tilde{\mathbf{X}}$ . . . . .	224
4.3.5. Bondad del Ajuste para la matriz $\mathbf{X}$ . . . . .	227
4.4. Varios conjuntos de variables observados sobre un único conjunto de individuos . . . . .	229
4.4.1. Primera Etapa: Análisis Individuales . . . . .	230
4.4.2. Segunda Etapa: Biplot ponderado de la matriz $\tilde{\mathbf{X}}$ . . . . .	231
4.4.3. Tercera Etapa: Medidas de la calidad de la representación Biplot para matriz $\tilde{\mathbf{X}}$ . . . . .	231
4.4.4. Cuarta Etapa: Bondad del Ajuste para la matriz $\mathbf{X}$ . . . . .	233
4.5. Bootstrap sobre datos de tres vías . . . . .	234
4.5.1. Bootstrap sobre Biplot Múltiple . . . . .	235
4.6. Programa <i>MultibiplotGUI</i> . . . . .	236
4.7. Aplicación a Datos . . . . .	243
<b>5. CLUSTERING DISJOINT BILOT</b>	<b>261</b>
5.1. Clustering Biplot . . . . .	263
5.1.1. Notación . . . . .	265

---

5.1.2. Modelo Clustering Biplot . . . . .	266
5.1.3. Algoritmo CBiplot . . . . .	271
5.2. Disjoint Biplot . . . . .	274
5.2.1. Notación . . . . .	276
5.2.2. Modelo Disjoint Biplot . . . . .	277
5.2.3. Algoritmo DBiplot . . . . .	280
5.3. Clustering Disjoint Biplot . . . . .	284
5.3.1. Notación . . . . .	285
5.3.2. Modelo Clustering Disjoint Biplot . . . . .	287
5.3.3. Algoritmo CDBiplot . . . . .	287
5.4. Programa <i>CDBiplot</i> . . . . .	292
5.5. Aplicación a Datos . . . . .	297
<b>Conclusiones</b>	<b>305</b>
<b>Artículos Publicados</b>	<b>309</b>
<b>Bibliografía</b>	<b>343</b>
<b>Apéndice A</b>	<b>377</b>

---

# Índice de figuras

2.1. Tipos de Biplot. . . . .	77
2.2. Representación Biplot de una matriz que contiene información de 6 variables medidas sobre 9 individuos. . . . .	78
2.3. Esquema del algoritmo Bootstrap. . . . .	103
2.4. Ventana principal. . . . .	114
2.5. Ventana de opciones. . . . .	116
2.6. Ventana con el gráfico de barras representando la inercia absorbida por cada eje. . . . .	117
2.7. Ventana que muestra la representación Biplot dos dimensiones. . .	117
2.8. Gráfico que muestra la representación HJ Biplot para los datos simulados en las dos primeras dimensiones. . . . .	123
2.9. Histograma de las calidades de aproximación de las variables para datos simulados. . . . .	125
2.10. Histograma de los valores propios para datos simulados. . . . .	126
2.11. Histograma de los ángulos entre variables para datos simulados. .	128
2.12. Histograma de los ángulos entre variables y ejes para datos simulados.	129
2.13. Histograma de la longitud de las variables para datos simulados. .	130
2.14. Histograma de las contribuciones relativas a la variabilidad total de las variables para datos simulados. . . . .	131

2.15. Histograma de las contribuciones relativas de los ejes a las variables para datos simulados. . . . .	133
2.16. Histograma de las contribuciones relativas de las variables a los ejes para datos simulados. . . . .	134
2.17. Coordenadas bootstrap para las variables datos simulados. . . . .	135
2.18. Gráfico que muestra la representación HJ Biplot para los datos iris en las dos primeras dimensiones. . . . .	137
2.19. Histograma de las calidades de aproximación de las variables para datos iris. . . . .	139
2.20. Histograma de los valores propios para datos iris. . . . .	140
2.21. Histograma de los ángulos entre variables para datos iris. . . . .	141
2.22. Histograma de los ángulos entre variables y ejes para datos iris. . . . .	142
2.23. Histograma de las contribuciones relativas de las variables a la variabilidad total para datos iris. . . . .	143
2.24. Histograma de la longitud de las variables para datos iris. . . . .	144
2.25. Histograma de las contribuciones relativas de las variables a los ejes para datos iris. . . . .	146
2.26. Histograma de las contribuciones relativas de los ejes a las variables para datos iris. . . . .	147
2.27. Histograma de la calidad de aproximación de los individuos para datos iris. . . . .	148
2.28. Coordenadas bootstrap para las variables datos iris. . . . .	149
3.1. Esquema de la obtención de $\mathbf{W}$ . . . . .	160
3.2. Esquema de la obtención de $\tilde{\mathbf{P}}^*$ y $\mathbf{L}$ . . . . .	173
3.3. Esquema del análisis de la Coinercia. . . . .	183
3.4. Ventana principal. . . . .	186



---

3.5. Ventana para cambio de denominación. . . . .	187
3.6. Ventana de opciones. . . . .	188
3.7. Ventana de barplot. . . . .	189
3.8. Ventana que muestra la representación de las coordenadas obtenidas mediante el método elegido en dos dimensiones. . . . .	190
3.9. Ventana para el análisis Bootstrap. . . . .	193
3.10. Gráfico que muestra la representación de los datos en las dos primeras dimensiones. . . . .	201
3.11. Histogramas y gráficos de normalidad para los valores propios. . .	203
3.12. Histogramas y gráficos de normalidad para la proporción de inercia explicada por cada eje en el espacio proyectado. . . . .	204
3.13. Coordenadas de variables, especies y lugares para las 1000 submuestras seleccionadas mediante Bootstrap. . . . .	206
4.1. Ventana para seleccionar el tipo de datos. . . . .	237
4.2. Ventana de opciones. . . . .	238
4.3. Ventana con el gráfico de barras representando la inercia absorbida por cada eje. . . . .	239
4.4. Ventana que muestra la representación Biplot en dos dimensiones. . . . .	239
4.5. Ventana para el análisis Bootstrap. . . . .	243
4.6. Gráfico en las dos primeras dimensiones de las coordenadas resultantes del Biplot Múltiple. . . . .	247
4.7. Histogramas y gráficos de normalidad para los valores propios. . .	248
4.8. Histogramas y gráficos de normalidad para las contribuciones relativas a la variabilidad total de las variables. . . . .	249
4.9. Histogramas y gráficos de normalidad para las contribuciones relativas de las variables a la conformación de los ejes factoriales. .	251

4.10. Histogramas y gráficos de normalidad para las contribuciones relativas de los ejes factoriales a las variables. . . . .	253
4.11. Histogramas y gráficos de normalidad para las contribuciones relativas a la variabilidad total de las matrices. . . . .	254
4.12. Histogramas y gráficos de normalidad para las contribuciones relativas de las matrices a la conformación de los ejes factoriales. .	256
4.13. Histogramas y gráficos de normalidad para las contribuciones relativas de los ejes factoriales a las matrices. . . . .	258
4.14. Coordenadas de las variables para las 1000 muestras bootstrap. .	259
5.1. Esquema del algoritmo CBiplot. . . . .	274
5.2. Esquema de construcción de la submatriz $\mathbf{W}_q$ . . . . .	281
5.3. Esquema del algoritmo DBiplot. . . . .	283
5.4. Esquema del algoritmo CDBiplot. . . . .	292
5.5. Ventana principal de la función CDBiplot. . . . .	293
5.6. Ventana para elegir los parámetros para ejecutar el algoritmo. . .	293
5.7. Ventana con la representación conjunta de objetos y variables. . .	294
5.8. Ventana para elegir las opciones respecto del polígono envolvente.	296
5.9. Gráfico que muestra la representación HJ Biplot para los datos iris en las dos primeras dimensiones. . . . .	298
5.10. Gráfico que muestra los clusters obtenidos mediante el algoritmo CDBiplot en las dos primeras dimensiones. . . . .	298
5.11. Gráfico que muestra los clusters y las variables obtenidas mediante el algoritmo CDBiplot en las dos primeras dimensiones. . . . .	299
5.12. Gráfico que muestra la representación HJ Biplot para los datos Vinho Verde en las dos primeras dimensiones. . . . .	301

---

5.13. Gráfico que muestra los clusters y las variables obtenidas mediante el algoritmo CDBiplot en las dos primeras dimensiones. Datos Vinho Verde. . . . .	301
5.14. Gráfico que muestra los clusters obtenidos mediante el algoritmo CDBiplot en las dos primeras dimensiones. Datos Vinho Verde. . .	302



---

# Índice de tablas

2.1. Obtención de marcadores y sus calidades de representación. . . . .	90
2.2. Biplot en R. . . . .	96
2.3. Paquetes de R que aluden a la palabra Biplot. . . . .	97
2.4. Matriz sigma para la generación de los datos simulados. . . . .	121
2.5. Valores propios y variabilidad explicada (%) por cada eje para los datos simulados. . . . .	122
2.6. Contribuciones relativas del factor al elemento columna para datos simulados. . . . .	122
2.7. Longitud de las variables para datos simulados. . . . .	123
2.8. Ángulos entre variables para datos simulados. . . . .	123
2.9. Calidades de aproximación de las columnas para datos simulados.	125
2.10. Valores propios para datos simulados. . . . .	125
2.11. Ángulos entre variables para datos simulados. . . . .	127
2.12. Ángulos entre variables y ejes para datos simulados. . . . .	127
2.13. Longitud de las variables para datos simulados. . . . .	130
2.14. Contribuciones relativas a la variabilidad total de las variables para datos simulados. . . . .	131
2.15. Contribuciones relativas de los ejes a las variables para datos simulados. . . . .	132

2.16. Contribuciones relativas de las variables a los ejes para datos simulados. . . . .	132
2.17. Valores propios y variabilidad explicada (%) por cada eje para los datos iris. . . . .	136
2.18. Contribuciones relativas del factor al elemento fila para datos iris.	137
2.19. Contribuciones relativas del factor al elemento columna para datos iris. . . . .	137
2.20. Ángulos entre variables para datos iris. . . . .	138
2.21. Ángulos entre variables y ejes para datos iris. . . . .	138
2.22. Calidades de aproximación de las columnas para datos iris. . . . .	139
2.23. Valores propios para datos iris. . . . .	139
2.24. Ángulos entre variables para datos iris. . . . .	141
2.25. Ángulos entre variables y los dos primeros ejes para datos iris. . .	142
2.26. Contribuciones a la variabilidad total de las variables para datos iris. . . . .	143
2.27. Longitud de las variables para datos iris. . . . .	144
2.28. Contribuciones relativas del elemento columna al factor para datos iris. . . . .	145
2.29. Contribuciones relativas del factor al elemento columna para datos iris. . . . .	145
2.30. Calidad de aproximación de las filas para datos iris. . . . .	148
3.1. Distancias al cuadrado (especies y lugares), con respecto al espacio proyectado ( <i>PS</i> ) y al espacio original ( <i>OS</i> ). . . . .	175
3.2. Calidades de representación para las puntuaciones de ejes factoriales y elementos (especies y lugares), con respecto al espacio proyectado ( <i>PS</i> ) y al espacio original ( <i>OS</i> ). $\alpha = 1, \dots, v$ . . . . .	175

---

3.3. Valores propios y variabilidad explicada (%) y acumulada para los tres primeros ejes retenidos. Espacio Projectado. . . . .	197
3.4. Variabilidad explicada (%) y acumulada para los tres primeros ejes retenidos. Espacio Original. . . . .	197
3.5. Contribuciones relativas de los primeros tres factores a las especies. Espacio Projectado. . . . .	199
3.6. Contribuciones relativas de los primeros tres factores a las especies. Espacio Original. . . . .	200
3.7. Resultados bootstrap para valores propios. . . . .	202
3.8. Resultados bootstrap para la inercia absorbida por cada eje en el espacio proyectado. . . . .	204
4.1. Valores propios y variabilidad explicada (%) por cada eje para los datos iris. . . . .	244
4.2. Contribuciones relativas de las variables a la variabilidad total. . .	245
4.3. Contribuciones relativas de las variables a la conformación de los tres primeros ejes. . . . .	245
4.4. Contribuciones relativas de los ejes a las variables. . . . .	245
4.5. Contribuciones relativas de las matrices a la variabilidad total. . .	246
4.6. Contribuciones relativas de las matrices a la conformación de los tres primeros ejes. . . . .	246
4.7. Contribuciones relativas de los ejes a las matrices. . . . .	246
4.8. Resultados bootstrap para valores propios. . . . .	248
4.9. Resultados bootstrap para las contribuciones relativas de cada variable a la variabilidad total. . . . .	249
4.10. Resultados bootstrap para las contribuciones relativas de las variables a los ejes factoriales. . . . .	250

4.11. Resultados bootstrap para la contribuciones relativas de los ejes factoriales a las variables. . . . .	252
4.12. Resultados bootstrap para las contribuciones relativas de las matrices a la variabilidad total. . . . .	254
4.13. Resultados bootstrap para las contribuciones relativas de las matrices a los ejes factoriales. . . . .	255
4.14. Resultados bootstrap para las contribuciones relativas de los ejes factoriales a las matrices. . . . .	257



# INTRODUCCIÓN



Los métodos de ordenación y de reducción de la dimensión ofrecen diferentes parámetros para poder presentar los resultados obtenidos a partir de una muestra de datos multivariante. Sin embargo, estos métodos únicamente nos muestran esos resultados de una manera incompleta ya que sólo se obtienen estimaciones puntuales de tales parámetros, sin ninguna información acerca de la incertidumbre proporcionada por los mismos. Este hecho, hace que una parte de la comunidad científica considere estas técnicas multivariantes como exploratorias y sin ninguna certeza de que lo que se deduzca de las muestras analizadas se pueda extender a la población de la que proceden. Para proporcionar unos resultados completos que puedan ser considerados válidos para la población y no sólo a nivel muestral es necesario proporcionar una forma de decidir cómo de exacta es la información que se presenta. El método más común para proporcionar una indicación de la cantidad de incertidumbre de un parámetro son los intervalos de confianza representados por los límites de confianza.

Efron, 1979, 1987; Efron y Tibshirani, 1993 propusieron los métodos Bootstrap, cuya idea principal es que la inferencia sobre una población a partir de una muestra se puede obtener realizando sucesivos remuestreos sobre ella y haciendo inferencia sobre esta nueva “muestra”. La forma de realizar el remuestreo es, a partir de la muestra original de tamaño  $n$ , extraer diferentes submuestras de  $n$  elementos tomados con reposición. En las submuestras extraídas se calcula el parámetro que se desea estudiar y con todos los resultados obtenidos se realiza la inferencia. Los métodos Bootstrap proporcionan diferentes formas de calcular intervalos de confianza para los parámetros calculados a partir de una muestra de datos multivariantes. Tienen la ventaja además de que son métodos sencillos que no requieren del conocimiento de la distribución teórica de la población de partida y tampoco necesitan un tamaño de muestra elevado para realizar las estimaciones.

Debido al problema de muchas técnicas multivariantes de obtener únicamente estimaciones puntuales, numerosos autores han desarrollado versiones inferenciales basándose en las ideas de los métodos Bootstrap. Gifi, 1990; Greenacre, 1984; Meulman, 1982 introdujeron la idea de remuestreo Bootstrap en el caso de matrices de dos vías y en el Análisis de Correspondencias Múltiple; Chatterjee, 1984; Lambert et al., 1990, 1991 lo aplican en el contexto del Análisis Factorial; Daudin et al., 1988; Diaconis y Efron, 1983; Holmes, 1985, 1989; Stauffer et al., 1985 utilizaron la metodología Bootstrap en el caso del Análisis de Componentes Principales para proponer intervalos de confianza para los puntos representados en el subespacio de los ejes principales intentando resolver el problema de la elección del número de ejes a retener; Milan y Whittaker, 1995 lo proponen en el caso de modelos bilineales que incorporan Descomposición en Valores Singulares; Raykov y Little, 1999 utilizan los métodos Bootstrap para evaluar el ajuste de las rotaciones Procrustes; Linting et al., 2007 en el análisis de Componentes Principales no lineal; Timmerman et al., 2009 utilizan los métodos Bootstrap para estimar intervalos de confianza en el Análisis de Componentes Multinivel.

En el contexto del análisis de Correspondencias también se encuentran numerosas referencias que utilizan esta metodología para presentar medidas de precisión de los resultados puntuales que proporcionan sus versiones clásicas. Por ejemplo, Meulman, 1982 propone volver a ejecutar en cada muestra bootstrap un análisis de Correspondencias y poner todas las soluciones juntas en el mismo gráfico. Greenacre, 1984 propuso usar los resultados obtenidos a partir de las submuestras como elementos suplementarios en el análisis de la matriz original. Knox y Peet, 1989 aplican los métodos bootstrap sobre el análisis de Correspondencias Detrended. También se ha estudiado la estabilidad de los resultados del análisis no simétrico de Correspondencias utilizando los métodos Bootstrap (Balbi, 1992). Markus y Visser, 1992 aplican

dichos métodos para generar regiones de confianza en el análisis Múltiple de Correspondencias. Reiczigel, 1996 desarrollaron un test bootstrap para el análisis de Correspondencias que incluye la construcción de intervalos de confianza. Lombardo y Ringrose, 2012; Lombardo et al., 2012 también proponen la construcción de regiones de confianza para los análisis de Correspondencias simétrico y no simétrico.

Si se hace una revisión de las técnicas que analizan datos de tres vías también se encuentran versiones inferenciales basadas en los métodos Bootstrap. Así, Kiers, 2004 propone el cálculo de intervalos de confianza para los métodos CANDECOM/PARAFAC y TUCKER3 (Tucker, 1966). El muestreo que proponen se basa en la descomposición de los cubos de datos en capas que contienen la información de cada individuo. De esta forma, remuestrea matrices completas que contienen la información de cada individuo en las otras dos vías estudiadas. Abdi et al., 2013; Husson et al., 2005; Pagès y Husson, 2005 remuestran tablas completas. Otros autores como Abdi et al., 2009 utilizan otro tipo de remuestreo que denominan *remuestreo dividido a la mitad* ya que las observaciones de los datos que analizan no siempre son independientes y pueden tener una correlación temporal. (Dehlholm et al., 2012) presentan una función que forma parte del paquete **FactoMineR** que permite remuestrear filas dentro de cada matriz pero partiendo de las coordenadas resultantes del análisis Factorial Múltiple en lugar de los datos originales. Con las filas remuestreadas calculan los centroides para cada matriz. Cadoret y Husson, 2013 proponen un método para construir elipses de confianza que denominan *bootstrap total truncado*, implementado en el paquete **SensoMineR** (Le y Husson, 2008). El esquema que siguen es contruir las muestras bootstrap, aplicar el método y luego utilizar rotaciones Procrustes en un número reducido de dimensiones.

Debido a que los métodos Bootstrap están pensados mayoritariamente para

una muestra univariante, es necesario adaptarlos para el caso de técnicas multivariantes. De este modo, se presentan estrategias a seguir para obtener versiones inferenciales de diferentes técnicas multivariantes en los tres contextos señalados anteriormente.

En el capítulo 1 se presentan los principales aspectos teóricos de la aproximación de matrices en dimensión reducida y la interpretación geométrica.

En el segundo capítulo se realiza una revisión bibliográfica de los métodos Biplot (Gabriel, 1971; Galindo, 1986) así como del software disponible para su utilización. A continuación, se realiza la propuesta de una versión inferencial basada en los métodos Bootstrap (Efron, 1979, 1987; Efron y Tibshirani, 1993). Para facilitar su utilización se presenta el software desarrollado en el entorno R, `biplotbootGUI` que se ha implementado como una interfaz gráfica de usuario que permite al usuario interactuar a través de ventanas, botones y menús sin el requisito de tener un amplio conocimiento del lenguaje de programación. Por último, se ilustra su manejo y la interpretación de los resultados mediante su utilización en dos ejemplos.

En el tercer capítulo, se presentan las principales técnicas que existen en el contexto del análisis de Correspondencias. También se han analizado los softwares que existen para su uso. Como se ha explicado anteriormente, estas técnicas también presentan la carencia de versiones inferenciales así que después de analizar las versiones basadas en los métodos Bootstrap que hay para estas técnicas, se propone una nueva versión para el análisis Canónico de Correspondencias simétrico (Ter Braak, 1986) y no simétrico (Willems y Galindo, 2008). Para permitir su aplicación práctica, se ha desarrollado una interfaz gráfica en el lenguaje R que permite por un lado, la utilización del análisis Canónico de Correspondencias no Simétrico (que carecía de software específico) y por otro, presenta versiones inferenciales tanto para este análisis de Correspondencias

no simétrico como para el simétrico. Dicha interfaz se denomina `cncaGUI`. La utilización de este programa se presenta mediante la ejecución de un ejemplo con datos reales.

Siguiendo con el desarrollo de versiones inferenciales, el siguiente paso esperado es ver como se realiza en el contexto de tres vías. Así, en el capítulo 4 se realiza una revisión de las técnicas de tres vías y el software desarrollado para su utilización. A continuación, se revisan las diferentes estrategias elaboradas por diversos autores para proporcionar versiones inferenciales de algunas de ellas y por último, se presenta una nueva estrategia diseñada para proporcionar una versión inferencial del análisis Biplot Múltiple, un método basado en las ideas del análisis Factorial Múltiple pero que tiene la ventaja de proporcionar representaciones Biplot de los resultados obtenidos. Como en los capítulos anteriores, se ha implementado una interfaz gráfica que permite la utilización de esta técnica tanto en su versión clásica como inferencial y se ha puesto en práctica con un ejemplo. La interfaz gráfica se ha denominado `multibiplotGUI`.

Esta tesis presenta otra parte bien diferenciada de la anterior en la que se estudian las técnicas que buscan combinar la clasificación de los individuos con métodos de reducción de la dimensión y que pretenden además, buscar una mejor interpretación de los ejes factoriales extraídos. Relacionado con estas ideas se encuentran las técnicas de biclustering (Hartigan, 1972, 1975). Sin embargo, estos métodos no tienen definido un problema de optimización. Por ello, Bock, 1979, 2003; DeSarbo, 1982 proponen alternativas al biclustering basadas en diferentes criterios de optimización. Vichi, 2000 presenta el algoritmo Doble k-means. Witten y Tibshirani, 2010 proponen el *Sparse k-means Clustering*, que pretende encontrar una clasificación de los objetos mediante el algoritmo k-means y buscar las variables que tengan más influencia en dicha clasificación utilizando la penalización *lasso* (Tibshirani, 1996). En Vichi y Saporta, 2009

se propone un método iterativo que busca la partición de los objetos alrededor de un conjunto de centroides y la de las variables en torno a un conjunto de componentes. Dicho método se denomina CDPCA (*Clustering Disjoint Principal Component Analysis*). Este método tiene la ventaja, respecto del sparse PCA, de obtener componentes en las que cada variable únicamente va a contribuir a la conformación de una componente de manera similar a la presentada en [Vigneau y Qannari, 2004](#). Recientemente, [Macedo y Freitas, 2015](#) desarrollan un programa para la puesta en práctica de este algoritmo.

En el capítulo 5 se presentan tres métodos con sus correspondientes algoritmos en este contexto que utilizan como técnica de reducción de la dimensión el método HJ Biplot: *Clustering Biplot*, cuyo objetivo es la búsqueda simultánea de la mejor clasificación de los individuos utilizando el algoritmo k-means en un espacio de dimensión reducida que explique la mayor parte de la variabilidad presente en los datos utilizando el HJ Biplot; *Disjoint Biplot*, cuyo objetivo es encontrar la mejor representación en dimensión reducida pero forzando a que las variables únicamente contribuyan a la conformación de uno sólo de los ejes factoriales; *Clustering Disjoint Biplot*, cuyo objetivo es una combinación de los dos anteriores, es decir, busca la mejor clasificación de los individuos mediante el algoritmo k-means a la vez que maximiza la variabilidad explicada por los ejes extraídos de tal forma que cada variable sólo tenga contribución en la conformación de un eje. Para poder poner en práctica estos tres algoritmos se ha desarrollado una función en el lenguaje R en forma de interfaz gráfica llamada `CDBiplot` y se ha utilizado sobre un conjunto de datos reales para mostrar su utilización e interpretación.

Por último, se presentan las principales conclusiones para finalizar este trabajo.



# Capítulo 1

## APROXIMACIÓN DE MATRICES E INTERPRETACIONES GEOMÉTRICAS



Dada una matriz de datos con información de  $n$  individuos observados sobre  $p < n$  variables, los métodos de análisis multivariante estudian la manera de obtener un subespacio de menor dimensión que el original donde representar la información de la forma más clara posible y con una pérdida mínima de información. Esta idea se basa en el hecho de que si se tiene una matriz con rango menor a  $p$ , es posible representarla en un subespacio de dimensión menor eliminando el ruido existente en los datos. La mejor manera de obtener ese subespacio está basada en el teorema de Descomposición en valores singulares (Eckart y Young, 1936). Se introducirán previamente aspectos del álgebra matricial necesarios para su posterior demostración.

## 1.1. Valores y Vectores Propios

Los valores propios son las medidas básicas del tamaño de una matriz invariantes por transformaciones lineales de dicha matriz. Por ejemplo, si se efectúa una rotación de los ejes, los valores propios no se modifican. Debido a ello hay medidas globales del tamaño de una matriz como la traza o el determinante que sólo dependen de estos valores para su cálculo.

Sea  $\mathbf{X}$  una matriz cuadrada de orden  $n$ . Se dice que el escalar  $\lambda$  perteneciente a un campo  $\mathbb{K}$  es un autovalor o valor propio si existe un vector  $x$  perteneciente a  $\mathbb{K}^n$  no nulo que satisface la ecuación:

$$\mathbf{X}x = \lambda x.$$

El vector  $x$  se denomina autovector o vector propio de la matriz  $\mathbf{X}$  asociado al valor propio  $\lambda$ . Un autoespacio es el conjunto de todos los vectores propios de  $\mathbf{X}$  asociados al valor propio  $\lambda$  unido al vector nulo, es decir:

$$V_\lambda(\mathbf{X}) = \{x \in \mathbb{K}^n : x \neq \emptyset \wedge \mathbf{X}x = \lambda x\} \cup \emptyset.$$

**Teorema 1.1.** *Una condición necesaria y suficiente para que el escalar  $\lambda$  sea valor propio de la matriz  $\mathbf{X}$  es que:*

$$|\mathbf{X} - \lambda\mathbf{I} = 0|$$

donde  $|\mathbf{X} - \lambda\mathbf{I}|$  es lo que se denomina autopolinomio.

*Demostración.* Condición Necesaria. Si  $\lambda$  es un valor propio de la matriz  $\mathbf{X}$  se cumple:

$$\exists x \neq \emptyset \text{ tal que } \mathbf{X}x = \lambda x$$

es decir,

$$\exists x \neq \emptyset \text{ tal que } (\mathbf{X} - \lambda\mathbf{I})x = \emptyset$$

con lo que la matriz  $(\mathbf{X} - \lambda\mathbf{I})$  tiene determinante 0 ya que es singular.

Condición Suficiente. Si  $|\mathbf{X} - \lambda\mathbf{I}| = 0$ , entonces se tiene que:

$$\exists x \neq \emptyset \text{ tal que } (\mathbf{X} - \lambda\mathbf{I})x = \emptyset$$

por lo tanto,

$$\exists x \neq \emptyset \text{ tal que } \mathbf{X}x = \lambda x$$

y se puede afirmar que  $\lambda$  es valor propio de la matriz  $\mathbf{X}$ .

□

Se denomina espectro de la matriz  $\mathbf{X}$  al conjunto de todos sus valores propios, es decir:

$$\Lambda(\mathbf{X}) = \{\lambda \in \mathbb{R} \text{ tal que } |\mathbf{X} - \lambda\mathbf{I}| = 0\}.$$

Se llama radio espectral al radio del menor círculo que contiene a todos los valores propios de la matriz  $\mathbf{X}$ . Es decir:

$$\rho(\mathbf{X}) = \max | \lambda_i | .$$

**Teorema 1.2.** *Sea  $\mathbf{X}$  una matriz cuadrada de orden  $n$  y  $(x_1, x_2, \dots, x_r)$  un conjunto de vectores propios asociados a los valores propios  $\lambda_1, \lambda_2, \dots, \lambda_r$  respectivamente con  $\lambda_i \neq \lambda_j \forall i \neq j$ . Entonces los  $r$  vectores propios son linealmente independientes.*

*Demostración.* Supongamos que los  $r$  vectores son linealmente dependientes y que el conjunto  $(x_1, x_2, \dots, x_{j-1})$  son vectores linealmente independientes. Entonces existe un vector  $x_j$  que se puede expresar como combinación lineal de  $(x_1, x_2, \dots, x_{j-1})$ , es decir, existen escalares  $\alpha_{1j}, \alpha_{2j}, \dots, \alpha_{(j-1)j}$ , tales que:

$$x_j = \alpha_{1j}x_1 + \alpha_{2j}x_2 + \alpha_{(j-1)j}x_{j-1}.$$

Multiplicando por la matriz  $\mathbf{X}$ :

$$\mathbf{X}x_j = \sum_{i=1}^{j-1} \alpha_{ij} \mathbf{X}x_i = \sum_{i=1}^{j-1} \alpha_{ij} \lambda_i x_i.$$

Multiplicando por el valor propio  $\lambda_j$ :

$$\lambda_j x_j = \sum_{i=1}^{j-1} \alpha_{ij} \lambda_j x_i.$$

Como  $x_j$  es vector propio de  $\mathbf{X}$  asociado al valor  $\lambda_j$  se tiene:

$$\sum_{i=1}^{j-1} \alpha_{ij} \lambda_j x_i = \sum_{i=1}^{j-1} \alpha_{ij} \lambda_i x_i.$$

Es decir:

$$\sum_{i=1}^{j-1} \alpha_{ij}(\lambda_i - \lambda_j)x_i = 0.$$

Como el conjunto de vectores  $(x_1, x_2, \dots, x_{j-1})$  son linealmente independientes entonces todos los escalares  $\alpha_{ij}(\lambda_i - \lambda_j)$  son nulos  $\forall i = 1, \dots, j-1$ . Como  $\lambda_i \neq \lambda_j$   $\forall i \neq j$  se tiene que  $\alpha_{ij} = 0 \forall i = 1, \dots, j-1$ . Es decir, el vector  $x_j$  es el vector nulo, lo que contradice la condición de ser vector propio y por lo tanto el conjunto de vectores propios  $(x_1, x_2, \dots, x_r)$  es linealmente independiente.  $\square$

**Teorema 1.3.** *Si  $x$  es un vector propio de la matriz  $\mathbf{X}$  asociado al valor propio  $\lambda$ , para cualquier entero positivo  $p$  se tiene que:*

- $\lambda^p$  es valor propio de  $\mathbf{X}^p$ .
- $x$  es vector propio de  $\mathbf{X}^p$  asociado a  $\lambda^p$ .

*Demostración.* Por hipótesis  $\mathbf{X}x = \lambda x$ . Si se multiplica por la matriz  $\mathbf{X}$ :

$$\mathbf{X}^2x = \mathbf{X}\lambda x = \lambda(\lambda x) = \lambda^2x.$$

Supongamos que se cumple para  $p-1$ :

$$\mathbf{X}^{p-1}x = \lambda^{p-1}x.$$

Vamos a demostrar que se cumple para  $p$ :

$$\begin{aligned} \mathbf{X}^p x &= \mathbf{X}\mathbf{X}^{p-1}x \\ &= \mathbf{X}(\lambda^{p-1}x) \\ &= \lambda^{p-1}(\mathbf{X}x) \end{aligned}$$

$$= \lambda^{p-1}(\lambda x)$$

$$= \lambda^p x.$$

□

### 1.1.1. Valores y Vectores Propios de una Matriz y su Traspuesta

A continuación se introducen las relaciones existentes entre los vectores y valores propios de una matriz y su traspuesta.

Se sabe que si  $\lambda$  es un valor propio de la matriz  $\mathbf{X}$  entonces se cumple que:

$$| \mathbf{X} - \lambda \mathbf{I} | = 0.$$

Como el determinante de una matriz es igual al de su traspuesta se tiene que:

$$| (\mathbf{X} - \lambda \mathbf{I}) | = 0$$

$$= | \mathbf{X}^\top - \lambda \mathbf{I} | = 0.$$

Con lo que se puede concluir que los valores propios de una matriz  $\mathbf{X}$  y los valores propios de su traspuesta  $\mathbf{X}^\top$  son iguales.

En el caso de los vectores propios no sucede así pero se puede enunciar la siguiente propiedad:

**Teorema 1.4.** *Si  $x_i$  es un vector propio de  $\mathbf{X}$  asociado al valor propio  $\lambda_i$  e  $y_j$  es un vector propio de  $\mathbf{X}^\top$  asociado al valor propio  $\lambda_j$  con  $i \neq j$ , entonces se tiene*

que  $x_i$  e  $y_j$  son ortogonales, es decir:

$$x_i^\top y_j = 0.$$

*Demostración.* Como  $x_i$  e  $y_j$  son vectores propios de  $\mathbf{X}$  e  $\mathbf{X}^\top$  respectivamente se tiene:

$$\mathbf{X}x_i = \lambda_i x_i \text{ y } \mathbf{X}^\top y_j = \lambda_j y_j.$$

Multiplicando cada ecuación por  $y_j^\top$  y  $x_i^\top$  respectivamente:

$$y_j^\top \mathbf{X}x_i = \lambda_i y_j^\top x_i$$

y

$$x_i^\top \mathbf{X}^\top y_j = \lambda_j x_i^\top y_j.$$

Tomando traspuesta en la primera se obtiene:

$$x_i^\top \mathbf{X}^\top y_j = \lambda_i x_i^\top y_j$$

e igualando términos:

$$\lambda_j x_i^\top y_j = \lambda_i x_i^\top y_j$$

de donde:

$$(\lambda_i - \lambda_j)x_i^\top y_j = 0$$

pero como  $\lambda_i \neq \lambda_j$ , forzosamente:

$$x_i^\top y_j = 0.$$

□



Supongamos que todos los valores propios de la matriz  $\mathbf{X}$  son distintos entre sí y se eligen vectores  $x_i$  e  $y_i$  vectores propios de  $\mathbf{X}$  y  $\mathbf{X}^\top$  respectivamente de tal forma que  $y_i^\top x_i = 1 \forall i = 1, \dots, n$ . Por otro lado, es posible encontrar dos matrices no singulares  $\mathbf{A}$  y  $\mathbf{B}$  cuyas columnas son los vectores propios de  $\mathbf{X}$  y  $\mathbf{X}^\top$ .

Entonces, se tiene que el producto de las  $\mathbf{B}^\top \mathbf{A}$  es la matriz identidad, con lo que se puede decir que la matriz  $\mathbf{B}$  es la inversa de la matriz  $\mathbf{A}$ .

Ahora bien, si se multiplica:

$$\begin{aligned}
 \mathbf{XA} &= \mathbf{X}(x_1, x_2, \dots, x_n) \\
 &= (\mathbf{X}x_1, \mathbf{X}x_2, \dots, \mathbf{X}x_n) \\
 &= (\lambda_1 x_1, \lambda_2 x_2, \dots, \lambda_n x_n) \\
 &= (x_1, x_2, \dots, x_n) = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_n \end{pmatrix} = \mathbf{AD} \quad (1.1)
 \end{aligned}$$

Si se multiplica por la inversa de la matriz  $\mathbf{A}$  se obtiene:

$$\mathbf{A}^{-1} \mathbf{XA} = \mathbf{D}.$$

siendo  $\mathbf{D}$  una matriz diagonal cuyos valores propios son iguales que los de la matriz de partida.

Este resultado nos conduce a la siguiente definición: Dos matrices  $\mathbf{A}$  y  $\mathbf{B}$  cuadradas del mismo orden son semejantes si y sólo si existe una matriz  $\mathbf{P}$

invertible que cumpla

$$\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{B}.$$

Es decir, si se tiene una matriz cuyos vectores propios son todos diferentes entre sí, existe una matriz diagonal semejante a ella que contiene los valores propios en su diagonal.

**Teorema 1.5.** *Si las matrices  $\mathbf{A}$  y  $\mathbf{B}$  son semejantes entonces se cumple:*

- $\mathbf{A}$  y  $\mathbf{B}$  tienen los mismos valores propios.
- $\text{traza}(\mathbf{A}) = \text{traza}(\mathbf{B})$ .
- $|\mathbf{A}| = |\mathbf{B}|$ .
- $\text{rango}(\mathbf{A}) = \text{rango}(\mathbf{B})$ .

*Demostración.*    ▪ Teniendo en cuenta que

$$\begin{aligned} |\mathbf{B} - \lambda\mathbf{I}| &= |\mathbf{P}^{-1}\mathbf{A}\mathbf{P} - \lambda\mathbf{P}^{-1}\mathbf{P}| \\ &= |\mathbf{P}^{-1}(\mathbf{A} - \lambda\mathbf{I})\mathbf{P}| \\ &= |\mathbf{P}^{-1}| |(\mathbf{A} - \lambda\mathbf{I})| |\mathbf{P}| \\ &= |(\mathbf{A} - \lambda\mathbf{I})| \end{aligned}$$

y que dos polinomios son iguales si y sólo si sus raíces son iguales, se tiene que los valores propios de  $\mathbf{A}$  y  $\mathbf{B}$  son iguales.

▪

$$\begin{aligned} \text{traza}(\mathbf{B}) &= \text{traza}(\mathbf{P}^{-1}\mathbf{A}\mathbf{P}) \\ &= \text{traza}(\mathbf{P}^{-1}\mathbf{P}\mathbf{A}) \end{aligned}$$

$$= \text{traza}(\mathbf{A}).$$

■

$$\begin{aligned} |\mathbf{B}| &= |\mathbf{P}^{-1}\mathbf{A}\mathbf{P}| \\ &= |\mathbf{P}^{-1}| |\mathbf{A}| |\mathbf{P}| \\ &= |\mathbf{A}|. \end{aligned}$$

- Teniendo en cuenta que el rango de una matriz no se altera si se multiplica dicha matriz por matrices no singulares:

$$\text{rango}(\mathbf{B}) = \text{rango}(\mathbf{P}^{-1}\mathbf{A}\mathbf{P}) = \text{rango}(\mathbf{A}).$$

□

**Teorema 1.6.** Si  $\mathbf{A}$  y  $\mathbf{B}$  son matrices semejantes de tal forma que  $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{B}$ , entonces se tiene que:

$$\mathbf{P}^{-1}\mathbf{A}^k\mathbf{P} = \mathbf{B}^k.$$

para cualquier entero positivo  $k$ . En particular, si  $\mathbf{A}$  es invertible se cumple:

$$\mathbf{P}^{-1}\mathbf{A}^1\mathbf{P} = \mathbf{B}^1.$$

*Demostración.* Para cualquier entero positivo  $k$  se cumple

$$\begin{aligned} \mathbf{B}^k &= (\mathbf{P}^{-1}\mathbf{A}\mathbf{P})(\mathbf{P}^{-1}\mathbf{A}\mathbf{P}) \dots (\mathbf{P}^{-1}\mathbf{A}\mathbf{P}) \\ &= \mathbf{P}^{-1}\mathbf{A}(\mathbf{P}\mathbf{P}^{-1})\mathbf{A}(\mathbf{P}\mathbf{P}^{-1}) \dots (\mathbf{P}\mathbf{P}^{-1})\mathbf{A}\mathbf{P} \\ &= \mathbf{P}^{-1}\mathbf{A}^k\mathbf{P}. \end{aligned}$$

Además como  $\mathbf{A}$  y  $\mathbf{B}$  son semejantes y la primera es invertible, entonces:

$$\begin{aligned}\mathbf{B}^{-1} &= (\mathbf{P}^{-1}\mathbf{A}\mathbf{P})^{-1} \\ &= \mathbf{P}^{-1}\mathbf{A}^{-1}(\mathbf{P}^{-1})^{-1} \\ &= \mathbf{P}^{-1}\mathbf{A}^{-1}\mathbf{P}.\end{aligned}$$

□

Dado que toda matriz cuadrada con valores propios diferentes entre sí es semejante a una matriz diagonal, esta propiedad proporciona una manera sencilla de calcular las potencias de dicha matriz y en particular su inversa sin más que aplicar las potencias a los valores propios situados en la diagonal.

**Teorema 1.7** (Lema de Schur). *Toda matriz cuadrada  $\mathbf{X}$  es semejante a una matriz triangular superior. Es decir,*

$$\mathbf{B}\mathbf{X}\mathbf{B} = \mathbf{T}$$

*Demostración.* La triangularización de  $\mathbf{X}$  se realiza por etapas. Para ello se tiene en cuenta que toda matriz de orden  $n \times n$  tiene al menos un autovalor  $\lambda_1$ , que se puede repetir  $n$  veces y para el cual se asegura la existencia de al menos un autovector asociado  $x_1$ .

En la primera etapa se construye una matriz  $\mathbf{B}_1$  para aplicar una transformación de semejanza sobre la matriz  $\mathbf{X}$ . Dicha matriz es de orden  $n \times n$  y contiene en su primera columna el autovector  $x_1$ , el resto de columnas se eligen de tal forma que los vectores sean linealmente independientes para garantizar la invertibilidad de la matriz  $\mathbf{B}_1$ .

$$\mathbf{B}_1 = (x_1, z_2, \dots, z_n) = (x_1, \mathbf{Z}_1)$$

Como por construcción,  $\mathbf{B}_1$  es invertible, se tiene que:

$$\mathbf{B}_1^{-1}\mathbf{B}_1 = (\mathbf{B}_1^{-1}x_1, \mathbf{B}_1^{-1}\mathbf{Z}_1) = (e^1, e^2, \dots, e^n)$$

Si se efectúa la transformación sobre  $\mathbf{X}$ , se obtiene:

$$\begin{aligned} \mathbf{B}_1^{-1}\mathbf{X}\mathbf{B}_1 &= \mathbf{B}_1^{-1}(\mathbf{X}x_1, \mathbf{X}\mathbf{Z}_1) = \mathbf{B}_1^{-1}(\lambda_1 x_1, \mathbf{X}\mathbf{Z}_1) \\ &= (\lambda_1 \mathbf{B}_1^{-1}x_1, \mathbf{B}_1^{-1}\mathbf{X}\mathbf{Z}_1) = (\lambda_1 e^1, \mathbf{B}_1^{-1}\mathbf{X}\mathbf{Z}_1) \\ &= \begin{pmatrix} \lambda_1 & \mathbf{C}_{11 \times (n-1)} \\ \theta_{(n-1) \times 1} & \mathbf{X}_{1(n-1) \times (n-1)} \end{pmatrix} \end{aligned}$$

Esta matriz tiene los mismos autovalores de  $\mathbf{X}$  ya que por construcción, son semejantes.  $\lambda_1$  es un autovalor de la transformada y el resto se pueden calcular a partir de la submatriz  $\mathbf{X}_1$ , que es de orden  $(n-1) \times (n-1)$ . Dicha matriz tiene por lo menos un autovalor  $\lambda_2$  que se puede repetir  $n-1$  veces, y para el cual es posible encontrar un autovector  $x_2$ .

En la segunda etapa, se construye una matriz  $\mathbf{B}_2$ , de orden  $(n-1) \times (n-1)$  que contenga en su primera columna el autovector  $x_2$  y el resto de columnas se eligen de tal forma que los vectores sean linealmente independientes para garantizar la invertibilidad de la matriz  $\mathbf{B}_2$ . Así se obtiene:

$$\mathbf{B}_2^{-1}\mathbf{X}_1\mathbf{B}_2 = \begin{pmatrix} \lambda_2 & \mathbf{C}_{21 \times (n-2)} \\ \theta_{(n-2) \times 1} & \mathbf{X}_{2(n-2) \times (n-2)} \end{pmatrix}$$

Pero lo que se pretende es aplicar transformaciones sobre  $\mathbf{X}$  en lugar de  $\mathbf{X}_1$ .

Para ello, se construye una matriz no singular  $\mathbf{B}_2^*$ , que efectúe transformaciones que modifiquen la estructura obtenida en la primera etapa. Dicha matriz es de la forma:

$$\mathbf{B}_2^* = \begin{pmatrix} 1 & \theta_{1 \times (n-1)} \\ \theta_{(n-1) \times 1} & \mathbf{B}_{2(n-1) \times (n-1)} \end{pmatrix}$$

La matriz  $\mathbf{B}_2^*$  es invertible, ya que las submatrices diagonales son no singulares y las no diagonales son nulas. Su inversa se puede calcular fácilmente sustituyendo  $\mathbf{B}_2$  por su inversa y manteniendo iguales el resto de submatrices. Por otro lado, al aplicar la transformación asociada con  $\mathbf{B}_2^*$  sobre la matriz obtenida en la primera etapa, se tiene:

$$\begin{aligned} (\mathbf{B}_2^*)^{-1} \mathbf{B}_1^{-1} \mathbf{X} \mathbf{B}_1 \mathbf{B}_2^* &= \begin{pmatrix} 1 & \theta \\ \theta & \mathbf{B}_2^{-1} \end{pmatrix} \begin{pmatrix} \lambda_1 & \mathbf{C}_1 \\ \theta & \mathbf{X}_1 \end{pmatrix} \begin{pmatrix} 1 & \theta \\ \theta & \mathbf{B}_2 \end{pmatrix} \\ &= \begin{pmatrix} \lambda_1 & \mathbf{C}_1 \mathbf{B}_2 \\ \theta & \mathbf{B}_2^{-1} \mathbf{X}_1 \mathbf{B}_2 \end{pmatrix} = \begin{pmatrix} \lambda_1 & \mathbf{C}_1 \mathbf{B}_2 \\ 0 & \lambda_2 & \mathbf{C}_2 \\ \theta & \theta & \mathbf{A}_2 \end{pmatrix} \end{aligned}$$

Esta matriz, por construcción, es semejante a la matriz original  $\mathbf{X}$  y consecuentemente tiene los mismos autovalores. En particular, es evidente que  $\lambda_1$  y  $\lambda_2$  son autovalores de la matriz transformada y el resto se pueden obtener a partir de la submatriz  $\mathbf{X}_2$ .

Repitiendo  $n-1$  veces este proceso, se construye la matriz  $\mathbf{B}$  que triangulariza  $\mathbf{X}$  como producto de las matrices de transformación calculadas en las sucesivas etapas:

$$\mathbf{B} = \mathbf{B}_1 \begin{pmatrix} 1 & \theta \\ \theta & \mathbf{B}_2 \end{pmatrix} \begin{pmatrix} \mathbf{I}_2 & \theta \\ \theta & \mathbf{B}_3 \end{pmatrix} \cdots \begin{pmatrix} \mathbf{I}_{n-2} & \theta \\ \theta & \mathbf{B}_{n-1} \end{pmatrix}$$

Siendo  $\mathbf{B}_{n-1}$  la matriz correspondiente a la transformación de la etapa  $(n-1)$ -ésima y que afecta a la submatriz  $\mathbf{X}_{n-2}$ :

$$\mathbf{B}_{n-1}^{-1} \mathbf{X}_{n-2} \mathbf{B}_{n-1} = \begin{pmatrix} \lambda_{n-1} & \mathbf{C}_{n-2} \\ 0 & \lambda_n \end{pmatrix}$$

Finalmente, al aplicar la transformación sobre la matriz de partida  $\mathbf{X}$  la transformación de semejanza asociada a la matriz  $\mathbf{B}$  se obtiene:

$$\mathbf{B}^{-1} \mathbf{X} \mathbf{B} = \begin{pmatrix} \lambda_1 & & \mathbf{F}_1 \\ 0 & \lambda_2 & \mathbf{F}_2 \\ \vdots & \vdots & \vdots \\ 0 & 0 & \lambda_n \end{pmatrix}$$

□

**Teorema 1.8.** Para cualquier matriz cuadrada  $\mathbf{X}$  de orden  $n$  se cumple:

- $|\mathbf{X}| = \prod_{i=1}^n \lambda_i$ .
- $\text{traza}(\mathbf{X}) = \sum_{i=1}^n \lambda_i$ .

*Demostración.* Según el lema de Schur, toda matriz cuadrada es semejante a una matriz triangular. Además toda matriz triangular tiene sus valores propios en la diagonal luego se cumple:

- $|\mathbf{X}| = \prod_{i=1}^n \lambda_i$ .
- $\text{traza}(\mathbf{X}) = \sum_{i=1}^n \lambda_i$ .

□

### 1.1.2. Valores y Vectores Propios de una Matriz Simétrica

Se dice que  $\mathbf{X}$  es una matriz simétrica si se cumple:

$$x_{ij} = x_{ji} \forall i \neq j.$$

**Teorema 1.9.** *Si  $x_1, x_2, \dots, x_s$  son vectores propios de una matriz simétrica asociados a los valores propios  $\lambda_1, \lambda_2, \dots, \lambda_s$  respectivamente con  $\lambda_i \neq \lambda_j \forall i \neq j$ , entonces son ortogonales entre sí.*

*Demostración.* Por el apartado anterior se sabe que los vectores propios de una matriz y su traspuesta son ortogonales entre sí. Por lo tanto, los vectores propios asociados con valores propios distintos también son ortogonales ya que la matriz es simétrica.  $\square$

**Teorema 1.10.** *Toda matriz simétrica  $\mathbf{X}$  es semejante a una matriz diagonal mediante transformaciones ortogonales.*

*Demostración.* Según el lema de Schur,  $\mathbf{X}$  es semejante a una matriz triangular mediante transformaciones ortogonales.

$$\mathbf{P}^\top \mathbf{X} \mathbf{P} = \mathbf{T}_\lambda.$$

Tomando traspuestas:

$$(\mathbf{P}^\top \mathbf{X} \mathbf{P})^\top = \mathbf{T}_\lambda^\top$$

$$\mathbf{P}^\top \mathbf{X}^\top \mathbf{P} = \mathbf{T}_\lambda^\top$$

pero como  $\mathbf{X}$  es simétrica se cumple que  $\mathbf{X} = \mathbf{X}^\top$  y como consecuencia también  $\mathbf{T}_\lambda^\top$  es simétrica y al ser también triangular sólo puede ser una matriz diagonal.  $\square$



**Teorema 1.11.** Si  $\mathbf{X}$  es una matriz simétrica de orden  $n$ , se puede expresar de la forma:

$$\mathbf{X} = \sum_{i=1}^n \lambda_i p_i p_i^\top$$

siendo  $p_i$  el vector propio asociado al valor propio  $\lambda_i \forall i = 1, \dots, n$ .

*Demostración.* Dado que  $\mathbf{X}$  es simétrica, se cumple:

$$\mathbf{P}^\top \mathbf{X} \mathbf{P} = \mathbf{D}_\lambda$$

Como  $\mathbf{P}$  es ortogonal se puede expresar:

$$\begin{aligned} \mathbf{X} &= \mathbf{P} \mathbf{D}_\lambda \mathbf{P}^\top \\ &= (\lambda_1 P_1, \lambda_2 P_2, \dots, \lambda_n P_n) = \begin{pmatrix} (P_1)^\top \\ (P_2)^\top \\ \vdots \\ (P_n)^\top \end{pmatrix} \end{aligned} \quad (1.2)$$

Con lo que:

$$\mathbf{X} = \lambda_1 P_1 P_1^\top + \lambda_2 P_2 P_2^\top + \dots + \lambda_n P_n P_n^\top.$$

□

**Teorema 1.12.** Si  $\mathbf{X}$  es una matriz simétrica e invertible, se puede calcular su inversa como:

$$\mathbf{X}^{-1} = \mathbf{P} \mathbf{D}^{-1} \mathbf{P}^\top.$$

*Demostración.* Dado que toda matriz simétrica es ortogonalmente semejante a una matriz diagonal se tiene:

$$\mathbf{P}^\top \mathbf{X} \mathbf{P} = \mathbf{D}.$$

Multiplicando la ecuación por  $\mathbf{P}$  y  $\mathbf{P}^\top$  a izquierda y derecha respectivamente,

$$\mathbf{PDP}^\top = \mathbf{X}.$$

Tomando inversas:

$$(\mathbf{X})^{-1} = (\mathbf{PDP}^\top)^{-1} = (\mathbf{P}^\top)^{-1}\mathbf{D}^{-1}\mathbf{P}^{-1}$$

Y por la ortogonalidad de  $\mathbf{P}$ :

$$\mathbf{X}^{-1} = \mathbf{PD}^{-1}\mathbf{P}^\top.$$

□

**Teorema 1.13.** *Si  $\mathbf{X}$  es una matriz de orden  $n \times p$  de rango  $r$  entonces  $\mathbf{X}^\top \mathbf{X}$  y  $\mathbf{X}\mathbf{X}^\top$  tienen los mismos vectores propios no nulos.*

*Demostración.* Las matrices  $\mathbf{X}^\top \mathbf{X}$  y  $\mathbf{X}\mathbf{X}^\top$  son simétricas de rango  $r$ . Si consideramos el valor propio no nulo  $\lambda_i$  de la matriz  $\mathbf{X}^\top \mathbf{X}$  se cumple:

$$\mathbf{X}^\top \mathbf{X}y_i = \lambda_i y_i$$

siendo  $y_i$  un vector propio de dicha matriz. Si multiplicamos por  $\mathbf{X}$ :

$$\mathbf{X}\mathbf{X}^\top(\mathbf{X}y_i) = \lambda_i(\mathbf{X}y_i).$$

Con lo que se deduce que  $\lambda_i$  es valor propio asociado a  $\mathbf{X}\mathbf{X}^\top$  si el vector  $\mathbf{X}y_i$  es no nulo. Pero si fuera nulo se tendría que  $\lambda_i y_i = \emptyset$  lo que no puede ser ya que  $\lambda_i$  se ha escogido como un valor propio no nulo y  $y_i$  su vector propio asociado. □

**Teorema 1.14.** Si  $\mathbf{X}$  es una matriz de rango  $r$  y  $\mathbf{V}_{p \times r}$  y  $\mathbf{U}_{n \times r}$  las matrices que contienen en sus columnas los vectores propios de  $\mathbf{X}^\top \mathbf{X}$  y  $\mathbf{X}\mathbf{X}^\top$  respectivamente asociados a los valores propios no nulos, entonces:

$$v_i = (1/\sqrt{\lambda_i})\mathbf{X}^\top u_i$$

$$u_i = (1/\sqrt{\lambda_i})\mathbf{X}v_i.$$

*Demostración.* En virtud de la demostración anterior se tienen las siguientes igualdades:

$$\mathbf{X}^\top \mathbf{X}v_i = \lambda_i v_i$$

$$\mathbf{X}\mathbf{X}^\top (\mathbf{X}v_i) = \lambda_i (\mathbf{X}v_i)$$

De aquí se puede deducir que si  $v_i$  es vector propio de la matriz  $\mathbf{X}^\top \mathbf{X}$  asociado al valor propio  $\lambda_i$ , entonces los vectores  $u_i = \alpha \mathbf{X}v_i \forall \alpha \neq 0 \forall i = 1, \dots, r$  son vectores propios de la matriz  $\mathbf{X}\mathbf{X}^\top$  asociados al valor propio  $\lambda_i$ .

De manera análoga se puede obtener que los vectores  $v_i = \beta \mathbf{X}^\top u_i \forall \beta \neq 0 \forall i = 1, \dots, r$  son vectores propios de la matriz  $\mathbf{X}^\top \mathbf{X}$  asociados con el valor propio  $\lambda_i$ .

Como los vectores  $v_i$  y  $u_i$  son normales:

$$\begin{aligned} 1 &= v_i^\top v_i = \beta^2 u_i^\top \mathbf{X}\mathbf{X}^\top u_i \\ &= \beta^2 u_i^\top \lambda_i u_i = \beta^2 \lambda_i \end{aligned}$$

Con lo que se tiene que  $\beta = 1/\sqrt{\lambda_i}$ . Si se hace el mismo razonamiento para  $u_i$  se obtienen las ecuaciones planteadas.  $\square$

**Teorema 1.15.** Sea  $\mathbf{X}$  una matriz de rango  $r$ ,  $\mathbf{V}^i$  y  $\mathbf{U}^i$  vectores propios

ortonormalizados de las matrices  $\mathbf{X}^\top \mathbf{X}$  y  $\mathbf{X}\mathbf{X}^\top$ , respectivamente, asociados con el valor propio común  $\lambda_i$  ( $\forall i = 1, 2, \dots, r$ ). Entonces:

$$\mathbf{X} = \sum_{i=1}^r \sqrt{\lambda_i} \mathbf{U}^i (\mathbf{V}^i)^\top$$

*Demostración.* Si se considera la relación definida en el teorema 1.14:

$$\mathbf{U}^i = \left(1/\sqrt{\lambda_i}\right) \mathbf{X}\mathbf{V}^i$$

Multiplicando por  $\sqrt{\lambda_i} (\mathbf{V}^i)^\top$  y aplicando sumatorio sobre  $i$ :

$$\sum_{i=1}^p \mathbf{X}\mathbf{V}^i (\mathbf{V}^i)^\top = \sum_{i=1}^p \sqrt{\lambda_i} \mathbf{U}^i (\mathbf{V}^i)^\top \quad (1.3)$$

Si se tiene en cuenta que los vectores propios  $\mathbf{V}^i$  son una base ortonormal de  $\mathbb{R}^p$ , se tiene que:

$$\sum_{i=1}^p \mathbf{V}^i (\mathbf{V}^i)^\top = \mathbf{V}\mathbf{V}^\top = \mathbf{I}_p$$

Entonces el primer término de la ecuación 1.3 se puede expresar como:

$$\sum_{i=1}^p \mathbf{X}\mathbf{V}^i (\mathbf{V}^i)^\top = \mathbf{X} \sum_{i=1}^p \mathbf{V}^i (\mathbf{V}^i)^\top = \mathbf{X}\mathbf{V}\mathbf{V}^\top = \mathbf{X}\mathbf{I}_p = \mathbf{X}$$

y así:

$$\mathbf{X} = \sum_{i=1}^p \sqrt{\lambda_i} \mathbf{U}^i (\mathbf{V}^i)^\top$$

y como la matriz  $\mathbf{X}$  es de rango  $r$ , quiere decir existen exactamente  $r$  valores propios distintos de cero y consecuentemente, la matriz  $\mathbf{X}$  se puede reconstruir a partir de dichos  $r$  valores propios y sus vectores propios asociados. De esta

manera se tiene que:

$$\mathbf{X} = \sum_{i=1}^r \sqrt{\lambda_i} \mathbf{U}^i (\mathbf{V}^i)^\top.$$

□

## 1.2. Aproximación de Matrices

El objetivo de este apartado es encontrar un procedimiento para representar vectores de  $\mathbb{R}^n$  sobre un espacio de dimensión menor, generalmente dos o tres.

En primer lugar, se va a explicar como se representan vectores de  $\mathbb{R}^2$  sobre el sistema de coordenadas formado por los vectores directores  $e_1 = (1, 0)$ ,  $e_2 = (0, 1)$ . Estos vectores forman lo que se denomina base canónica de  $\mathbb{R}^2$ . Sea  $y = (y_1, y_2)$  un vector perteneciente a  $\mathbb{R}^2$ . Dicho vector se puede descomponer de la siguiente manera:

$$\begin{aligned} y &= y_1 e_1 + y_2 e_2 \\ &= \pi_{y e_1} e_1 + \pi_{y e_2} e_2 \end{aligned}$$

donde  $\pi_{y e_i}$  es la proyección ortogonal de  $y$  sobre la recta  $r_{\emptyset e_i} \forall i = 1, 2$ . Dicha recta resulta de la unión del origen de coordenadas con el correspondiente vector director de la base canónica.

La distancia del origen de coordenadas a la proyección  $\pi_{y e_i}$  es:

$$d(\pi_{y r_{\emptyset e_i}}, \emptyset) = \sqrt{y_i^2} = y_i = \pi_{y e_i}.$$

Utilizando el teorema de Pitágoras se puede ver que la distancia del vector  $y$  al origen de coordenadas es:

$$d(y, \emptyset) = \sqrt{y_1^2 + y_2^2}.$$

Si se considera un sistema de coordenadas en el que los ejes estén generados por vectores directores ortonormales  $(u_1, u_2)$  distintos de la base canónica, el vector  $y$  se podrá expresar como la suma de los vectores directores cuyos coeficientes se calculan como las proyecciones del vector  $y$  sobre los ejes correspondientes. Es decir:

$$y = (y^\top u_1)u_1 + (y^\top u_2)u_2 = \pi_{yu_1}u_1 + \pi_{yu_2}u_2.$$

Del mismo modo que en el caso de la base canónica, se pueden calcular las distancias de las proyecciones al origen de coordenadas:

$$d(\pi_{y\theta e_i}, \emptyset) = y^\top u_i = \pi_{yu_i}$$

y la distancia del vector  $y$  al origen de coordenadas:

$$d(y, \emptyset) = \sqrt{\pi_{yu_1}^2 + \pi_{yu_2}^2}.$$

Con estas consideraciones se puede concluir que para representar gráficamente un vector sobre un sistema generado por vectores ortonormales de  $\mathbb{R}^2$  sólo es necesario conocer la proyección del vector  $y$  sobre cada uno de los ejes del sistema.

### 1.2.1. Subespacio de Mejor Ajuste

Si se parte de una matriz  $\mathbf{X}$  de orden  $n \times p$  que contiene información de  $n$  individuos observados sobre  $p$  variables se puede considerar las filas de la matriz como vectores en el espacio  $\mathbb{R}^p$  y las columnas como vectores de  $\mathbb{R}^n$ . De esta manera, si se considera el espacio de las filas, se puede buscar un espacio de dimensión  $q < p$  donde se representen los vectores fila de tal forma que su representación sea lo más parecida a su representación en el espacio original. Para resolver esta cuestión se utiliza el criterio de los mínimos cuadrados.

Se define el subespacio de mejor ajuste a la nube de puntos fila de dimensión  $q$  como el hiperplano  $P_q^p$  generado por vectores ortonormales que hace mínima la suma de cuadrados de las distancias entre los puntos y su proyección en el hiperplano:

$$\min_{P_q^p} \sum_{i=1}^n d^2(x_i, P_q^p).$$

El procedimiento para encontrar el subespacio de mejor ajuste se realiza por etapas. En primer lugar se busca la recta de mejor ajuste y a partir de ahí, se van añadiendo sucesivamente dimensiones hasta llegar al subespacio de dimensión deseada.

Para calcular la recta de mejor ajuste se utiliza el siguiente teorema:

**Teorema 1.16.** *Sea  $\mathbf{X}$  una matriz de datos de orden  $n \times p$ . La recta de  $\mathbb{R}^p$  que mejor se ajusta al espacio de las filas de dicha matriz es la recta  $r_v$  cuya dirección viene dada por el vector propio normalizado de la matriz  $\mathbf{X}^\top \mathbf{X}$  asociado a su mayor valor propio  $\lambda_1$ .*

*Demostración.* Por definición, la recta que mejor se ajusta a la nube de puntos fila tiene que cumplir el criterio de los mínimos cuadrados, es decir, que  $\sum_{i=1}^n d^2(x_i, r_{\theta v})$  sea mínima.

Pero basta con que lo cumpla su vector director:

$$\sum_{i=1}^n d^2(x_i, P_{x_i/v})$$

con  $v^\top v = 1$  para que sea ortonormal.

Según el Teorema de Pitágoras se tiene la relación:

$$d^2(x_i, \emptyset) = d^2(x_i, P_{x_i/v}) + d^2(P_{x_i/v}, \emptyset)$$

$$= d^2(x_i, P_{x_i/v}) + \pi_{x_i v}^2$$

Si se despeja la distancia del vector  $x_i$  a la recta y se sustituye la distancia del vector  $x_i$  al origen:

$$d^2(x_i, P_{x_i/v}) = x_i^\top x_i - \pi_{x_i v}^2.$$

Entonces la expresión a minimizar se puede escribir como:

$$\sum_{i=1}^n d^2(x_i, P_{x_i/v}) = \sum_{i=1}^n x_i^\top x_i - \sum_{i=1}^n \pi_{x_i v}^2.$$

Teniendo en cuenta que la expresión  $\sum_{i=1}^n x_i^\top x_i$  es constante, el problema se reduce a maximizar:

$$\begin{aligned} \sum_{i=1}^n \pi_{x_i v}^2 &= \sum_{i=1}^n (x_i^\top v)^2 \\ &= \sum_{i=1}^n v^\top x_i x_i^\top v \\ &= v^\top \left( \sum_{i=1}^n x_i x_i^\top \right) v \\ &= v^\top \mathbf{X}^\top \mathbf{X} v. \end{aligned}$$

Por lo tanto hay que maximizar la expresión  $v^\top \mathbf{X}^\top \mathbf{X} v$  con la restricción  $v^\top v = 1$ . Utilizando los multiplicadores de Lagrange:

$$L = v^\top \mathbf{X}^\top \mathbf{X} v - \lambda(v^\top v - 1).$$

Derivando respecto a  $v$  e igualando a cero:

$$\frac{\partial L}{\partial v} = 2\mathbf{X}^\top \mathbf{X} v - 2\lambda v = 0$$



de donde:

$$\mathbf{X}^\top \mathbf{X}v = \lambda v.$$

Con lo cual, el vector  $v$  es vector propio de la matriz  $\mathbf{X}^\top \mathbf{X}$  asociado al valor propio  $\lambda$ . Como se busca el vector que maximice la expresión, se elige  $\lambda$  como el mayor valor propio de la matriz  $\mathbf{X}^\top \mathbf{X}$ .  $\square$

**Teorema 1.17.** *Sea  $y$  un vector perteneciente a  $\mathbb{R}^n$  y  $P_2^n$  un plano generado por los vectores columna ortonormales de la matriz  $\mathbf{V} = (v_1, v_2)$ . Se tiene:*

- *La proyección ortogonal de  $y$  sobre  $P_2^n$  es:*

$$\Pi_{yP_2^n} = \pi_{yv_1}v_1 + \pi_{yv_2}v_2.$$

- *La distancia de la proyección ortogonal de  $y$  sobre el plano al origen de coordenadas es:*

$$d^2(\Pi_{yP_2^n}, \emptyset) = \pi_{yv_1}^2 + \pi_{yv_2}^2.$$

- *La distancia de  $y$  al plano es:*

$$d^2(y, P_2^n) = y^\top y - \pi_{yv_1}^2 - \pi_{yv_2}^2.$$

*Demostración.* ■ La proyección ortogonal de  $y$  sobre el plano generado por las columnas de la matriz  $\mathbf{V}$  es:

$$\Pi_{yP_2^n} = \mathbf{V}(\mathbf{V}^\top \mathbf{V})^{-1} \mathbf{V}^\top y$$

y como las columnas de  $\mathbf{V}$  son ortonormales se cumple que  $\mathbf{V}^\top \mathbf{V} = \mathbf{I}$  y se tiene:

$$\begin{aligned}\Pi_{yP_2^n} &= \begin{pmatrix} v_1 & v_2 \end{pmatrix} \begin{pmatrix} v_1^\top y \\ v_2^\top y \end{pmatrix} \\ &= \pi_{yv_1} v_1 + \pi_{yv_2} v_2.\end{aligned}$$

- La distancia de la proyección ortogonal al origen es:

$$\begin{aligned}d^2(\Pi_{yP_2^n}, \emptyset) &= \Pi_{yP_2^n}^\top \Pi_{yP_2^n} \\ &= \pi_{yv_1}^2 v_1^\top v_1 + \pi_{yv_2}^2 v_2^\top v_2 \\ &= \pi_{yv_1}^2 + \pi_{yv_2}^2.\end{aligned}$$

- La distancia del vector  $y$  al plano es por el teorema de Pitágoras:

$$d^2(y, P_2^n) = d^2(y, \emptyset) - d^2(\Pi_{yP_2^n}, \emptyset)$$

y por el apartado anterior:

$$= y^\top y - (\pi_{yv_1}^2 + \pi_{yv_2}^2).$$

□

Como extensión del teorema 1.2.1 se puede enunciar:

**Teorema 1.18.** *Sea una matriz de datos  $\mathbf{X}$  de orden  $n \times p$ . El subespacio de dimensión dos que mejor se ajusta al espacio de las filas de la matriz es el plano  $P_2^p$  cuyos vectores directores son los vectores propios ortonormales de la matriz  $\mathbf{X}^\top \mathbf{X}$  asociados con los dos mayores valores propios  $\lambda_1$  y  $\lambda_2$ .*

*Demostración.* Por el teorema 1.2.1 se sabe que el vector propio de  $\mathbf{X}^\top \mathbf{X}$  asociado

al mayor valor propio  $\lambda_1$  nos da la dirección de la recta de mejor ajuste. La segunda dirección  $v_2$  debe ser un vector normalizado y ortogonal al primero que denotaremos por  $v_1$ . El plano de mejor ajuste será aquel que cumpla:

$$P_2^p = \{z \in \mathbb{R}^p : z = \alpha_1 v_1 + \alpha_2 v_2, \text{ con } v_1^\top v_2 = 0 \text{ y } v_2^\top v_2 = 1\}.$$

Para que sea el de mejor ajuste tiene que garantizar además que la suma  $\sum_{i=1}^n d^2(x_i, P_2^p)$  sea mínima. Por definición la distancia de un punto a un plano es la distancia del punto a la proyección del punto en el plano. Por el teorema 1.2.1, esa distancia se puede expresar como:

$$\sum_{i=1}^n x_i^\top x_i - \sum_{i=1}^n \pi_{x_i v_1}^2 - \sum_{i=1}^n \pi_{x_i v_2}^2.$$

Teniendo en cuenta que:

- El primer término es una constante.
- El segundo término es  $\lambda_1$ .
- El tercer término es  $v_2^\top \mathbf{X}^\top \mathbf{X} v_2$ .

El problema se reduce a maximizar  $v_2^\top \mathbf{X}^\top \mathbf{X} v_2$  con las restricciones  $v_1^\top v_2 = 0$  y  $v_2^\top v_2 = 1$ . Por las propiedades variacionales de los valores propios (Horn y Johnson, 1985) se tiene que ese máximo es  $\lambda_2$ , el segundo mayor valor propio de la matriz  $\mathbf{X}^\top \mathbf{X}$ . Así queda definido el plano de mejor ajuste mediante los vectores propios de  $\mathbf{X}^\top \mathbf{X}$  asociados a los dos mayores valores propios.  $\square$

Análogamente se puede calcular el subespacio de mejor ajuste de dimensión  $q$  mediante los vectores propios de  $\mathbf{X}^\top \mathbf{X}$  asociados a los  $q$  mayores valores propios.

### 1.2.2. Descomposición en Valores Singulares

El siguiente paso en la aproximación de matrices es encontrar una descomposición de una matriz de datos  $\mathbf{X}$  en función de sus valores y vectores propios.

**Teorema 1.19.** *Si  $\mathbf{X}$  es una matriz de datos de orden  $n \times p$  y de rango  $r$ , se puede expresar:*

$$\mathbf{X} = \mathbf{U}_r \mathbf{\Lambda}_r \mathbf{V}_r^\top$$

donde  $\mathbf{U}_r$  y  $\mathbf{V}_r$  son matrices cuyas columnas son los vectores propios ortonormales de  $\mathbf{X}^\top \mathbf{X}$  y  $\mathbf{X} \mathbf{X}^\top$  asociados a los valores propios no nulos comunes y  $\mathbf{\Lambda}_r$  una matriz diagonal que contiene las raíces cuadradas de los valores propios de  $\mathbf{X}^\top \mathbf{X}$ .

*Demostración.* Utilizando la expresión definida en el teorema 1.15 para reconstruir la matriz  $\mathbf{X}$  a partir de los valores y vectores propios de las matrices  $\mathbf{X}^\top \mathbf{X}$  y  $\mathbf{X} \mathbf{X}^\top$ :

$$\mathbf{X} = \sum_{i=1}^r \sqrt{\lambda_i} \mathbf{U}^i (\mathbf{V}^i)^\top$$

y expandiendo:

$$\begin{aligned} \mathbf{X} &= \left( \sqrt{\lambda_1} \mathbf{U}^1, \sqrt{\lambda_2} \mathbf{U}^2, \dots, \sqrt{\lambda_r} \mathbf{U}^r \right) \begin{pmatrix} (\mathbf{V}^1)^\top \\ (\mathbf{V}^2)^\top \\ \vdots \\ (\mathbf{V}^r)^\top \end{pmatrix} \\ &= (\mathbf{U}^1, \mathbf{U}^2, \dots, \mathbf{U}^r) \begin{pmatrix} \sqrt{\lambda_1} & 0 & \dots & 0 \\ 0 & \sqrt{\lambda_2} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \sqrt{\lambda_r} \end{pmatrix} \begin{pmatrix} (\mathbf{V}^1)^\top \\ (\mathbf{V}^2)^\top \\ \vdots \\ (\mathbf{V}^r)^\top \end{pmatrix} \end{aligned}$$

$$= \mathbf{U}_r \mathbf{\Lambda}_r \mathbf{V}_r^\top$$

quedando así demostrado. □

Sea  $M_{n \times p}(\mathbb{R})$  el conjunto de las matrices reales de orden  $n \times p$ . Se denomina norma de Frobenius de una matriz a la función:

$$\| \mathbf{X} \|_F: M_{n \times p}(\mathbb{R}) \longrightarrow \mathbb{R}$$

definida por:

$$\| \mathbf{X} \|_F = \sqrt{\text{traza}(\mathbf{X}^\top \mathbf{X})} = \sqrt{\sum_{i=1}^p \lambda_i}$$

**Teorema 1.20.** *Sea  $\mathbf{X}$  una matriz de orden  $n \times p$  de rango  $r$  y  $\mathbf{Z}$  una matriz del mismo orden pero de rango  $k$  ( $k < r$ ) tal que  $\mathbf{Z} = \mathbf{U}_k \mathbf{\Lambda}_k \mathbf{V}_k^\top$ , entonces:*

$$\| \mathbf{X} - \mathbf{Y} \|_F = \sum_{i=k+1}^r \lambda_i$$

*Demostración.* Por la definición de la norma de Frobenius:

$$\| \mathbf{X} - \mathbf{Y} \|_F^2 = [(\mathbf{X} - \mathbf{Y})^\top (\mathbf{X} - \mathbf{Y})]$$

$$= \text{traza}(\mathbf{X}^\top \mathbf{X}) - 2\text{traza}(\mathbf{X}^\top \mathbf{Y}) + \text{traza}(\mathbf{Y}^\top \mathbf{Y})$$

Como:

$$\mathbf{U}_k \mathbf{\Lambda}_k \mathbf{V}_k^\top = (\mathbf{U}_k, \mathbf{U}_{r-k}) \begin{pmatrix} \mathbf{\Lambda}_k & \theta \\ \theta & \theta \end{pmatrix} \begin{pmatrix} (\mathbf{V}_k)^\top \\ (\mathbf{V}_{r-k})^\top \end{pmatrix}$$

y por la descomposición singular de la matriz  $\mathbf{X}$ , se tiene:

$$\begin{aligned}
\mathbf{X}^\top \mathbf{Y} &= (\mathbf{U}_r \mathbf{\Lambda}_r \mathbf{V}_r^\top)^\top (\mathbf{U}_k \mathbf{\Lambda}_k \mathbf{V}_k^\top) \\
&= (\mathbf{V}_r \mathbf{\Lambda}_r \mathbf{U}_r^\top) \mathbf{U}_k \mathbf{\Lambda}_k \mathbf{V}_k^\top = (\mathbf{U}_k, \mathbf{U}_{r-k}) \begin{pmatrix} \mathbf{\Lambda}_k & \theta \\ \theta & \theta \end{pmatrix} \begin{pmatrix} (\mathbf{V}_k)^\top \\ (\mathbf{V}_{r-k})^\top \end{pmatrix} \\
&= \mathbf{V}_r \mathbf{\Lambda}_r (\mathbf{U}_r^\top \mathbf{U}_r) \begin{pmatrix} \mathbf{\Lambda}_k & \theta \\ \theta & \theta \end{pmatrix} \mathbf{V}_k^\top \\
&= (\mathbf{V}_k, \mathbf{U}_{r-k}) \begin{pmatrix} \mathbf{\Lambda}_k & \theta \\ \theta & \mathbf{\Lambda}_{r-k} \end{pmatrix} \begin{pmatrix} \mathbf{\Lambda}_k & \theta \\ \theta & \theta \end{pmatrix} \begin{pmatrix} (\mathbf{V}_k)^\top \\ (\mathbf{V}_{r-k})^\top \end{pmatrix} \\
&= \mathbf{V}_k \mathbf{\Lambda}_k \mathbf{V}_k^\top = \mathbf{V}_k \mathbf{\Lambda}_k \mathbf{U}_k^\top \mathbf{U}_k \mathbf{\Lambda}_k \mathbf{V}_k^\top = \mathbf{Y}^\top \mathbf{Y}
\end{aligned}$$

Por lo tanto:

$$\|\mathbf{X} - \mathbf{Y}\|_F^2 = \text{traza}(\mathbf{X}^\top \mathbf{X}) - \text{traza}(\mathbf{Y}^\top \mathbf{Y})$$

$$= \sum_{i=1}^r \lambda_i - \sum_{i=1}^k \lambda_i = \sum_{i=k+1}^r \lambda_i$$

□

Sea  $\mathbf{X}_{n \times p}$  una matriz de rango  $r$  y  $\mathbf{X}_k = \mathbf{U}_k \mathbf{\Lambda}_k \mathbf{V}_k^\top$  una aproximación para dicha matriz. Se define como medida de la bondad de aproximación a la siguiente expresión:

$$\frac{\sum_{i=1}^k \lambda_i}{\sum_{i=1}^r \lambda_i}$$

**Teorema 1.21.** La aproximación  $\mathbf{X}_{n \times p}^*$  de rango  $k$  para la matriz  $\mathbf{X}_{n \times p}$  de rango

$r$  ( $k < r$ ) que minimiza la suma de cuadrados:

$$\| \mathbf{X} - \mathbf{X}^* \|^2 = \sum_{i=1}^n \sum_{j=1}^p (x_{ij} - x_{ij}^*)^2$$

está expresada por  $\mathbf{X}^* = \mathbf{U}_k \mathbf{\Lambda}_k \mathbf{V}_k^\top$  (Eckart y Young, 1936).

*Demostración.* Según el teorema 1.15, la matriz  $\mathbf{X}^*$  queda definida por los primeros  $k$  sumandos de la descomposición en valores singulares de la matriz  $\mathbf{X}$ . Consecuentemente, la suma de cuadrados de los errores de la aproximación queda definida por la suma de los últimos  $r - k$  valores propios de la matriz  $\mathbf{X}^\top \mathbf{X}$ :

$$\| \mathbf{X} - \mathbf{X}^* \|^2 = \sum_{i=k+1}^r \lambda_i$$

□





# Capítulo 2

## BIPLOT



## 2.1. Introducción y Revisión Bibliográfica

El análisis de matrices de datos de grandes dimensiones, individuos por variables, puede ser abordado mediante técnicas multivariantes, las cuales reducen la dimensionalidad proyectando los datos sobre un subespacio óptimo, conservando los patrones de similitud entre individuos y los patrones de covariación entre variables. Las diferencias entre estas técnicas dependen del tipo de variables y métricas utilizadas en los respectivos subespacios. Los métodos Biplot propuestos por Gabriel, 1971 son parte de estas técnicas, pero su extensión no ha sido tan rápida debido a la ausencia de software específico para su utilización. Los Biplot son una representación gráfica, en el contexto del análisis de Componentes Principales, representando conjuntamente una matriz de datos multivariante mediante marcadores fila (individuos) y marcadores columna (variables), permitiendo que las interrelaciones entre ellos puedan ser capturadas visualmente en un espacio de dimensión menor. Los Biplot permiten describir los datos pero también hacer modelos y diagnosis (Bradu y Gabriel, 1978) y son una herramienta de visualización potente que puede ser considerada como la versión multivariante del scatterplot dado que usualmente se representan en un espacio de dos dimensiones. Los Biplot clásicos de Gabriel, tienen dos fases. Por un lado, se aproxima la matriz de partida utilizando la descomposición en valores singulares y por otro, se factoriza para obtener un mapa euclídeo en baja dimensión mediante marcadores fila y columna representados por puntos/vectores. La interpretación en los métodos Biplot está basada en las propiedades geométricas del producto escalar entre filas y columnas lo que permite una aproximación de los elementos de la matriz de partida.

Gower y Harding, 1988, Gower, 1992 y Gower y Hand, 1996 proporcionan un enfoque diferente de los Biplot clásicos, ordenando los individuos, escalando

y luego superponiendo las variables, lo que permite una interpretación gráfica conjunta al igual que en los Biplot clásicos. Los métodos Biplot más utilizados son conocidos como GH y JK. Estos Biplot consiguen una óptima calidad de representación para variables y para individuos respectivamente. Galindo, 1986 prueba que, con una elección adecuada de los marcadores, es posible representar simultáneamente filas y columnas en el mismo espacio euclídeo con óptima calidad, denominando a este tipo HJ Biplot. Sus coordenadas para las columnas son los marcadores columna en el GH Biplot y las coordenadas para las filas son los marcadores fila en el JK Biplot. El HJ Biplot de Galindo ha sido aplicado en numerosos campos.

Orfao et al., 1988b aplicó los Biplot en histopatología; Orfao et al., 1988a en inmunología; Rivas-Gonzalo et al., 1993; Santos et al., 1991 en enología; Galante et al., 1991 en entomología; Galindo et al., 1996 en ecología; Dorado et al., 1999 en estudios agropecuarios; García-Talegón et al., 1999 en geología química; Iñigo et al., 2005 en ingeniería civil; González et al., 2006 en ciencias ambientales; Pinto, 2006 en sociología; Correa et al., 2007 en microbiología; Alcántara y Rivas, 2007 en ciencia política; Martínez-Ruiz et al., 2007 en ingeniería geológica; Cadaviz, 2008 en neuropsicología; Mendes et al., 2009 en limnología; Marreiros et al., 2010 en gestión hospitalaria; Castela y Galindo, 2010 en inferencia ecológica; Patino et al., 2011 en relaciones laborales; Vázquez et al., 2011 en estudios de seguridad; García et al., 2012 en ciencias pecuarias; Ochoa et al., 2012 en cultivo de cítricos; Vilorio et al., 2012 en biotecnología; Gallego-Álvarez et al., 2013 en datos ambientales; García-Sánchez et al., 2013 en estudios de corporación social; Serafim et al., 2012 en salud ambiental y toxicología; Díaz-Faes et al., 2013 en bibliometría; Vázquez-de Aldana et al., 2013 en nutrición vegetal; Gallego-Álvarez et al., 2014a en estudios de responsabilidad corporativa en Brasil; Gallego-Álvarez et al., 2014b,c en sostenibilidad; Morillo et al., 2014 en cienciometría; Felício y Galindo, 2014 en

gobierno corporativo.

Otro tipo de Biplot es el denominado GGE Biplot (genotype main effects (G) and genotype x environment interaction effects (GE) (Yan et al., 2000). Suele ser utilizado en el contexto de la evaluación de cultivos. Este tipo de datos se suelen presentar en tablas de dos vías de genotipos por ambientes y el objetivo es evaluar el rendimiento de diferentes cultivos en distintos ambientes. El modelo GGE aplica la descomposición en valores singulares a los datos a los que previamente se les ha sustraído el efecto ambiente, de manera que estos Biplot visualizan simultáneamente el efecto genotipo (G) y la interacción (GE), las cuales se consideran las dos principales fuentes de variación en el contexto de la evaluación de cultivos.

Detalles de GH, HJ y JK Biplot se muestran en la Sección 2.2 y del GGE Biplot en Frutos et al., 2014.

Ter Braak, 1986 utilizó Biplot ajustados con modelos lineales en el contexto del Biplot directo del gradiente, lo que permite que un conjunto de especies sea ordenado de acuerdo a sus relaciones con un conjunto de variables ambientales. Gauch, 1988 empleó los Biplot para validar y seleccionar modelos cuando se estudia la interacción entre genotipo y ambiente. Ter Braak, 1990 y Ter Braak y Looman, 1994 aprovecharon las relaciones entre los Biplot y los métodos de regresión para introducir el Biplot de la matriz de coeficientes de regresión y proponer un Biplot basado en regresión en rango reducido. Cárdenas y Galindo, 2003 investigaron los aspectos inferenciales de los Biplot utilizando los modelos bilineales generalizados, extendiendo sus ajustes con información externa para variables relativas a la familia exponencial.

Vairinhos, 2003 muestra que los Biplot son una base perfecta para el desarrollo de un sistema de minería de datos, dado que la mayoría de las técnicas pueden ser expresadas como casos particulares de los Biplot. Amaro

et al., 2004 estudiaron las propiedades de los MANOVA Biplot dentro del contexto de los modelos lineales generalizados multivariantes, desarrollando métodos para su interpretación. Hernández, 2005 estudió el funcionamiento de los Biplot en presencia de outliers y Ramírez et al., 2005 propusieron los Biplot para la detección de la multicolinealidad. Como alternativa al análisis de Correspondencias Múltiple en el caso de variables de presencia/ausencia con distribución binomial, Vicente-Villardón et al., 2006 consideraron los Biplot de predicción y los aplicaron a Biplot ajustados por regresión lineal generalizada, proponiendo los Biplot logísticos, extendidos posteriormente por Demey et al., 2008 y utilizados en Gallego-Álvarez y Vicente-Villardón, 2012 y recientemente en Vicente et al., 2015.

Bradú y Gabriel, 1974 y Bradú y Gabriel, 1978 estudiaron el ajuste de los modelos bilineales de tablas de dos vías, analizando la colinealidad entre filas y columnas en los Biplot. Gabriel y Zamir, 1979 también trabajaron el ajuste de dichos modelos bilineales pero propusieron procesos iterativos para obtener aproximaciones a bajo rango utilizando mínimos cuadrados ponderados. Denis, 1991, Falguerolles, 1995, Choulakian, 1996 y Gabriel et al., 1998 utilizaron los Biplot para estudiar interacciones en tablas de dos y tres vías. Gabriel et al., 1998 desarrollaron diagnosis en modelos basados en tablas de contingencia. Sepúlveda et al., 2008 utilizaron los Biplot como herramienta de diagnóstico para dependencia local en modelos de clases latentes.

Los métodos de análisis de datos de tres vías han sido presentados como una extensión del análisis de Componentes Principales de una supermatriz de dos vías, siendo las más comunes: (i) TUCKER3 (Tucker, 1966) y (ii) STATIS (L'Hermier des Plantes, 1976). En (i), se trabaja con componentes principales de tres vías y se genera una "core matrix" que nos permite explicar las interacciones entre las variables de cada modo. En (ii), se parte de matrices de operadores (matrices de

covarianzas por ejemplo) y se calcula la correlación vectorial entre operadores a partir del producto de Hilbert-Schmidt. A esta matriz se le aplica un análisis de Componentes Principales y cada uno de los operadores aparece representado como un punto en un espacio de Hilbert-Schmidt. Puntos próximos indican estructuras de covarianza similares en las matrices originales.

Basándose en los modelos TUCKER3, Carlier y Kroonenberg, 1996 generalizan la descomposición en valores singulares a tablas de tres vías proponiendo Biplot interactivos y Biplot conjuntos para capturar la información proveniente de los datos. La diferencia entre estos dos Biplot es la manera en que se trata la matriz inicial, en el Biplot interactivo se combinan dos modos mientras que en el Biplot conjunto se condiciona uno de los modos. Martín-Rodríguez et al., 2002 propusieron los Meta-biplots siguiendo la filosofía de los métodos Procrustes y las Meta-componentes Principales, permitiendo que los Biplot sean comparados para estudiar individuos y variables, como alternativa a los Biplot conjuntos e interactivos.

Vallejo-Arboleda et al., 2006 y Vallejo-Arboleda et al., 2008 propusieron el STATIS Canónico, un Biplot para datos multi-tabla. Frecuentemente, los datos multivariantes se encuentran recogidos sobre múltiples ocasiones produciendo un experimento multi-tabla. Ni el análisis separado de cada ocasión utilizando MANOVA o análisis canónico, ni el análisis conjunto de las tablas múltiples utilizando STATIS, son adecuados para capturar la estructura real de dichas matrices. Esto es debido a que unas técnicas recogen la estructura de grupo pero no la evolución en el tiempo mientras que las últimas confunden la variabilidad dentro y entre grupos. El STATIS Canónico permite recoger al mismo tiempo tanto la estructura de grupos como la evolución temporal, obteniendo variables canónicas estables a lo largo de las diferentes ocasiones o conjuntos de datos.

Nos hemos centrado en los Biplot clásicos de Gabriel, 1971; una revisión se

puede ver en Cárdenas et al., 2007. Sin embargo, Gower, 1992 propone otros tipos de Biplot basándose en la obtención de los marcadores columna a partir de la regresión multivariante. Gower y Harding, 1988 y Gower, 1992 proponen los Biplot no lineales. Gower y Hand, 1996 definen los Biplot de interpolación y predicción. Según Gower, los Biplot se refieren a la presentación de información de las variables y unidades, de una matriz de datos  $\mathbf{X}$ . El Biplot arquetipo está representado por los ejes de coordenadas cartesianas en el que los ejes calibrados (generalmente ortogonales) representan las variables y las unidades se representan como puntos. Hay dos preguntas que se relacionan con los ejes cartesianos (i) dado un caso de  $\mathbf{X}$ , que es una fila de  $\mathbf{X}$ , ¿dónde está el punto  $P$  correspondiente? y (ii) dado un punto  $P$  ¿cuáles son los valores asociados de  $\mathbf{X}$ ? Los Biplot estadísticos pueden diferir de esta definición clásica en varios sentidos. En primer lugar por lo general tienen sólo una aproximación a la representación cartesiana del espacio completo. Esto induce ejes no ortogonales y complica las respuestas a las preguntas (i) y (ii). En segundo lugar, podemos utilizar las métricas basadas en coeficientes de disimilaridad o en variantes de la distancia de Mahalanobis. Por lo tanto, se necesitan extensiones para abarcar metodologías donde  $\mathbf{X}$  esté representado por diversas formas de escalamiento multidimensional métrico y no métrico. En tercer lugar, es posible que sea necesario incluir variables categóricas en  $\mathbf{X}$ , tanto nominales como ordinales. Por lo tanto es necesaria una extensión del sistema cartesiano para poder representar las categorías.

El desarrollo de los Biplot en el sentido de Gower muestra cómo manejar todas estas situaciones de una manera unificada. La proyección ortogonal es un concepto clave en el uso de ejes cartesianos, el concepto más general del punto más cercano a un conjunto se utiliza para manejar las generalizaciones. Para variables cuantitativas utiliza ejes calibrados, posiblemente no lineales, y para las variables categóricas, conjuntos de puntos marcados que representan a los diferentes niveles



de categoría. Los casos siguen siendo representados por puntos. En Gower y Hand, 1996, Greenacre, 2010 y Gower et al., 2011 se pueden consultar más detalles.

## 2.2. Formulación Teórica

### 2.2.1. Representaciones Biplot

Una matriz de datos  $\mathbf{X}$  de orden  $I \times J$  puede ser expresada como producto de dos matrices:  $\mathbf{A}$  con  $I$  filas y  $S$  columnas y  $\mathbf{B}$  con  $S$  filas y  $J$  columnas. Si  $S$  es igual a dos, cada fila de  $\mathbf{A}$  y cada columna de  $\mathbf{B}$  son las coordenadas de un punto en un gráfico de dos dimensiones. Cuando las  $I$  filas de  $\mathbf{A}$  y las  $J$  columnas de  $\mathbf{B}$  se representan en un único gráfico, se denomina Biplot.

Por lo tanto, un Biplot es una representación gráfica de una matriz  $\mathbf{X}_{I \times J}$  mediante marcadores fila  $\mathbf{a}_1, \dots, \mathbf{a}_I$  y marcadores columna  $\mathbf{b}_1, \dots, \mathbf{b}_J$ , elegidos de tal forma que el producto interno  $\mathbf{a}_i^\top \mathbf{b}_j$  aproxime el elemento  $x_{ij}$  de  $\mathbf{X}$  lo mejor posible. Las filas y las columnas de estas matrices de marcadores son las coordenadas de los puntos en un espacio euclídeo respecto de los mismo ejes ortogonales. Una propiedad del Biplot es que cada uno de los  $I \times J$  valores puede ser recuperado inspeccionando los  $I + J$  puntos representados en el gráfico.

La descomposición de una matriz  $\mathbf{X}$  en las componentes  $\mathbf{A}$  y  $\mathbf{B}$  se denomina descomposición en valores singulares y se obtienen como resultado  $S$  componentes. Una matriz de dos vías raramente suele tener rango dos, por lo tanto, aproximar  $\mathbf{X}$  por una matriz de rango dos significa que solamente las dos primeras componentes van a ser utilizadas para su representación. Si dichas componentes explican una proporción importante de la variabilidad total de la matriz  $\mathbf{X}$ , entonces es posible aproximar la matriz original por la matriz de rango dos y mostrar la información en un Biplot.

Sea  $\mathbf{X}$  una matriz de datos compuesta por  $I$  individuos sobre los que se han medido  $J$  variables. La descomposición en valores singulares de  $\mathbf{X}$  se define como:

$$\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top,$$

donde:

$\mathbf{U}$  es una matriz cuyos vectores columna son ortonormales y corresponden a los vectores propios de  $\mathbf{X}\mathbf{X}^\top$ ,

$\mathbf{V}$  es una matriz cuyos vectores columna son también ortonormales y se corresponden con los vectores propios de  $\mathbf{X}^\top\mathbf{X}$ ,

$\mathbf{\Lambda}$  es una matriz diagonal que contiene los valores singulares de la matriz  $\mathbf{X}$ , que son las raíces cuadradas no negativas de los valores propios de  $\mathbf{X}^\top\mathbf{X}$ , ordenados de manera decreciente.

Para que las matrices  $\mathbf{U}$  y  $\mathbf{V}$  sean ortonormales debe cumplirse que  $\mathbf{U}^\top\mathbf{U} = \mathbf{V}^\top\mathbf{V} = \mathbf{I}$ . Esta propiedad asegura la unicidad de la factorización.

Un elemento de  $\mathbf{X}$  se puede escribir genéricamente como

$$x_{ij} = \sum_{s=1}^{\min(I,J)} \lambda_s u_{is} v_{js} \quad (2.1)$$

Para encontrar la aproximación en un espacio de baja dimensión de la matriz  $\mathbf{X}$  es necesario minimizar la distancia entre la matriz original  $\mathbf{X} = x_{ij}$  y la matriz aproximada  $\tilde{\mathbf{X}} = \tilde{x}_{ij}$ . Esta distancia (Euclídea) entre dos matrices se define como:

$$d(\mathbf{X}, \tilde{\mathbf{X}}) = \sqrt{\sum_{i=1}^I \sum_{j=1}^J (x_{ij} - \tilde{x}_{ij})^2}.$$

El teorema de Eckart y Young, 1936, que también puede encontrarse en otros autores como Young, 1938, Gabriel, 1971 o Greenacre, 1984 demuestra que la mejor aproximación  $S$ -dimensional de la matriz  $\mathbf{X}$  en el sentido de los mínimos

cuadrados se puede obtener mediante la descomposición en valores singulares de la matriz  $\mathbf{X}$  sumando únicamente los  $S$  primeros términos en la ecuación 2.1.

Los primeros  $S$  elementos de  $\mathbf{u}_s$  y de  $\mathbf{v}_s$  combinados con los valores singulares  $\lambda_s$  de diferentes formas son utilizados como las coordenadas para mostrar gráficamente los datos. Los tipos más comunes de Biplot se muestran en la Figura 2.1.

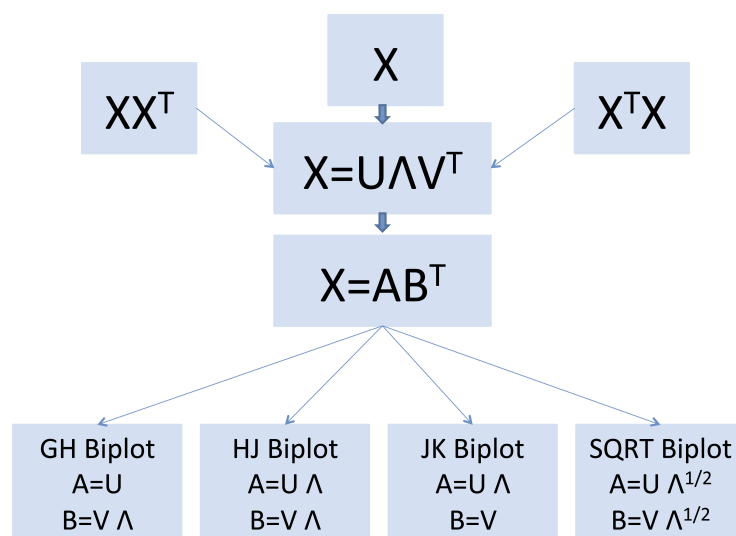


Figura 2.1: Tipos de Biplot.

### 2.2.2. Interpretaciones Geométricas

En un Biplot, los marcadores columna  $\mathbf{b}_j$  se representan como vectores y los marcadores fila  $\mathbf{a}_i^T$  como puntos. Este tipo de representación facilita la proyección de los marcadores fila sobre los marcadores columna. Y esta proyección es, precisamente, la que nos permite estudiar las relaciones entre individuos y variables, es decir,  $x_{ij} \approx \mathbf{a}_i^T \mathbf{b}_j$  implica

$$x_{ij} \approx \|\text{proy } \mathbf{a}_i / \mathbf{b}_j\| \|\text{signo } \mathbf{b}_j\| \|\mathbf{b}_j\|$$

donde:

- $\| \text{proy } \mathbf{a}_i / \mathbf{b}_j \|$  es la longitud del segmento que une el origen con el punto  $\mathbf{a}_i$  (longitud de la proyección desde  $\mathbf{a}_i$  hasta  $\mathbf{b}_j$ )
- signo  $\mathbf{b}_j$  es el signo de  $\mathbf{b}_j$
- $\| \mathbf{b}_j \|$  es el módulo de  $\mathbf{b}_j$  (longitud del segmento que une el origen con  $\mathbf{b}_j$ )

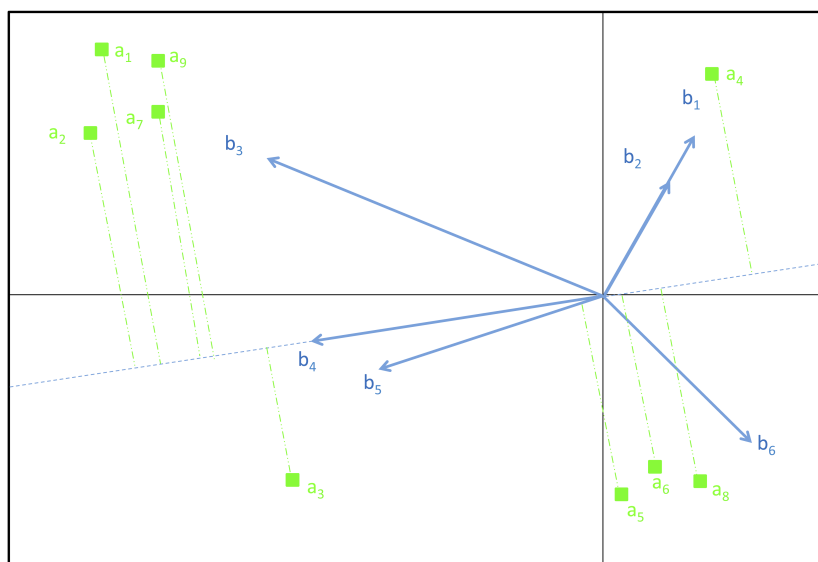


Figura 2.2: Representación Biplot de una matriz que contiene información de 6 variables medidas sobre 9 individuos.

Esto significa que  $x_{ij}$  se aproxima por el módulo de la proyección de  $\mathbf{a}_i$  sobre  $\mathbf{b}_j$  multiplicado por la longitud de  $\mathbf{b}_j$ , con el signo correspondiente. El sentido del vector  $\mathbf{b}_j$  muestra el sentido en el que se incrementan los valores de la variable representada por dicho vector. Las proyecciones de los puntos  $\mathbf{a}_i$  sobre un vector columna aproximan la  $j$ -ésima columna de  $\mathbf{X}$  y proporcionan una manera de ordenar los individuos respecto a esa variable. Una vez definida la forma de representación en un Biplot es posible explicar su interpretación. Así:

- La distancia entre puntos puede ser interpretada como disimilaridad entre los correspondientes individuos, especialmente si están bien representados. Los individuos que están muy lejos unos de otros tienen una mayor distancia Euclídea entre ellos y viceversa. En la Figura 2.2, la mayor distancia se observa entre los individuos  $\mathbf{a}_1$  y  $\mathbf{a}_8$  y la menor distancia se obtiene para los individuos  $\mathbf{a}_5$  y  $\mathbf{a}_6$ .
- En el JK Biplot, la longitud del vector aproxima la variabilidad de la variable correspondiente. Por lo tanto, cuanto mayor longitud, mayor variabilidad. Siguiendo con el ejemplo representado en la Figura 2.2, la variable  $\mathbf{b}_3$  posee la mayor variabilidad entre todas las variables, mientras que la variable  $\mathbf{b}_2$  tiene la menor. El coseno del ángulo entre dos vectores aproxima la correlación entre las variables que representan. Así, ángulos cercanos a 90 (o 270) grados indican que la correlación entre las variables representadas por ellos es prácticamente inexistente. Un ángulo de 0 o 180 grados refleja una correlación de 1 o  $-1$ , respectivamente. El Biplot representado en la Figura 2.2 muestra una relación fuerte entre las variables  $\mathbf{b}_4$  y  $\mathbf{b}_5$ , y una relación débil entre las variables  $\mathbf{b}_2$  y  $\mathbf{b}_3$ , y de  $\mathbf{b}_1$  con  $\mathbf{b}_3$ . La correlación entre las variables  $\mathbf{b}_3$  y  $\mathbf{b}_6$  es negativa. Las variables con la misma dirección indican la presencia de multicolinealidad, tal como se puede observar en la Figura 2.2 para las variables  $\mathbf{b}_1$  y  $\mathbf{b}_2$ . También, es posible detectar outliers multivariantes y cluster entre individuos, en la Figura 2.2 se puede observar el grupo formado por los individuos  $\mathbf{a}_1$ ,  $\mathbf{a}_2$ ,  $\mathbf{a}_7$  y  $\mathbf{a}_9$ .
- Las relaciones entre individuos y variables se pueden interpretar en términos de producto escalar, es decir, a través de las proyecciones de los puntos sobre los vectores. Esto nos permite identificar qué variables son las responsables de las diferencias entre grupos de individuos. Si la proyección está cerca

del origen, el valor de esa observación es aproximadamente la media de la variable en cuestión. En consecuencia, cuanto más se aleje la proyección de un individuo en el sentido del vector que representa la variable observada, más se alejará ese individuo de la media de la variable y viceversa. Observando la Figura 2.2, el individuo  $\mathbf{a}_2$  destaca con el mayor valor para la variable  $\mathbf{b}_4$ , seguido por  $\mathbf{a}_1$ ,  $\mathbf{a}_7$  y  $\mathbf{a}_9$ .

- En el HJ Biplot, la búsqueda de variables que diferencian individuos se realiza mediante la interpretación de los ejes factoriales, es decir, las nuevas variables que son combinaciones lineales de las variables originales y las relaciones entre dichas nuevas variables con las observadas.
- La medida de las relaciones entre los ejes de los Biplot y cada una de las variables observadas se llama contribución relativa del factor al elemento, representa la proporción de variabilidad de cada una de las variables explicada por cada factor. Esta medida se interpreta como el coeficiente de determinación en regresión. De hecho, si los datos están centrados, es el coeficiente de determinación en la regresión de cada variable sobre el eje correspondiente. Las contribuciones relativas permiten detectar qué variables están más relacionadas con cada eje y, por consiguiente, nos permite saber qué variables son responsables del orden de la proyección de los individuos sobre cada eje. Debido a que los ejes se construyen de manera que sean independientes, las contribuciones relativas de cada eje a cada variable son independientes y por lo tanto, es posible calcular la contribución relativa de un plano sumando las contribuciones relativas de los ejes que lo forman.

En las representaciones Biplot también es posible mostrar combinaciones lineales de filas y columnas. Ejemplos de estas combinaciones son las medias

de filas y de columnas.

- Las medias de las filas estarán ordenadas como las proyecciones de los marcadores fila sobre el vector que representa al marcador columna medio, puesto que:

$$x_{\bullet j} \approx a_{\bullet}^{\top} b_j \approx \| \text{proy } b_j / a_{\bullet} \| \text{ signo } a_{\bullet} \| a_{\bullet} \| = \| \text{proy } a_{\bullet} / b_j \| \text{ signo } b_j \| b_j \| .$$

- Las medias de las columnas estarán ordenadas como las proyecciones de los marcadores columna sobre el vector que representa al marcador fila medio, ya que:

$$x_{i \bullet} \approx a_i^{\top} b_{\bullet} \approx \| \text{proy } a_i / b_{\bullet} \| \text{ signo } b_{\bullet} \| b_{\bullet} \| = \| \text{proy } b_{\bullet} / a_i \| \text{ signo } a_i \| a_i \| .$$

- La media total es la proyección del marcador fila medio sobre el vector que representa el vector columna promedio, es decir:

$$x_{\bullet \bullet} \approx a_{\bullet}^{\top} b_{\bullet} \approx \| \text{proy } a_{\bullet} / b_{\bullet} \| \text{ signo } b_{\bullet} \| b_{\bullet} \| = \| \text{proy } b_{\bullet} / a_{\bullet} \| \text{ signo } a_{\bullet} \| a_{\bullet} \| .$$

### 2.2.3. Propiedades de los marcadores

Las propiedades de los marcadores dependen del tipo de Biplot elegido. Estas propiedades son fundamentales a la hora de elegir el tipo de Biplot que se va utilizar así como la interpretación posterior de los resultados.

**Propiedades de los marcadores en el JK Biplot** En este Biplot, se utiliza la métrica  $\mathbf{B}^{\top} \mathbf{B} = \mathbf{I}$  en el espacio de las filas de la matriz  $\mathbf{X}$ , de tal manera que:

- Los productos escalares de los individuos de  $\mathbf{X}$  con la métrica identidad son

los productos escalares de los marcadores fila de la matriz  $\mathbf{A}$  en el espacio completo.

$$\mathbf{X}\mathbf{X}^\top = \mathbf{A}\mathbf{A}^\top.$$

Dado que  $\mathbf{X} = \mathbf{A}\mathbf{B}^\top$  y  $\mathbf{B}^\top\mathbf{B} = \mathbf{I}$ ,

$$\mathbf{X}\mathbf{X}^\top = \mathbf{A}\mathbf{B}^\top\mathbf{B}\mathbf{A}^\top = \mathbf{A}\mathbf{A}^\top$$

- La distancia Euclídea entre dos individuos de  $\mathbf{X}$  y la distancia Euclídea entre los marcadores fila en el espacio completo son la misma, es decir,

$$(x_i - x_j)^\top(x_i - x_j) = (a_i - a_j)^\top(a_i - a_j).$$

Como  $x_i = \mathbf{B}a_i$ ,

$$(x_i - x_j)^\top(x_i - x_j) = (\mathbf{B}a_i - \mathbf{B}a_j)^\top(\mathbf{B}a_i - \mathbf{B}a_j) = (a_i - a_j)^\top\mathbf{B}^\top\mathbf{B}(a_i - a_j) = (a_i - a_j)^\top(a_i - a_j).$$

- Los marcadores fila y las coordenadas de los individuos son iguales en el espacio de las componentes principales.

Si  $\boldsymbol{\psi}$  es la matriz que contiene las coordenadas de los individuos en el espacio de las componentes principales entonces  $\boldsymbol{\psi} = \mathbf{A}$ .

$$\boldsymbol{\psi} = \mathbf{X}\mathbf{V} = (\mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^\top)\mathbf{V} = \mathbf{U}\boldsymbol{\Lambda} = \mathbf{A}.$$

- Las coordenadas de las columnas de la matriz  $\mathbf{X}$  son las proyecciones de los ejes originales sobre el espacio de componentes principales. La proyección



de cada marcador fila sobre los marcadores columna es una aproximación de los valores de los individuos sobre las variables correspondientes.

- El producto escalar de marcadores columna es el producto escalar de las columnas de la matriz  $\mathbf{X}$  con la métrica  $\mathbf{X}\mathbf{X}^\top$ . Es decir,

$$x_j^\top (\mathbf{X}\mathbf{X}^\top) x_j = b_j^\top b_j.$$

$$x_j^\top (\mathbf{X}\mathbf{X}^\top) x_j =$$

$$b_j^\top \mathbf{A}^\top (\mathbf{X}\mathbf{X}^\top) \mathbf{A} b_j =$$

$$b_j^\top \mathbf{A}^\top (\mathbf{A}\mathbf{A}^\top) \mathbf{A} b_j =$$

$$b_j^\top b_j.$$

- La similaridad entre columnas se mide utilizando la inversa de la matriz de dispersión de los individuos. Es imposible interpretar los ángulos en términos de correlación debido a:

$$(x_i - x_j)^\top (\mathbf{X}\mathbf{X}^\top)^{-1} (x_i - x_j) = (b_i - b_j)^\top (b_i - b_j)$$

- La calidad de representación es mejor para las filas que para las columnas.

**Propiedades de los marcadores en el GH Biplot** En este Biplot, se impone la métrica  $\mathbf{A}^\top \mathbf{A} = \mathbf{I}$ , de tal forma que:

- Los productos escalares entre las columnas de  $\mathbf{X}$  son los productos escalares

entre los marcadores columnas.

$$\mathbf{X}^\top \mathbf{X} = \mathbf{B}\mathbf{B}^\top.$$

Dado que  $\mathbf{X} = \mathbf{A}\mathbf{B}^\top$ ,

$$\mathbf{X}^\top \mathbf{X} = \mathbf{B}\mathbf{A}^\top \mathbf{A}\mathbf{B}^\top = \mathbf{B}\mathbf{B}^\top.$$

- Si la matriz  $\mathbf{X}$  ha sido centrada por columnas, el cuadrado de la longitud de los vectores que representan los marcadores columna aproximan la covarianza entre las correspondientes variables. Como consecuencia de esta propiedad se derivan las tres propiedades siguientes:

- La longitud al cuadrado del vector que representa un marcador columna aproxima la varianza de la variable correspondiente y la longitud aproxima la desviación estándar.

$$\|b_j\| = \|x_j\| = \sqrt{\text{Var}(x_j)}.$$

- El coseno del ángulo formado por dos marcadores columna aproxima la correlación entre las variables correspondientes.

$$\cos(b_i, b_j) = \text{corr}(x_i, x_j).$$

- La distancia Euclídea entre dos variables es la distancia entre los correspondientes marcadores columna.

$$\begin{aligned} d^2(x_i, x_j) &= (x_i - x_j)^\top (x_i - x_j) = \|x_i\|^2 + \|x_j\|^2 - 2(x_i^\top x_j) = \|b_i\|^2 \\ &+ \|b_j\|^2 - 2(b_i^\top b_j) = d^2(b_i, b_j). \end{aligned}$$

- Las coordenadas en  $\mathbf{B}$  son la importancia de las variables sobre los ejes principales.
- La distancia de Mahalanobis entre dos individuos se puede aproximar por la distancia Euclídea entre los marcadores fila, es decir,  $(\mathbf{x}_i - \mathbf{x}_j)^\top \hat{\Sigma}^{-1} (\mathbf{x}_i - \mathbf{x}_j) = (\mathbf{a}_i - \mathbf{a}_j)^\top (\mathbf{a}_i - \mathbf{a}_j)$ , donde  $\hat{\Sigma}$  es una estimación de la matriz de varianzas-covarianzas correspondiente.  $x_i$  puede ser escrito como  $\mathbf{B}a_i$ , entonces:

$$\begin{aligned} (x_i - x_j)^\top \hat{\Sigma}^{-1} (x_i - x_j) &= \\ (\mathbf{B}a_i - \mathbf{B}a_j)^\top \hat{\Sigma}^{-1} (\mathbf{B}a_i - \mathbf{B}a_j) &= \\ (a_i - a_j)^\top \mathbf{B}^\top (\mathbf{B}^\top \mathbf{B})^{-1} \mathbf{B} (a_i - a_j) &= (a_i - a_j)^\top (a_i - a_j). \end{aligned}$$

- Si  $\mathbf{X}$  está centrada por columnas, las coordenadas para los marcadores fila son las coordenadas de los individuos en el espacio de las componentes principales y  $\mathbf{A}$  contiene los pesos en las componentes principales estandarizadas.
- Los productos escalares de los marcadores fila son iguales que los productos escalares de las filas de la matriz  $\mathbf{X}$  con la métrica  $(\mathbf{X}^\top \mathbf{X})^{-1}$  en el espacio de las columnas.

$$\mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top = \mathbf{A} \mathbf{A}^\top.$$

Dado que  $\mathbf{X} = \mathbf{A} \mathbf{B}^\top$ ,

$$\mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top = \mathbf{A} \mathbf{B}^\top (\mathbf{B}^\top \mathbf{B})^{-1} \mathbf{B} \mathbf{A}^\top = \mathbf{A} \mathbf{A}^\top.$$

- La calidad de representación es mejor para las columnas que para las filas.

**Propiedades de los marcadores en el HJ Biplot** En este Biplot, las propiedades de los marcadores fila son las mismas que las explicadas en el JK Biplot, mientras que las propiedades de los marcadores columna son las mismas que las explicadas en el GH Biplot. Las reglas para interpretar el HJ Biplot son una combinación de las reglas utilizadas en los Biplot clásicos, en el análisis de Correspondencias, en el análisis Factorial y en el Multidimensional Scaling. Específicamente, se tiene que:

- Las distancias entre marcadores fila se interpretan como inversas de similitudes, en este sentido, cuanto más próximos estén dos individuos, serán más similares, lo que permite identificar cluster de individuos con perfiles similares.
- Las longitudes de los vectores columna aproximan las desviaciones estándar de las variables.
- Los cosenos de los ángulos entre vectores columna aproximan las correlaciones entre las variables. Por lo tanto, ángulos muy agudos indican alta correlación positiva entre las variables; ángulos obtusos cercanos a 180 grados están asociados con variables altamente correlacionadas negativamente; y ángulos rectos indican variables no correlacionadas; análogamente los cosenos de los ángulos entre los marcadores columna y las componentes principales aproximan las correlaciones entre ellos, mientras que en el caso de datos estandarizados aproximan las cargas factoriales en el análisis factorial.
- El orden de las proyecciones ortogonales de los marcadores fila sobre un marcador columna aproxima el orden de los elementos en dicha columna. Así, cuanto más lejos esté la proyección de un punto (individuo) del centro

de gravedad (media de las coordenadas de los puntos), más lejano será el valor que toma dicho individuo de la media de la variable correspondiente.

- Marcadores fila y columna se pueden representar en el mismo sistema de referencia con óptima calidad de representación. En el contexto del análisis de Correspondencias, Greenacre, 1984 y Lebart et al., 1995 demuestran que las nubes de puntos fila y columna tienen los mismos valores propios y consecuentemente, existen relaciones baricéntricas entre ellos. Las relaciones propuestas por Galindo, 1986 son similares, es decir, las relaciones entre los vectores propios  $\mathbf{U}$  y  $\mathbf{V}$  son  $\mathbf{U} = \mathbf{XV}\mathbf{\Lambda}^{-1}$  y  $\mathbf{V} = \mathbf{X}^{\top}\mathbf{U}\mathbf{\Lambda}^{-1}$ . Por lo tanto, los marcadores se pueden expresar como:

$$\mathbf{A} = \mathbf{U}\mathbf{\Lambda} = \mathbf{XV}\mathbf{\Lambda}^{-1}\mathbf{\Lambda} = \mathbf{XV} = \mathbf{XX}^{\top}\mathbf{U}\mathbf{\Lambda}^{-1} = \mathbf{XB}\mathbf{\Lambda}^{-1}$$

y

$$\mathbf{B} = \mathbf{V}\mathbf{\Lambda} = \mathbf{X}^{\top}\mathbf{U}\mathbf{\Lambda}^{-1}\mathbf{\Lambda} = \mathbf{X}^{\top}\mathbf{U} = \mathbf{X}^{\top}\mathbf{XV}\mathbf{\Lambda}^{-1} = \mathbf{X}^{\top}\mathbf{A}\mathbf{\Lambda}^{-1}$$

Es decir, las coordenadas para las filas se obtienen como medias ponderadas de las columnas donde los pesos son los valores de  $\mathbf{X}$  y lo mismo aplica para las columnas.

#### 2.2.4. Bondad de ajuste

Para evaluar la calidad de la aproximación  $S$ -dimensional es necesario saber qué cantidad de la variabilidad original contenida en la matriz  $\mathbf{X}$  es explicada por la matriz aproximada  $\tilde{\mathbf{X}}$ . La variabilidad total de una matriz es la suma de todos sus elementos al cuadrado:

$$\text{Variabilidad Total} = \|\mathbf{X}\|^2 = \sum_{i=1}^I \sum_{j=1}^J x_{ij}^2$$

Por las propiedades de la descomposición en valores singulares, esta variabilidad se puede descomponer en una parte explicada y una parte residual:

$$\|\mathbf{X}\|^2 = \|\tilde{\mathbf{X}}\|^2 + \|\mathbf{X} - \tilde{\mathbf{X}}\|^2.$$

Usando la ortonormalidad de  $\mathbf{U}$  y  $\mathbf{V}$ , podemos expresar esta ecuación en términos de los cuadrados de los valores singulares:

$$\sum_{s=1}^{\tilde{S}} \lambda_s^2 = \sum_{s=1}^S \lambda_s^2 + \sum_{s=S+1}^{\tilde{S}} \lambda_s^2.$$

Esta ecuación muestra que la suma de los  $S$  primeros vectores singulares al cuadrado dividido por la suma total de los cuadrados de los valores singulares proporciona una forma de evaluar la cantidad de variabilidad total explicada por los  $S$  primeros vectores. Si la cantidad es grande, indica que el gráfico construido por los  $S$  primeros vectores singulares da una buena representación de la estructura de la matriz de partida. Si solo una pequeña parte de la variabilidad es explicada por los primeros vectores singulares hay que tener en cuenta que parte de la estructura de la matriz puede estar representada en dimensiones superiores. Si los datos están centrados por variables, individuos situados cerca del origen de coordenadas pueden tener valores próximos a las medias de las variables o su variabilidad es explicada en otras dimensiones. Del mismo modo, las variables situadas cerca del origen pueden tener una variabilidad pequeña o pueden no estar bien representadas en esas dimensiones.

Para calcular la calidad de representación para columnas se parte de la matriz

de varianzas-covarianzas.

$$\hat{\Sigma} = \mathbf{V}\Lambda\mathbf{U}^{\top}\mathbf{U}\Lambda\mathbf{V}^{\top}.$$

En el caso de la representación GH Biplot, la métrica utilizada es:  $\mathbf{A}^{\top}\mathbf{A} = \mathbf{U}^{\top}\mathbf{U} = \mathbf{I}$ . Por lo tanto, la matriz de varianzas-covarianzas viene dada por:

$$\hat{\Sigma} = \mathbf{V}\Lambda^2\mathbf{V}^{\top}.$$

Es decir, la suma de los cuadrados de los elementos de  $\hat{\Sigma}$  es  $\sum_{s=1}^{\tilde{S}} \lambda^4$ . Por lo que la calidad de representación para las columnas de la aproximación  $S$ -dimensional de la matriz  $\mathbf{X}$  se calcula como:

$$\frac{\sum_{s=1}^S \lambda^4}{\sum_{s=1}^{\tilde{S}} \lambda^4}.$$

La calidad de representación para las filas es:  $S/\tilde{S}$  ya que la suma de cuadrados de los elementos de  $\mathbf{X}\hat{\Sigma}^{-1}\mathbf{X}^{\top}$  sobre el espacio de las filas de  $\mathbf{X}$  es igual a  $\tilde{S}$  y el de su aproximación es  $S$ . En el caso de la representación JK Biplot, la calidad de representación para columnas es:  $S/\tilde{S}$  y la calidad de representación para las filas es:

$$\frac{\sum_{s=1}^S \lambda^4}{\sum_{s=1}^{\tilde{S}} \lambda^4}.$$

En el caso de la representación HJ Biplot, la calidad de representación tanto para filas como para columnas es óptima e igual a:

$$\frac{\sum_{s=1}^S \lambda^4}{\sum_{s=1}^{\tilde{S}} \lambda^4}.$$

En la Tabla 2.1 se resume la obtención de los marcadores fila y columna para cada tipo de Biplot así como la calidad de representación en cada caso.

	Filas		Columnas	
	Coordenadas	Calidad	Coordenadas	Calidad
GH Biplot	$U$	$\frac{S}{\tilde{S}}$	$V\Lambda$	$\frac{\sum_{s=1}^S \lambda_s^4}{\sum_{s=1}^{\tilde{S}} \lambda_s^4}$
JK Biplot	$U\Lambda$	$\frac{\sum_{s=1}^S \lambda_s^4}{\sum_{s=1}^{\tilde{S}} \lambda_s^4}$	$V$	$\frac{S}{\tilde{S}}$
HJ Biplot	$U\Lambda$	$\frac{\sum_{s=1}^S \lambda_s^4}{\sum_{s=1}^{\tilde{S}} \lambda_s^4}$	$V\Lambda$	$\frac{\sum_{s=1}^S \lambda_s^4}{\sum_{s=1}^{\tilde{S}} \lambda_s^4}$

Tabla 2.1: Obtención de marcadores y sus calidades de representación.

### 2.2.5. Contribuciones

Las calidades de representación explicadas anteriormente en la subsección 2.2.4 son una forma de evaluar globalmente los ajustes de la aproximación pero también es posible medir el ajuste de individuos y variables a nivel individual, lo cual es importante a la hora de interpretar los resultados. Estas medidas individuales están basadas en los conceptos de contribución (absoluta y relativa) (Galindo, 1986 y Jambu, 1991). La inercia total es la suma de los valores propios de una matriz, es decir, la traza de la matriz, que se utiliza como una medida de la variabilidad total de una matriz de datos. Está directamente relacionada con el concepto físico de inercia, que es la tendencia de un objeto en movimiento a permanecer en movimiento, y de un objeto en reposo a seguir en reposo.

Se tiene:

Varianza total de la nube de individuos =



Varianza total de la nube de variables =

$$\begin{aligned} \text{traza}(\mathbf{X}\mathbf{X}^\top) &= \text{traza}(\mathbf{X}^\top\mathbf{X}) = \\ &= \sum_{s=1}^S \lambda_s^2 = \\ &= \sum_{j=1}^J d^2(b_j, 0) = \sum_{s=1}^S \sum_{j=1}^J b_{js}^2 = \\ &= \sum_{i=1}^I d^2(a_i, 0) = \sum_{s=1}^S \sum_{i=1}^I a_{is}^2. \end{aligned}$$

Las contribuciones absolutas de los individuos para la varianza del eje  $s$ :

$$CAE_i F_s = a_{is}^2.$$

Las contribuciones absolutas de las variables para la varianza del eje  $s$ :

$$CAE_j F_s = b_{js}^2.$$

La inercia total del factor  $s$  considerando las contribuciones de los individuos:

$$\sum_{i=1}^I a_{is}^2 = \lambda_s^2.$$

La inercia total del factor  $s$  considerando las contribuciones de las variables:

$$\sum_{j=1}^J b_{js}^2 = \lambda_s^2.$$

La contribución relativa del elemento  $i$  al factor  $s$ :

$$CRE_i F_s = \frac{CAE_i F_s}{\lambda_s^2}.$$

La contribución relativa del elemento  $j$  al factor  $s$ :

$$CRE_j F_s = \frac{CAE_j F_s}{\lambda_s^2}.$$

La contribución relativa del factor  $s$  al elemento  $i$ :

$$CRF_s E_i = \frac{a_{is}^2}{d^2(a_i, 0)} = \cos^2(\alpha_i).$$

La contribución relativa del factor  $s$  al elemento  $j$ :

$$CRF_s E_j = \frac{a_{js}^2}{d^2(b_j, 0)} = \cos^2(\beta_j).$$

Las contribuciones relativas del elemento al factor evalúan en qué medida ese factor puede ser explicado por ese individuo o esa variable.

Las contribuciones relativas del factor al elemento evalúan en qué medida el significado de ese factor puede estar relacionado con el significado de ese individuo o esa variable.

Puede ocurrir que un elemento (individuo o variable) sea el que más contribuya a un factor pero que no sea el más interesante para su interpretación.

### 2.2.6. Software para Biplot

La mayoría del software disponible para hacer Biplot se ha desarrollado para aplicaciones específicas, o como una parte dentro de un paquete con un propósito más general, por lo que no son muy flexibles y producen gráficos estáticos que limitan la interpretación de los resultados. Por ello, Vicente-Villardón, 2003 implementó en el software comercial Matlab ([www.mathworks.com/products/matlab](http://www.mathworks.com/products/matlab)), un programa para realizar Biplot llamado **multBiplot**. Este programa contiene

gran cantidad de técnicas multivariantes, entre ellas se encuentran los Biplot Clásicos, HJ Biplot y los Biplot Logísticos. También es posible realizar análisis Factorial, análisis de Correspondencias Simple y Múltiple, análisis de Coordenadas Principales, Escalado Multidimensional y Unfolding, Biplot Canónico/ MANOVA Biplot para una y dos vías, análisis de la Redundancia, análisis Canónico de Correspondencias, Coinercia o Unfolding Restringido. Si se dispone de matrices de tres vías, el programa también tiene implementado diferentes técnicas para su análisis como el Meta-Biplot, el STATIS y el STATIS dual, el análisis Factorial Múltiple, el Doble PCA, el análisis Triádico Parcial o los Biplot Consenso. Es posible obtener Biplot para datos binarios, ordinales y nominales. En presencia de datos faltantes el programa permite aplicar distintas formas de imputación. Tiene la posibilidad además de transformar, recodificar o convertir las variables de las que se dispone. No es un programa estático, ya que es posible cambiar diferentes opciones gráficas tanto a priori como a posteriori. Permite además realizar transformaciones a los datos previas al propio análisis. Una vez que se ha obtenido el gráfico con la representación en dimensión reducida se tienen diferentes opciones para cambiar su aspecto visual. Además, para algunas técnicas es posible realizar técnicas de clustering a los datos y representarlas en el propio gráfico. Toda la información que aporta cada técnica se guarda en un archivo de texto para poder hacer una interpretación conjunta del gráfico con dicha información.

Existen otros softwares como el **GGEbiplot**, cuya principal funcionalidad es el GGE Biplot ([www.ggebiplot.com](http://www.ggebiplot.com)) aunque también tiene la posibilidad de generar Biplot clásicos. El programa **GGEbiplot** es un software comercial ampliamente utilizado por agrónomos, científicos agrícolas y genetistas (Yan y Kang, 2003 y Frutos et al., 2014). Respecto al entorno R (R-Team, 2014) se ha realizado una revisión de los paquetes en los que se han implementado descomposiciones y/o representaciones Biplot. En la tabla 2.2 se han recogido

información de cada uno de ellos. En dicha tabla, se presenta el nombre del paquete, la aproximación en la que se basa, es decir, Gabriel, 1971, Galindo, 1986 o Gower, 1992, las principales referencias, las fechas de creación y última actualización así como los principales contenidos y funcionalidades.

Paquete	Contenido	Creación
Método		Actualización
Referencia		
<b>BiplotGUI</b>	Proporciona una interfaz gráfica de usuario (GUI) para construir e interactuar con Biplot y muestra variables como ejes calibrados. Por lo tanto, no es posible interpretar las longitudes de las variables. Permite cambiar el título, mostrar o esconder etiquetas y puntos, cambiar el tipo, color y tamaño de líneas y fuentes, el color y la orientación de etiquetas y marcas de los ejes, dibujar convex-hulls y alpha bags cuando haya grupos definidos en los datos. También es posible ejecutar otro tipo de Biplot (no lineales y multidimensional scaling) y permite elegir la distancia a utilizar y la forma de calcular las coordenadas (coordenadas principales, matriz de covarianzas/correlaciones). Muestra las correlaciones de las variables y proporciona gráficos interactivos en 3 dimensiones.	13-08-08 19-03-13
Gower (La Grange et al., 2009, 2013)		
<b>bpca</b>	Muestra gráficos de Biplot en 2 y 3 dimensiones (interactivo), proporciona la longitud de las variables y las correlaciones y los ángulos entre ellas, coordenadas para individuos y variables, valores y vectores propios y calidad de representación. Presenta un gráfico con las correlaciones y sus aproximaciones.	17-08-08 24-11-13
Gabriel Galindo (Faria y Demetrio, 2012)		

Sigue en la página siguiente

Paquete	Contenido	Creación
Método		Actualización
Referencia		
<b>GGEbiplotGUI</b> Gabriel Galindo Yang (Frutos y Galindo, 2013; Frutos et al., 2014; Yan et al., 2000; Yan y Kang, 2003)	Es una GUI para construir e interaccionar con GGE Biplot. Proporciona valores propios, % de variabilidad explicada por cada uno de ellos, coordenadas para individuos y variables, contribuciones de factores a elementos. También, permite cambiar el color de fondo, las etiquetas de los genotipos y de los ambientes, el título y la fuente. Es posible mostrar genotipos y ambientes así como esconder el título, los ejes y los símbolos. Además, con la interfaz gráfica es posible mover las etiquetas con el ratón y cambiar su color y texto. También tiene la posibilidad de realizar gráficos interactivos en 3 dimensiones.	29-08-11 27-04-14
<b>multibiplotGUI</b> Gabriel Galindo (Nieto et al., 2012)	Proporciona una GUI con la que construir e interaccionar con Biplot Múltiples. Permite obtener calidades de representación, bondad de ajuste, contribuciones, valores propios y tiene la posibilidad de seleccionar el número de ejes que se van a retener. Muestra gráficos en 2 y 3 dimensiones (en 2 dimensiones se puede mover o eliminar etiquetas, cambiar el color, tamaño y símbolo de los puntos y seleccionar los ejes que se van a mostrar; en 3 dimensiones es posible rotar y hacer zoom).	29-10-12
<b>calibrate</b> Gower (Graffelman, 2012)	Calibra vectores de variables en Biplot y scatterplots, dibujando marcas a lo largo del vector y etiquetando esas marcas con los valores específicos. La calibración óptima se encuentra utilizando mínimos cuadrados. Una calibración no óptima se puede encontrar si se especifica un factor de calibración.	21-01-06 10-09-13
<b>nominallogisticBiplot</b> Gabriel, Galindo (Hernández y Vicente-Villardón, 2013a)	Produce una matriz de análisis de elementos politómicos utilizando Biplot logísticos nominales, extendiendo el Biplot logístico binario para datos nominales politómicos.	05-01-14
<b>biplotstats</b> Gabriel (R-Team, 2014)	Forma parte del paquete básico de <b>R</b> y produce un Biplot a partir de la salida de <b>princomp</b> o <b>prcomp</b> .	25-09-13

Sigue en la página siguiente

Paquete		Creación
Método	Contenido	Actualización
Referencia		
<b>ordinallogisticBiplot</b>		
Gabriel, Galindo	Produce una matriz de análisis de elementos politómicos utilizando Biplot logísticos ordinales, extendiendo el Biplot logístico binario para datos ordinales politómicos.	30-10-13 26-11-13
(Hernández y Vicente-Villardón, 2013b)		
<b>dynBiplotGUI</b>		
Gabriel, Galindo	Es una GUI para resolver Biplot dinámicos, clásicos y HJ con matrices de dos y tres vías.	04-11-13 25-04-15
(Egido, 2014)		

Tabla 2.2: Biplot en R.

En la Tabla 2.3, se proporciona una revisión de los paquetes de **R** que mencionan la palabra “biplot”, aunque dicha mención se refiere a la representación conjunta de coordenadas en un mismo gráfico obtenidas con otros métodos en lugar de mediante la descomposición Biplot.

Paquete		Creación
Método	Contenido	Actualización
Referencia		
<b>vegan</b>	Proporciona herramientas para la descripción de comunidades ecológicas. Este paquete tiene las funciones básicas para análisis de diversidad, ordenación de comunidades y disimilaridad. Además, muestra Biplot con los resultados provenientes de análisis de la redundancia, correlación canónica y canónico de correspondencias. Dichos análisis pueden ser utilizados con otro tipo de datos.	06-09-01 12-01-15
(Oksanen et al., 2013)		
<b>ade4</b>	Se caracteriza por la implementación de funciones gráficas y estadísticas, disponibilidad de datos numéricos y redacción de documentación técnica y temática. Se incluyen referencias bibliográficas y tiene funciones para mostrar Biplot de resultados del análisis implementado.	10-12-02 14-04-15
(Chessel et al., 2013, 2004; Dray y Dufour, 2007; Dray et al., 2007)		
<b>ade4TkGUI</b>	Es una Tcl/Tk GUI para varias funciones básicas del paquete <b>ade4</b> .	29-09-06 13-11-12
(Thioulose y Dray, 2007, 2012)		

Sigue en la página siguiente

Paquete	Contenido	Creación
Método		Actualización
Referencia		
<b>ca</b>		
(Greenacre y Nenadic, 2012; Nenadic y Greenacre, 2007)	Ejecuta y visualiza análisis de correspondencias simple, múltiple y conjunto y muestra Biplot con los resultados de dichos análisis.	28-07-07 31-12-14
<b>caGUI</b>		
(Markos, 2012)	Es una Tcl/Tk GUI para las funciones del paquete <b>ca</b> .	04-10-09 29-10-12
<b>ThreeWay</b>		
(Del Ferraro et al., 2013)	Permite realizar análisis de componentes para las matrices de datos de 3 vías por medio de modelos CANDECOMP / PARAFAC, Tucker1, Tucker2 y Tucker3 y muestra Biplot conjuntos de los resultados de modelos Tucker3.	29-10-12 09-04-14

Tabla 2.3: Paquetes de R que aluden a la palabra Biplot.

A pesar del amplio abanico de aplicaciones prácticas de los métodos Biplot, reciben críticas al considerarlos como métodos exploratorios. Esto es debido a que los resultados proporcionados por estos métodos son estimaciones puntuales calculadas a partir de una única muestra. Debido a que esta crítica puede ser generalizada al resto de técnicas de análisis multivariante, se han desarrollado versiones inferenciales para algunas de ellas. Uno de los métodos más utilizados para proporcionar versiones inferenciales son los métodos Bootstrap (Efron, 1979; Efron y Tibshirani, 1993). Gifi, 1990; Greenacre, 1984; Meulman, 1982 introdujeron estos métodos en el contexto de las dos vías o el análisis de Correspondencias Múltiple (MCA); Chatterjee, 1984; Lambert et al., 1990, 1991 lo utilizan en el contexto del análisis Factorial; en el caso del análisis de Componentes Principales (PCA) Daudin et al., 1988; Diaconis y Efron, 1983; Holmes, 1985, 1989; Stauffer et al., 1985 utilizaron la metodología Bootstrap para proponer intervalos de confianza para los puntos representados en el subespacio de los ejes principales intentando resolver el problema de la elección del número de ejes a retener; Milan y Whittaker, 1995 lo proponen en el caso de modelos

bilineales que incorporan Descomposición en Valores Singulares (dvs); (Raykov y Little, 1999) se valen de los métodos Bootstrap para evaluar el ajuste de las rotaciones Procrustes; Linting et al., 2007 en el PCA no lineal; Timmerman et al., 2009 utilizan los métodos Bootstrap para estimar intervalos de confianza en el análisis de Componentes Multinivel (MLSCA). Recientemente, Fisher et al., 2014 desarrollan una metodología para utilizar remuestreo bootstrap en el contexto del PCA. En este caso se dispone de un número de variables muy alto y a la vez mayor que el número de individuos. Para este propósito se han desarrollado paquetes en R tales como `bootSVD` (Fisher, 2015) o `Exposition` (Beaton et al., 2014).

Siguiendo la idea de utilizar inferencia en el análisis multivariante, se presenta una versión inferencial de los métodos Biplot basada en la utilización de los métodos Bootstrap (Efron, 1979; Efron y Tibshirani, 1993).

### 2.3. Bootstrap sobre Biplot

La teoría estadística intenta responder a tres preguntas básicas:

1. ¿Cómo debo recoger mis datos?
2. ¿Cómo debo recoger y analizar los datos recogidos?
3. ¿Con qué precisión se han analizado y resumido los datos recogidos?

La tercera cuestión corresponde a lo que se conoce como inferencia estadística.

La distribución muestral de un estadístico es la distribución que resulta de considerar todas las muestras posibles que pueden ser tomadas de una población. La mayoría de los procedimientos estadísticos requieren conocer la distribución muestral del estadístico que se va a utilizar en el análisis posterior. El conocimiento necesario depende del tipo de análisis que se va a realizar.



Por ejemplo, para la construcción de intervalos de confianza y test de hipótesis se necesita el conocimiento de la distribución muestral o de los percentiles de dicha distribución. Por otro lado, cuando se tiene un problema de estimación, es fundamental tener indicadores de la precisión de los estimadores puesto que un estimador debe tener un error de estimación. Por lo tanto, es necesario el conocimiento de medidas de precisión como la varianza, el sesgo o el error cuadrático medio de un estimador. Estas medidas dependen de la distribución muestral del estimador y pueden ser también utilizadas para elegir el mejor estimador a partir de un conjunto apropiado de ellos.

La distribución muestral de un estadístico y sus características dependen normalmente de la población subyacente y por lo tanto no son conocidas. Deben ser estimadas o aproximadas a partir de los datos observados. El Jackknife y el Bootstrap son dos métodos utilizados para estimar o aproximar la distribución muestral de un estadístico y sus características.

Desde un enfoque tradicional, una medida de precisión se estima mediante analogía empírica de una fórmula explícita teórica de la medida de precisión. Este enfoque tiene algunas desventajas y debilidades que se han ido descubriendo a lo largo de los años:

- A veces se requiere de un tamaño de muestra muy grande para poder calcular las medidas de precisión de los estimadores.
- La fórmula teórica o su aproximación se basa en un modelo. Cuando el modelo es ligeramente incorrecto, el estimador de la precisión no puede ser correcto.
- Para aplicar este enfoque a diversos problemas es necesario desarrollar una fórmula teórica para cada problema.

- A veces, este desarrollo es demasiado costoso e incluso imposible.
- La fórmula teórica puede ser demasiado complicada para ser útil en el proceso de estimación de la medida de precisión.

Quenouille, 1949 introdujo un método, posteriormente denominado Jackknife, para estimar el sesgo de un estimador prescindiendo de una observación cada vez del conjunto original de datos y recalculando el estimador a partir del resto de los datos. Sea  $T_n = T_n(X_1, \dots, X_n)$  un estimador de un parámetro desconocido  $\theta$ . El sesgo de  $T_n$  se define como:

$$\text{sesgo}(T_n) = E(T_n) - \theta.$$

Sea  $T_{n-1,i} = T_{n-1}(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n)$  el estadístico pero calculado sobre  $n - 1$  observaciones  $X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n$ ,  $i = 1, \dots, n$ . El sesgo jackknife del estimador es:

$$\text{sesgo}_{jack} = (n - 1)(\bar{T}_n - T_n),$$

donde  $\bar{T}_n = \frac{1}{n} \sum_{i=1}^n T_{n-1,i}$ . Esto conduce a un sesgo reducido del estimador jackknife de  $\theta$ ,

$$T_{jack} = T_n - \text{sesgo}_{jack} = nT_n - (n - 1)\bar{T}_n.$$

El Jackknife se ha convertido en una herramienta valiosa desde que Tukey, 1958 encontrara que se puede utilizar también para construir estimadores de la varianza.

El Jackknife es menos dependiente de suposiciones del modelo y no necesita una fórmula teórica como sucede en el caso del enfoque tradicional. Sin embargo, necesita calcular el estadístico  $n$  veces.

### 2.3.1. Bootstrap

Un conjunto de datos de tamaño  $n$  contiene  $2^n - 1$  subconjuntos no vacíos. El método Jackknife únicamente utiliza  $n$  de ellos. El desarrollo tecnológico de las últimas décadas ha sido muy rápido. Esa velocidad y la potencia de las nuevas generaciones de ordenadores han favorecido el desarrollo de nuevos métodos estadísticos que son más fiables y que tienen aplicaciones más amplias. Los métodos Bootstrap (Efron, 1979) es uno de estos métodos.

El Bootstrap se basa en la elección de sucesivas muestras aleatorias de tamaño  $n$ . Dada una población con  $N$  unidades, se define una muestra aleatoria simple de tamaño  $n$  al conjunto de  $n$  unidades escogidas de la población de partida en la que cada unidad del 1 a la  $N$  tiene una probabilidad de ser escogida de  $1/N$ . Este tipo de muestreo se realiza con reposición, es decir, cada unidad de la población puede ser elegida más de una vez.

Los métodos Bootstrap se utilizan para estudiar el sesgo y el error estándar de estimadores con la ventaja de calcular el sesgo y el error estándar de una manera automática sin tener en cuenta lo complicado que pueda resultar el cálculo del estimador.

#### Estimador Bootstrap del Error Estándar

Sea  $X_1, \dots, X_n$  vectores aleatorios de tamaño  $p$  procedentes de una distribución muestral desconocida  $F$  y  $\theta = T(F)$ . La siguiente cuestión que se plantea es cómo de preciso es dicho estimador. Los métodos Bootstrap fueron creados en 1979 (Efron, 1979) con el objetivo de calcular el error estándar de un estimador. Es un método completamente automático y puede ser utilizado con cualquier estimador independientemente de la complejidad matemática del mismo.

Los métodos Bootstrap están basados en la idea de *muestra bootstrap*. Si se

parte de una muestra aleatoria  $X_1, \dots, X_n$ , una muestra bootstrap  $X_1^*, \dots, X_n^*$  es una muestra aleatoria simple con reposición de  $n$  unidades procedentes de la muestra de partida. Si se define  $\hat{\theta}_n = T(\hat{F})$  como el estimador del parámetro de interés y  $\hat{F}$  el estimador de la distribución muestral desconocida  $F$ , se puede denominar *réplica bootstrap* a  $\hat{\theta}_n^* = T(X_1^*, \dots, X_n^*)$ , que es el cálculo del estimador a partir de la muestra bootstrap. Para calcular la precisión del estimador  $\hat{\theta}_n$  se calcula el estimador bootstrap del error estándar de  $\hat{\theta}_n$ ,  $se_{\hat{F}}(\hat{\theta}_n^*)$  que es el error estándar de  $\hat{\theta}_n$  calculado a partir de las réplicas bootstrap. Esta estimación se suele llamar estimador bootstrap ideal del error estándar de  $\hat{\theta}_n$ . Para conseguir una buena aproximación al valor numérico de dicho error se utiliza el algoritmo Bootstrap definido a continuación:

1. Se seleccionan  $B$  muestras bootstrap independientes  $X_{1b}^*, \dots, X_{nb}^*$  con  $b = 1, \dots, B$ .
2. Se evalúa la réplica bootstrap de cada muestra bootstrap,  $\hat{\theta}_{nb}^* = T_n(X_{1b}^*, \dots, X_{nb}^*)$ ,  $b = 1, \dots, B$ .
3. Se estima el error estándar como la desviación estándar de las  $B$  réplicas.

$$se_b = \frac{1}{B-1} \left( \sum_{b=1}^B \left( T_{n,b}^* - \frac{1}{B} \sum_{l=1}^B T_{n,l}^* \right)^2 \right)^{1/2},$$

donde  $T_{n,b}^* = T_n(X_{1b}^*, \dots, X_{nb}^*)$ .

Mientras mayor sea  $B$  más se aproxima el estimador del error estándar a su valor real. Este tipo de Bootstrap se denomina bootstrap no paramétrico ya que utiliza un estimador no paramétrico de la distribución muestral  $F$ . La figura 2.3 muestra esquemáticamente los pasos que sigue este algoritmo.

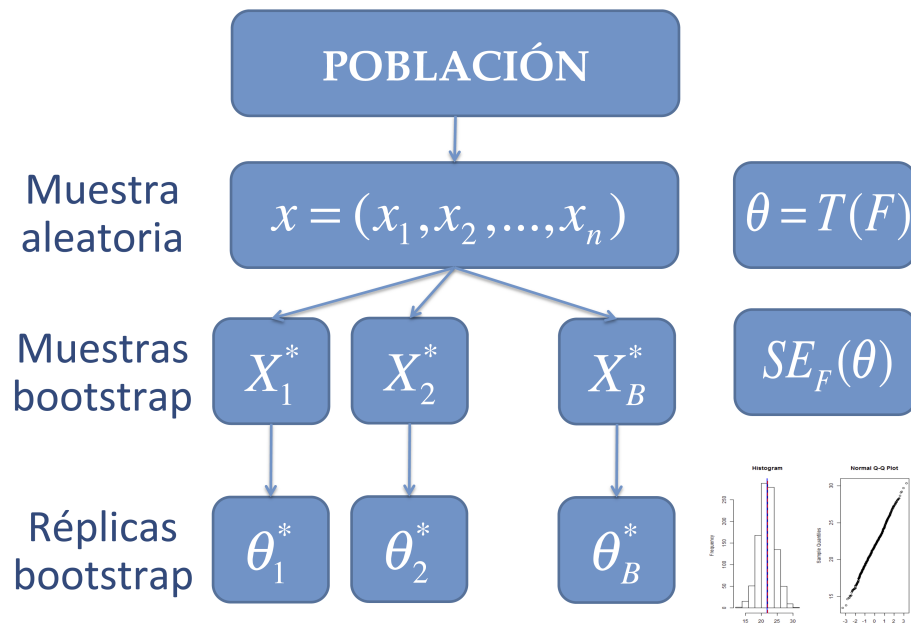


Figura 2.3: Esquema del algoritmo Bootstrap.

El algoritmo Bootstrap explicado anteriormente está basado en la estructura de datos más simple: una muestra procedente de una población con distribución muestral  $F$ . Sin embargo, hay análisis estadísticos que requieren del uso de estructuras más complejas como las series temporales, el análisis de la varianza, modelos de regresión... El algoritmo Bootstrap puede ser adaptado para poder utilizarse en el caso de tener unos datos cuya estructura es más compleja.

Hasta ahora se ha utilizado el error estándar como medida de precisión de un estimador, pero existen otras medidas que analizan diferentes aspectos del comportamiento de un estimador. Entre ellos se puede mencionar el sesgo, que se puede calcular de una forma análoga al caso de utilizar Jackknife.

Los errores estándar calculados anteriormente permiten obtener intervalos de confianza para un parámetro de interés. A continuación, se describen los principales intervalos de confianza descritos en Efron y Tibshirani, 1993.

### 2.3.2. Intervalos de Confianza Bootstrap

Sea  $X_1, \dots, X_n$  vectores aleatorios de tamaño  $p$  procedentes de una distribución muestral desconocida  $F$  y  $\theta = T(F)$  un parámetro de interés. Si  $C_n = C_n(X_1, \dots, X_n)$  es un subconjunto de  $\mathbb{R}$  que depende únicamente de  $X_1, \dots, X_n$  y

$$P(\theta \in C_n) \geq 1 - \alpha,$$

donde  $\alpha$  es una constante que satisface  $0 \leq \alpha \leq 1$ , entonces se dice que  $C_n$  es un intervalo de confianza de nivel  $1 - \alpha$ . Sea  $\underline{\theta} = \underline{\theta}_n(X_1, \dots, X_n)$  y  $\bar{\theta} = \bar{\theta}_n(X_1, \dots, X_n)$  dos estadísticos. Los intervalos  $(-\infty, \bar{\theta}]$  y  $[\underline{\theta}, \infty)$  son intervalos de confianza unilaterales y  $[-\infty, \infty]$  se denomina intervalo de confianza bilateral.  $\underline{\theta}$  (o  $\bar{\theta}$ ) se llama límite de confianza inferior (o superior). Si  $\underline{\theta}$  y  $\bar{\theta}$  son límites de confianza inferior y superior de  $\theta$  con nivel de confianza  $1 - \alpha$ , entonces  $[\underline{\theta}, \bar{\theta}]$  es un intervalo de confianza con nivel  $1 - 2\alpha$  para  $\theta$ .

En la mayoría de los casos, los intervalos de confianza se construyen considerando una cantidad pivote  $\mathfrak{R}_n = \mathfrak{R}_n(X_1, \dots, X_n, F)$  cuya distribución muestral  $G_n$  es conocida (independiente de  $F$ ). Si se puede deducir  $\underline{\theta} \leq \theta \leq \bar{\theta}$  a partir de la desigualdad  $L \leq \mathfrak{R}_n \leq U$ , entonces  $[\underline{\theta}, \bar{\theta}]$  es un intervalo de confianza de un determinado nivel (eligiendo convenientemente  $L$  y  $U$ ). Pero elegir dicha cantidad pivote para cada problema es generalmente difícil. Si  $G_n$  no se conoce, es necesario aproximarla. En el enfoque tradicional,  $G_n$  se aproxima por su límite pero si dicho límite depende de una variable desconocida  $\nu$  entonces el límite se sustituye por  $G_{\hat{\nu}}$  donde  $\hat{\nu}$  es un estimador consistente de  $\nu$ . Los métodos Bootstrap se pueden utilizar para obtener intervalos de confianza reemplazando simplemente  $G_n$  por su estimador bootstrap  $G_{boot}$ .

### Intervalos t-bootstrap

Sea  $X_1^*, \dots, X_n^*$  una muestra independiente e idénticamente distribuida procedente de  $\hat{F}$ , estimador de  $F$  (paramétrica o no paramétrica).

El método t-bootstrap (Efron, 1982) está basado en un pivote estudentizado  $\mathfrak{R}_n = (\hat{\theta}_n - \theta) / \hat{\sigma}_n$ , donde  $\hat{\theta}_n$  es un estimador de  $\theta$  y  $\hat{\sigma}_n^2$  es un estimador de la varianza de  $\hat{\theta}_n$ . Si la distribución  $G_n$  de  $\mathfrak{R}_n$  es desconocida, puede ser estimada a través del estimador bootstrap  $G_{boot}$  definido como:

$$G_{boot}(x) = P_*(\mathfrak{R}_n^* \leq x),$$

donde  $\mathfrak{R}_n^* = (\hat{\theta}_n^* - \hat{\theta}_n) / \hat{\sigma}_n^*$ , y  $\hat{\theta}_n^*$  y  $\hat{\sigma}_n^*$  son réplicas bootstrap de  $\hat{\theta}_n$  y  $\hat{\sigma}_n$  respectivamente. El límite de confianza inferior que resulta para  $\theta$  es:

$$\underline{\theta}_{bt} = \hat{\theta}_n - \hat{\sigma}_n G_{boot}^{-1}(1 - \alpha),$$

que se denomina límite de confianza inferior t-bootstrap. El método t-bootstrap es simple y fácil de entender. Sin embargo, a pesar de que los límites de confianza son altamente precisos, poseen las siguientes desventajas:

1. Para utilizar el método t-bootstrap es necesario conocer un estimador de la varianza  $\hat{\sigma}_n^2$ . Si no se dispone de estimador de la varianza, se puede utilizar el estimador bootstrap de dicha varianza  $\nu_{boot}$ . Sin embargo, si tanto  $G_{boot}$  como  $\nu_{boot}$  tienen que ser aproximados por Monte Carlo, se necesita hacer bootstrap anidado, es decir, si se generan  $B_1$  muestras para aproximar  $G_{boot}$  y por cada una de ellas es necesario  $B_2$  muestras para aproximar  $\nu_{boot}$ , en total es necesario utilizar  $B_1 B_2$  muestras. Por ejemplo, si  $B_1 = 1000$  y  $B_2 = 250$ , el total necesario sería  $B_1 B_2 = 250000$ .
2. No es invariante por reparametrización.

### Intervalos basados en Percentiles Bootstrap

Sea  $\hat{\theta}_n$  un estimador de  $\theta$  y  $\hat{\theta}_n^*$  una réplica bootstrap calculada a partir de  $X_1^*, \dots, X_n^*$ . Se define

$$K_{boot}(x) = P_*(\hat{\theta}_n^* \leq x).$$

El método Percentiles Bootstrap (Efron, 1981) calcula el límite inferior de confianza para  $\theta$  como:

$$\underline{\theta}_{bp} = K_{boot}^{-1}(\alpha).$$

El nombre percentil proviene del hecho de que  $K_{boot}^{-1}(\alpha)$  es un percentil de la distribución bootstrap  $K_{boot}$ . Si se supone que existe una transformación creciente  $\phi_n(x)$  tal que

$$P(\hat{\phi}_n - \phi_n(\theta) \leq x) = \Psi(x) \quad (2.2)$$

válido para todas las posibles  $F$  (incluido el caso  $F = \hat{F}$ ), donde  $\hat{\phi}_n = \phi_n(\hat{\theta}_n)$ , y  $\Psi$  es una distribución continua, creciente y simétrica ( $\Psi(x) = 1 - \Psi(-x)$ ). Cuando  $\Psi = \Phi$ , la distribución normal estándar, la función  $\phi_n$  es la transformación de normalización y estabilización de la varianza. Si  $\phi_n$  y  $\Psi$  son conocidas, se puede obtener el siguiente límite inferior de confianza para  $\theta$ :

$$\underline{\theta}_{exact} = \phi_n^{-1}(\hat{\phi}_n + z_\alpha)$$

donde  $z_\alpha = \Psi^{-1}(\alpha)$ .

A continuación se muestra que  $\underline{\theta}_{bp} = \underline{\theta}_{exact}$  y por consiguiente, se puede utilizar este límite de confianza incluso si  $\phi$  y/o  $\Psi$  son desconocidas. Sea



$w_n = \phi_n(\underline{\theta}_{bp}) - \hat{\phi}_n$ . Dado que la ecuación 2.2 es válida para  $F = \hat{F}$ ,

$$\Psi(w_n) = P_* \left( \hat{\phi}_n^* - \hat{\phi}_n \leq w_n \right) = P_* \left( \hat{\theta}_n^* \leq \underline{\theta}_{bp} \right) = \alpha,$$

donde  $\hat{\phi}_n^* = \phi_n(\hat{\theta}_n^*)$  y la última igualdad deriva de la definición de  $\underline{\theta}_{bp}$  y de los supuestos de  $\Psi$ . Por lo tanto,  $w_n = z_\alpha = \Psi^{-1}(\alpha)$  y

$$\underline{\theta}_{bp} = \phi_n^{-1} \left( \hat{\theta}_n + z_\alpha \right) = \underline{\theta}_{exact}.$$

Se ha demostrado que el límite inferior para el método Percentiles Bootstrap es exacto para toda  $n$  si la ecuación 2.2 es válida exactamente. Si dicha ecuación es válida aproximadamente para valores grandes de  $n$ , entonces el límite de confianza es asintóticamente válido y su cálculo depende de cómo de buena sea dicha aproximación. Esto funciona bien en el caso de que la transformación  $\phi_n$  sea lineal. Si  $\phi_n$  no es lineal, el sesgo de  $\hat{\phi}_n - \phi_n(\theta)$  no desaparece rápidamente cuando  $n \rightarrow \infty$ . La aproximación sólo es buena cuando  $n$  es muy grande y por lo tanto el límite no es muy preciso a no ser que  $n$  sea muy grande. Esto conduce al desarrollo de los dos siguientes métodos que también se basan en ciertos percentiles de la distribución  $K_{boot}$ .

### Intervalos basados en Percentiles Bootstrap corregido por sesgo

La última apreciación sugiere la adaptación de la ecuación 2.2 para incluir un término que tenga en cuenta el sesgo. Efron, 1982 propone reescribir la ecuación de la siguiente manera:

$$P \left( \hat{\phi}_n - \phi_n(\theta) + z_0 \leq x \right) = \Psi(x) \quad (2.3)$$

donde  $z_0$  es una constante que puede depender de  $F$  y de  $n$ ,  $\phi_n$  es una

transformación creciente y  $\Psi$  se sigue asumiendo continua, estrictamente creciente y simétrica. Cuando  $z_0 = 0$  la ecuación 2.3 se reduce a 2.2. Si  $\phi_n$ ,  $z_0$  y  $\Psi$  son conocidas, se puede obtener el límite inferior del intervalo de confianza como:

$$\underline{\theta}_{exact} = \phi_n^{-1} \left( \hat{\phi}_n + z_\alpha + z_0 \right).$$

Aplicando la ecuación 2.3 en el caso  $F = \hat{F}$ , se obtiene:

$$K_{boot}(\hat{\theta}_n) = P_* \left( \hat{\phi}_n - \phi_n(\theta) + z_0 \leq z_0 \right) = \Psi(z_0).$$

Esto implica que

$$z_0 = \Psi^{-1} \left( K_{boot} \left( \hat{\theta}_n \right) \right). \quad (2.4)$$

De 2.3

$$\begin{aligned} 1 - \alpha &= \Psi(-z_\alpha) \\ &= \Psi \left( \hat{\phi}_n - \phi_n(\underline{\theta}_{exact}) + z_0 \right) \\ &= P_* \left( \hat{\phi}_n^* - \hat{\phi}_n \leq \hat{\phi}_n - \phi_n(\underline{\theta}_{exact}) \right) \\ &= P_* \left( \hat{\theta}_n^* \leq \phi_n^{-1} \left( \hat{\phi}_n - z_\alpha - z_0 \right) \right), \end{aligned}$$

lo que implica

$$\phi_n^{-1} \left( \hat{\phi}_n - z_\alpha - z_0 \right) = K_{boot}^{-1} (1 - \alpha).$$

Puesto que esta última igualdad es válida para cada  $\alpha$ , esto implica que  $0 < x < 1$ ,

$$K_{boot}^{-1}(x) = \phi_n^{-1} \left( \hat{\phi}_n + \Psi^{-1}(x) - z_0 \right). \quad (2.5)$$

Por esto y por la definición de  $\underline{\theta}_{exact}$  se tiene:

$$\underline{\theta}_{exact} = K_{boot}^{-1} \left( \Psi(z_\alpha + 2z_0) \right).$$

Ahora, si se supone que  $\Psi$  es conocida y utilizando la ecuación 2.4, se obtiene el límite inferior del intervalo de confianza *bootstrap bias-corrected percentile (BC)* para  $\theta$  (Efron, 1981):

$$\underline{\theta}_{bc} = K_{boot}^{-1} \left( \Psi \left( z_\alpha + 2\Psi^{-1} \left( K_{boot} \left( \hat{\theta}_n \right) \right) \right) \right).$$

Normalmente, se suele utilizar  $\Psi = \Phi$ . Dado que  $\Psi^{-1} \left( \frac{1}{2} \right) = 0$ ,  $\underline{\theta}_{bc}$  se reduce a  $\underline{\theta}_{bp}$  si  $K_{boot} \left( \hat{\theta}_n \right) = \frac{1}{2}$ , es decir,  $\hat{\theta}_n$  es la mediana de la distribución bootstrap  $K_{boot}$ . Por lo tanto,  $\underline{\theta}_{bc}$  es una versión del método de los Percentiles Bootstrap ajustada para el sesgo. De nuevo,  $\underline{\theta}_{bc}$  es exacto para toda  $n$  si 2.3 se cumple exactamente y es asintóticamente válido si 2.3 se cumple aproximadamente.

Este método mejora el método basado en Percentiles Bootstrap teniendo en cuenta el sesgo. Sin embargo, hay muchos casos para los que la ecuación 2.3 no se puede cumplir y consecuentemente este método no funciona bien.

### **Intervalos basados en Percentiles Bootstrap corregido por sesgo y acelerado**

Efron, 1987 introdujo una suposición más exhaustiva que tiene en cuenta la asimetría:

$$P \left( \frac{\hat{\phi}_n - \phi_n(\theta)}{1 + a\phi_n(\theta)} + z_0 \leq x \right) = \Psi(x) \quad (2.6)$$

donde  $\phi_n$ ,  $z_0$  y  $\Psi$  son los mismos que en la ecuación 2.3 pero  $a$  es un parámetro nuevo que depende de  $F$  y  $n$ . Según Efron, 1987  $a$  mide la rapidez del cambio de la desviación estándar de  $\hat{\phi}_n$  con respecto a  $\phi_n(\theta)$  y por esa razón se denomina *constante de aceleración*. Claramente se puede observar que la ecuación 2.6 es más general que 2.2 y 2.3.

Si  $\phi_n$ ,  $z_0$  y  $\Psi$  son conocidos, un límite exacto de confianza para  $\theta$  es:

$$\underline{\theta}_{exact} = \phi_n^{-1} \left( \frac{\hat{\phi}_n + (z_\alpha + z_0)(1 + a\hat{\theta}_n)}{1 - a(z_\alpha + z_0)} \right).$$

A continuación se demuestra que  $\underline{\theta}_{exact}$  es un percentil de  $K_{boot}$ . En primer lugar, la ecuación 2.4 sigue siendo válida. Utilizando la ecuación 2.6 y tal como se probó en 2.5, se obtiene que para  $0 < x < 1$ ,

$$K_{boot}^{-1}(x) = \phi_n^{-1} \left( \hat{\phi} [\Psi^{-1}(x) - z_0] (1 + a\hat{\phi}_n) \right).$$

Si se sustituye  $x = \Psi \left( \frac{z_0 + (z_\alpha + z_0)}{1 - a(z_\alpha + z_0)} \right)$  en la ecuación anterior se obtiene que:

$$\underline{\theta}_{BCa}(a) = K_{boot}^{-1} \left( \Psi \left( \frac{z_0 + (z_\alpha + z_0)}{1 - a(z_\alpha + z_0)} \right) \right) = \underline{\theta}_{exact}. \quad (2.7)$$

Es decir, si  $\Psi$  y  $a$  son conocidos, entonces  $\underline{\theta}_{BCa}(a)$  es un límite inferior exacto de confianza para  $\theta$  para toda  $n$ . Si se puede estimar  $a$  a través de  $\hat{a}$  y sustituirla en la ecuación 2.7 entonces se tiene que:

$$\underline{\theta}_{BCa} = \underline{\theta}_{BCa}(\hat{a})$$

es el límite inferior del intervalo de confianza *bootstrap accelerated bias-corrected percentile (BCa)* para  $\theta$  (Efron, 1987).

El parámetro  $a$  no es fácil de estimar. En Efron, 1987 se muestran varios métodos para aproximarlos. Una de las formas de estimar  $a$  y  $z_0$  es utilizar

Jackknife. Dado que  $a$  y  $z_0$  son funciones de la varianza, el sesgo y la asimetría de  $\hat{\theta}_n$ , Tu y Zhang, 1992 sugieren la siguiente aproximación para  $a$  y  $z_0$ :

$$\hat{a}_{jack} = \frac{(n-1)^3}{6n^3\nu_{jack}^{3/2}} \sum_{i=1}^n \left( \hat{\theta}_{n-1,i} - \frac{1}{n} \sum_{j=1}^n \hat{\theta}_{n-1,j} \right)^3, \quad (2.8)$$

$$\hat{z}_{jack} = sesgo_{jack}/(n\nu_{jack}) - sk_{jack}/6, \quad (2.9)$$

donde

$$\hat{\theta}_{n-1,i} = T(F_{n-1,i}),$$

$$sesgo_{jack} = (n-1)(\bar{T}_n - T_n),$$

$$\nu_{jack} = \frac{n-1}{n} \sum_{i=1}^n \left( T_{n-1,i} - \frac{1}{n} \sum_{j=1}^n T_{n-1,j} \right)^2,$$

$$sk_{jack} = \frac{3(n-1)^2}{n^3 sesgo_{jack}^{3/2}} \sum_{i \neq j} (T_{n-1,i} - \bar{T}_n) (T_{n-1,j} - \bar{T}_n) \Delta_{ij} \\ - \frac{(n-1)^3}{n^3 sesgo_{jack}^{3/2}} \sum_{i=1}^3 (T_{n-1,i} - \bar{T}_n)^3$$

y  $T_n = \hat{\theta}_n$ .

El método BCa bootstrap es invariante bajo reparametrización. Puesto que el cálculo de  $a$  y  $z_0$  están pensados para conseguir una alta precisión, es esperable que los intervalos de confianza que produce este método sean precisos.

Una desventaja que presenta este método es que la determinación de  $a$  no es fácil, especialmente si el problema a tratar es complejo.

La versión inferencial para los métodos Biplot que se propone, considera como unidades de muestreo las filas que contienen la información de las variables

medidas sobre cada uno de los individuos. De tal forma, se extraen  $B$  muestras con reposición del mismo número de individuos que la matriz de partida y a cada una de ellas se le aplica el Biplot (GH, HJ o JK). Por tanto, para cada parámetro de interés, se dispone de un vector que contiene  $B$  valores calculados a partir de las muestras bootstrap. A partir de cada uno de los vectores es posible calcular la media, la desviación estándar, y los intervalos explicados a lo largo de esta sección: t-bootstrap, basado en percentiles y corregido por sesgo y acelerado (BCa). Los valores de cada vector también se pueden representar mediante histogramas y gráficos de normalidad con el fin de estudiar su distribución.

## 2.4. Paquete *biplotbootGUI*

Debido a la necesidad de tener disponible un software que permita obtener los resultados de un análisis Biplot de una forma inferencial, se ha desarrollado un nuevo paquete en el lenguaje R (R-Team, 2014) que incluye el análisis Biplot en el sentido en el que lo desarrollaron Gabriel, 1971 y Galindo, 1986 y la metodología Bootstrap para presentar sus resultados complementados con medidas de precisión.

R es un conjunto integrado de programas para manipulación de datos, cálculo y gráficos. Entre otras características dispone de:

- Almacenamiento y manipulación efectiva de datos.
- Operadores para cálculo sobre variables indexadas (Arrays), en particular matrices.
- Una amplia, coherente e integrada colección de herramientas para análisis de datos.

- Posibilidades gráficas para análisis de datos, que funcionan directamente sobre pantalla o impresora.
- Un lenguaje de programación bien desarrollado, simple y efectivo, que incluye condicionales, ciclos, funciones recursivas y posibilidad de entradas y salidas.

El término *entorno* lo caracteriza como un sistema completamente diseñado y coherente, antes que como una agregación incremental de herramientas muy específicas e inflexibles, como ocurre frecuentemente con otros programas de análisis de datos.

El nuevo paquete desarrollado se denomina `biplotbootGUI` (Nieto et al., 2014) y está disponible en <http://cran.r-project.org/web/packages/biplotbootGUI>. El nombre proviene de la contracción de las palabras *biplot* (**biplot**), *bootstrap* (**boot**) e *Interfaz Gráfica de Usuario* (**GUI**).

El paquete es una interfaz gráfica que permite interactuar mediante el uso de ventanas, botones y menús.

En primer lugar, es necesario bajar R de la web [cran.r-project.org](http://cran.r-project.org) e instalarlo. A continuación se descarga el paquete `biplotbootGUI` y sus dependencias que son los paquetes: `rgl`, `tcltk`, `tcltk2`, `tkrplot`, `shapes`, `cluster` y `dendroextras`; (Adler y Murdoch, 2012; Dryden, 2014; Grosjean, 2012; Jefferis, 2014; Maechler et al., 2015; Tierney, 2012).

Para cargar el paquete `biplotbootGUI` en el software R, se debe introducir el comando `library(biplotbootGUI)` en la ventana principal de R. Una vez realizadas estas acciones, se deben cargar los datos que se quieren analizar. Entonces, la interfaz se inicializa mediante el comando `biplotboot(data)` donde `data` corresponde al nombre de los datos que se van a analizar.

Una vez que se ha ejecutado la sentencia se abre la ventana principal. Dicha

ventana (figura 2.4) está dividida en dos bloques. El bloque de la izquierda permite introducir el número de submuestras que se van a extraer a partir de los datos de partida para calcular los intervalos de confianza; el nivel de confianza que se va a utilizar para calcularlos y la opción de guardar los gráficos que se generan a partir del remuestreo bootstrap tanto con extensión .eps como .pdf ambas en color o en blanco y negro. En el bloque de la derecha se encuentra un listado con todas los resultados que proporciona el análisis Biplot para los cuales el programa ofrece medidas de precisión:

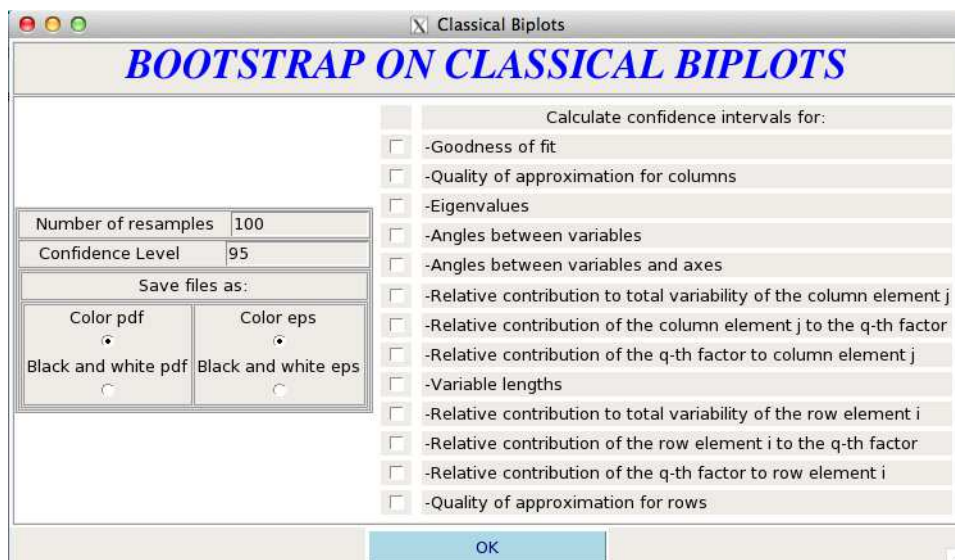


Figura 2.4: Ventana principal.

- Bondad de ajuste.
- Calidad de aproximación para columnas.
- Valores propios.
- Ángulos entre variables.
- Ángulos entre variables y ejes.
- Contribución relativa a la variabilidad total del elemento columna  $j$ .



- Contribución relativa del  $j$ -ésimo elemento columna al  $q$ -ésimo factor.
- Contribución relativa del  $q$ -ésimo factor al  $j$ -ésimo elemento columna.
- Longitud de las variables.
- Contribución relativa a la variabilidad total del elemento fila  $i$ .
- Contribución relativa del  $i$ -ésimo elemento fila al  $q$ -ésimo factor.
- Contribución relativa del  $q$ -ésimo factor al  $i$ -ésimo elemento fila.
- Calidad de aproximación para filas.

Una vez que el usuario ha terminado de elegir las opciones necesarias y pulse el botón OK aparece la ventana de opciones (figura 2.5). En ella es posible:

- Elegir el tipo de Biplot que se quiere realizar (HJ, GH o JK).
- Elegir una transformación previa a realizar sobre los datos originales:
  - Restar la media global
  - Centrar por columnas
  - Estandarizar por columnas
  - Centrar por filas
  - Estandarizar por filas
  - Doble centrado
  - No realizar ninguna transformación
- Cambiar color, tamaño, etiqueta y símbolo que van a representar a los individuos en los gráficos.

- Cambiar color, tamaño y etiqueta que van a representar a las variables en los gráficos.
- Mostrar los ejes de coordenadas en los gráficos.
- Cambiar el tamaño de las ventanas para que el programa sea adaptable a los diferentes tamaños de las pantallas.

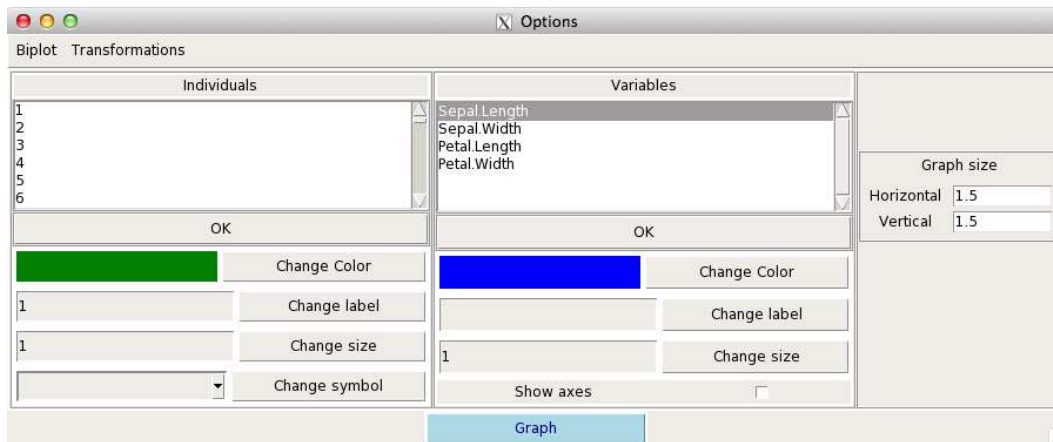


Figura 2.5: Ventana de opciones.

Cuando se pulsa el botón **Graph** emerge una nueva ventana (figura 2.6) en la que se muestra un diagrama de barras en el que se representa la inercia absorbida por cada eje y se puede elegir el número de ejes a retener para el posterior análisis.

Una vez elegidos el número de ejes deseado y pulsado el botón **Choose**, aparece una ventana (figura 2.7) que contiene el gráfico resultante sobre las dos primeras dimensiones.

En esta ventana se pueden distinguir dos bloques y varios menús. En el bloque de la derecha se tiene la representación gráfica de las coordenadas biplot tanto para individuos como para variables. En este gráfico es posible mover las etiquetas de los puntos con el botón izquierdo del ratón y se pueden cambiar las características gráficas de los puntos con el botón derecho del ratón. En el bloque de la izquierda se pueden observar tres listbox que sirven para elegir la

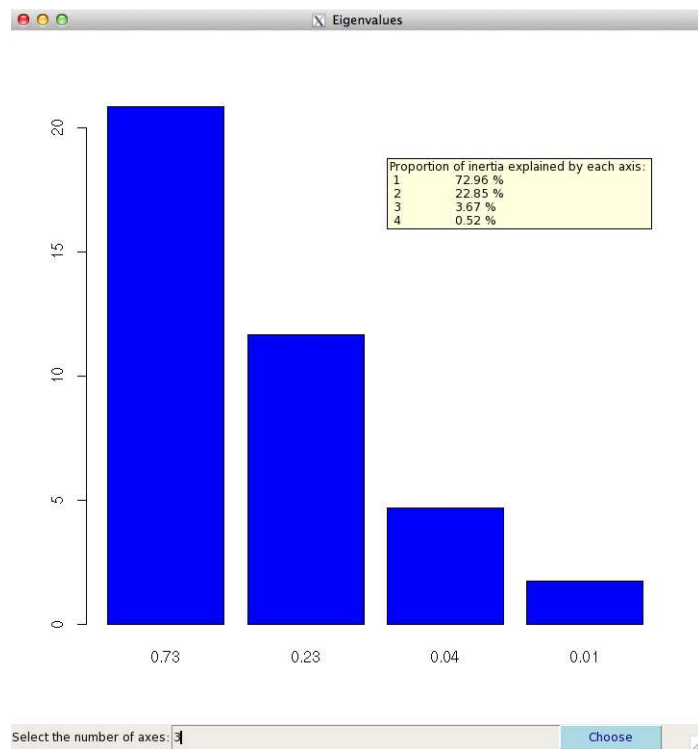


Figura 2.6: Ventana con el gráfico de barras representando la inercia absorbida por cada eje.

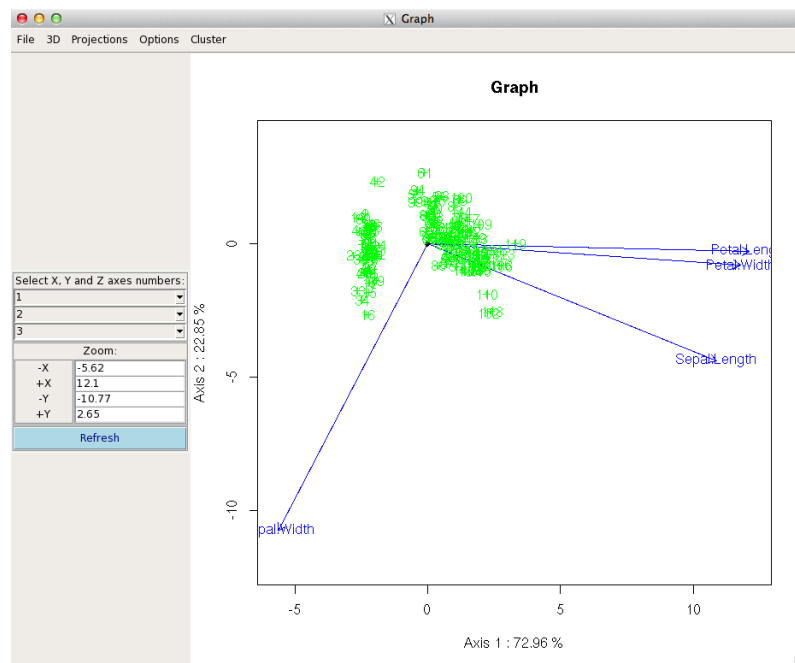


Figura 2.7: Ventana que muestra la representación Biplot dos dimensiones.

dimensión que se quiere ver en cada gráfico (2 y 3 dimensiones); cuatro textbox que permiten elegir los límites de los ejes  $x$  e  $y$  para poder observar parte del gráfico con más detalle o menos y un botón **Refresh** que es necesario pulsar para que se haga efectivo el cambio de los anteriores elementos. En la parte superior de la ventana se dispone de 5 menús con sus correspondientes submenús:

- File
  - Copy image
  - Save image
    - PDF file
    - Eps file
    - Png file
    - Jpg/Jpeg file
  - Exit
- 3D
  - 3D
- Projections
  - Variables
  - Back to original data
- Options
  - Change title
  - Show/Hide axes
- Cluster

- Hierarchical cluster with biplot coordinates
- K-means with biplot coordinates
- K-medoids with biplot coordinates
- Back to original graph

El primero de ellos permite copiar la imagen al portapapeles, guardarla en varios tipos de formato y salir del programa. El siguiente es el menú de 3 dimensiones. Tiene la opción de mostrar las coordenadas biplot en las tres dimensiones elegidas mediante los tres listbox. En este gráfico resultante es posible rotar la imagen con el botón izquierdo del ratón y ampliar o disminuir la imagen con el botón derecho del ratón. El tercer menú que se encuentra el usuario en esta ventana sirve para proyectar los puntos que representan a los individuos sobre una determinada variable. Si se elige el submenú **Variables** emerge una ventana con el listado de las variables que están siendo analizadas, al elegir una de ellas y pulsar el botón **OK**, se muestran en el gráfico las proyecciones de los individuos sobre dicha variable. Si se desea volver al gráfico anterior se elige el submenú **Back to original graph**. El siguiente menú disponible es el de opciones del gráfico. Contiene dos submenús: cambiar el título del gráfico y mostrar o quitar los ejes de coordenadas. El último menú es el que permite analizar las coordenadas biplot con técnicas de Cluster. Los métodos de los que se dispone son:

- Cluster Jerárquico. Se dispone de una ventana para elegir el número de clusters, la distancia que se quiere utilizar (Euclidean, Maximum, Manhattan, Canberra, Binary, Minkowski) y el método de agrupación (Ward.D, Ward.D2, Single, Complete, Average, Mcquitty, Median, Centroid). Se obtiene como resultado un dendrograma de la agrupación de los individuos y se muestra sobre el gráfico en dos dimensiones en distintos colores y encerrados en una línea poligonal cada uno de los clusters obtenidos.

- Cluster de k-medias. Se proporciona una ventana para elegir el número de clusters, el número máximo de iteraciones que se va a permitir realizar hasta llegar a la solución óptima, el número de conjuntos aleatorios de centroides que se van a utilizar como solución inicial en el algoritmo y el algoritmo que se va a aplicar para la búsqueda de la solución (Hartigan-Wong, Lloyd, Forgy, MacQueen). Este método muestra sobre el gráfico en dos dimensiones en distintos colores y encerrados en una línea poligonal tanto los clusters obtenidos como los centroides de cada uno de ellos.
- Cluster de K-medoides. En este método aparece una ventana para elegir el número de clusters deseado por el usuario y la distancia que se quiere utilizar (Euclidean, Maximum, Manhattan, Canberra, Binary, Minkowski).

En este menú también es posible volver atrás y recuperar el gráfico que se mostraba antes de aplicar las técnicas Cluster.

Junto a esta ventana emerge un texto con los resultados del análisis Biplot (coordenadas, bondad de ajuste, calidades de representación para filas y columnas, ángulos entre variables y entre variables y ejes, longitudes de los vectores que representan variables, contribuciones de los elementos a la variabilidad total, contribuciones de los elementos a los factores y de los factores a los elementos, inercia absorbida por cada eje y valores propios). En cuanto a los resultados bootstrap se refiere, aparece otro texto que muestra para cada parámetro elegido el valor observado, la media calculada a partir de los valores del parámetro en cada remuestreo, la desviación estándar, el sesgo y los extremos inferior y superior de los intervalos de confianza t-bootstrap, percentiles y BCa descritos en la sección 2.3.2.

Estos dos archivos de texto se guardan automáticamente en el directorio en el que se encuentre el usuario así como todos las figuras que contienen histogramas

y gráficos de normalidad de los parámetros seleccionados en la pantalla principal. Por último, se muestra un gráfico donde aparecen representadas todas las coordenadas de las variables que se han calculado a partir de las muestras bootstrap. Para ello se han realizado rotaciones Procrustes con el objetivo de eliminar el efecto espejo que pudiera existir entre los diferentes conjuntos de coordenadas. Cada grupo de coordenadas que representan a la misma variable se muestra en el mismo color y se encuentra encerrado bajo una línea poligonal cerrada.

## 2.5. Aplicación a Datos

A continuación se desarrollan dos ejemplos de aplicación e interpretación del paquete desarrollado anteriormente.

### 2.5.1. Datos Simulados

Para el siguiente ejemplo se utiliza una matriz de datos simulados con la función `mvrnorm` del paquete `MASS` de R. La matriz `sigma` necesaria para su generación es:

4	1	2	0.5	2
1	0.5	0	0	0
2	0	3	0	0
0.5	0	0	0.625	0
2	0	0	0	16

Tabla 2.4: Matriz sigma para la generación de los datos simulados.

La matriz consta de 100 observaciones sobre 5 variables.

Se ha realizado un HJ Biplot con la transformación *Doble centrado*.

En la tabla 2.5 se observa la información proporcionada por los valores propios. En ella se aprecia que el primer eje absorbe la mayor parte de la información

(casi el 77 %) y con los tres primeros ejes se explica prácticamente el 100 % de la información.

No.	Valor Propio	Variabilidad	Variabilidad Acumulada
1	36.98	76.76	76.76
2	15.17	12.91	89.67
3	11.86	7.89	97.56

Tabla 2.5: Valores propios y variabilidad explicada (%) por cada eje para los datos simulados.

A continuación se muestran las contribuciones relativas del factor al elemento para las variables simuladas en los tres ejes retenidos (tabla 2.6).

Variable	Eje 1	Eje 2	Eje 3
V1	54.13	141.19	478.88
V2	58.76	105.17	40.45
V3	52.38	332.88	408.65
V4	36.01	420.36	71.81
V5	798.71	0.40	0.20

Tabla 2.6: Contribuciones relativas del factor al elemento columna para datos simulados.

Según dicha tabla, la única variable que está bien representada en el eje 1 es V5; V1 presenta mejor representación en el eje 3; V4 en el eje 2 y V3 estaría representado con una calidad similar en los ejes 2 y 3.

La figura 2.8 muestra el plano formado por los dos primeros ejes con la representación Biplot de los datos simulados.

Según se puede apreciar en el gráfico, las variables V2 y V4 están altamente correlacionadas así como las variables V1 y V3. Sin embargo, entre dichos pares existe muy poca relación. La variable V5 presenta la mayor variabilidad y una relación inversa con el resto de variables. Las longitudes de las otras cuatro variables son similares. Dichas relaciones se recogen en las tablas 2.7 y 2.8.



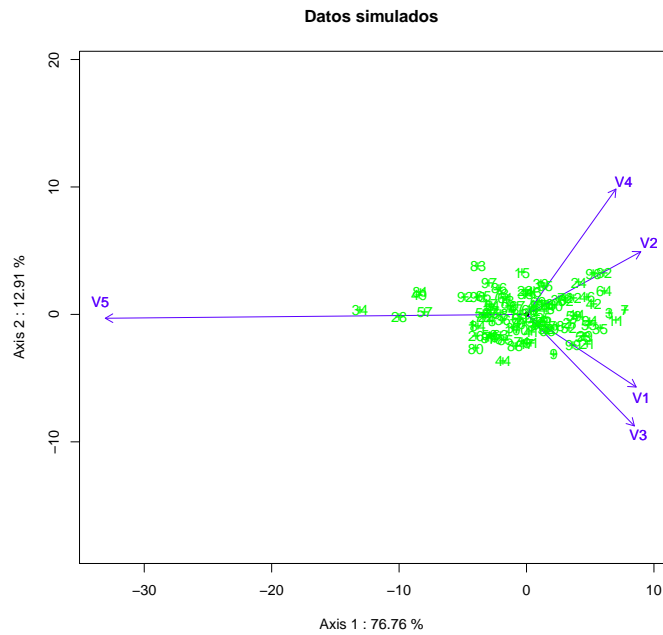


Figura 2.8: Gráfico que muestra la representación HJ Biplot para los datos simulados en las dos primeras dimensiones.

Variable	Longitud
V1	10.32
V2	10.23
V3	12.18
V4	12.08
V5	33.06

Tabla 2.7: Longitud de las variables para datos simulados.

Variable	V1	V2	V3	V4	V5
V1	0.00	62.27	12.44	88.01	145.96
V2	62.27	0.00	74.71	25.73	151.77
V3	12.44	74.71	0.00	100.45	133.52
V4	88.01	25.73	100.45	0.00	126.04
V5	145.96	151.77	133.52	126.04	0.00

Tabla 2.8: Ángulos entre variables para datos simulados.

A continuación se muestran los resultados obtenidos al aplicar la metodología Bootstrap a los parámetros que se obtienen del HJ Biplot. Para todos ellos se presentan:

- Histograma de los valores de las 1000 réplicas bootstrap. En ellos se ha resaltado en línea continua azul el valor de la media de dichos valores y en línea discontinua roja el valor calculado a partir de la muestra inicial.
- Gráfico de normalidad.
- Tabla en la que se muestra:
  - Valor observado del parámetro (calculado a partir de la muestra inicial).
  - Media de los valores obtenidos a partir de las réplicas bootstrap.
  - Desviación estándar de dichos valores.
  - Sesgo.
  - Extremos inferior y superior del intervalo t-bootstrap.
  - Extremos inferior y superior del intervalo basado en percentiles.
  - Extremos inferior y superior del intervalo basado en los percentiles BCa.

En primer lugar, se muestran los resultados de aplicar los métodos Bootstrap al HJ Biplot para la calidad de aproximación para las columnas. Según se observa en la tabla 2.9 y en la figura 2.9, la diferencia entre el valor obtenido mediante el HJ Biplot de la muestra inicial y el calculado a partir del remuestreo es muy pequeña y la amplitud de los intervalos también con lo que se deduce que la calidad de aproximación para las columnas es muy estable.

V. Obs	Media	SE	Sesgo	t-b inf	t-b sup	perc inf	perc sup	BCa inf	BCa sup
97.57	97.62	0.5	0.06	96.65	98.60	96.50	98.46	96.24	98.34

Tabla 2.9: Calidades de aproximación de las columnas para datos simulados.

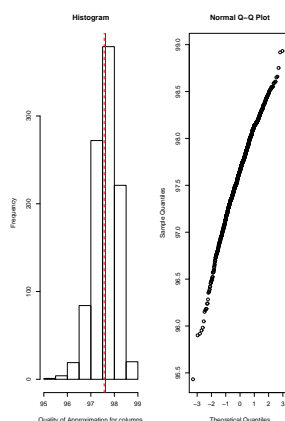


Figura 2.9: Histograma de las calidades de aproximación de las variables para datos simulados.

Lo mismo se puede deducir observando los datos detallados en la tabla 2.10 y figura 2.10 referentes a los resultados de los valores propios. Tienen unos sesgos muy pequeños y poca amplitud en los intervalos de confianza por lo que los datos que se han calculado para las 1000 muestras bootstrap no difieren demasiado entre sí y podemos deducir gran estabilidad para estos parámetros. Cabe destacar que la distribución del último valor propio es ligeramente asimétrica y se desvía más de la normalidad que los cuatro primeros.

	V. Obs	Media	SE	Sesgo	t-b inf	t-b sup	perc inf	perc sup	BCa inf	BCa sup
1	36.99	36.66	3.11	-0.33	30.56	42.76	30.76	42.93	31.94	44.64
2	15.17	15.14	0.95	-0.03	13.27	17.01	13.35	17.10	13.50	17.23
3	11.86	11.54	0.77	-0.31	10.03	13.06	9.99	13.13	10.83	13.72
4	6.59	6.39	0.49	-0.20	5.42	7.36	5.47	7.39	5.83	7.71
5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Tabla 2.10: Valores propios para datos simulados.

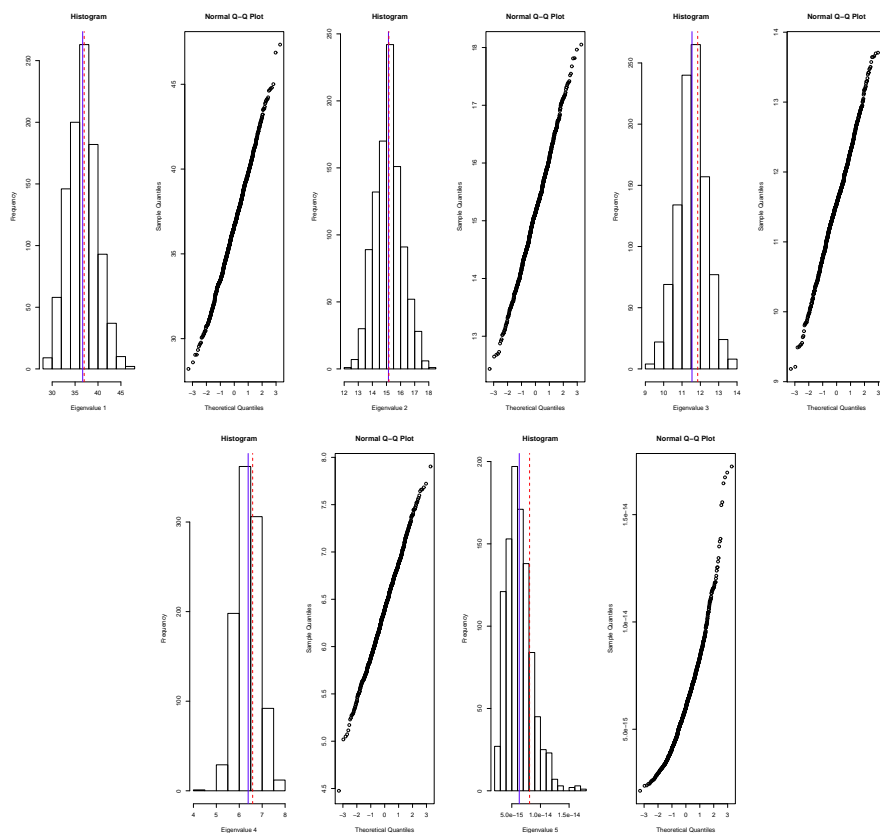


Figura 2.10: Histograma de los valores propios para datos simulados.

En cuanto a los ángulos entre las variables y entre las variables y los ejes se observa que las diferencias entre los valores observados y los estimados no son grandes y las amplitudes de los intervalos son pequeñas, así que se puede considerar que las relaciones entre las variables y entre las variables y los ejes son estables (tablas 2.11 y 2.12 y figuras 2.11 y 2.12). Las distribuciones de los ángulos de la variable 1 con el resto son bastante asimétricas y se desvían de la normalidad. Lo mismo sucede con los ángulos entre las variables 3 y 4. Pero esto es debido a las transformaciones que se hacen sobre los ángulos obtenidos con objeto de que sean todos positivos entre 0 y 180 grados. Estas transformaciones se realizan para que las diferencias entre los ángulos no se distorsionen en caso de tener ángulos parecidos con distinto signo o distinta orientación. Esta misma situación se observa en los ángulos que forma la variable 5 con los ejes 1 y 2.

	V. Obs	Media	SE	Sesgo	t-b inf	t-b sup	perc inf	perc sup	BCa inf	BCa sup
V1										
V2	62.27	60.46	11.01	-1.81	38.85	82.07	35.39	79.15	40.72	81.58
V3	12.44	17.76	12.97	5.33	-7.68	43.21	0.97	48.69	0.18	38.13
V4	88.01	86.45	14.02	-1.56	58.93	113.96	52.67	108.16	52.00	108.01
V5	145.96	146.96	10.98	1.01	125.42	168.51	128.65	172.83	128.89	173.47
V2										
V3	74.71	73.26	13.98	-1.45	45.82	100.70	43.59	95.41	43.84	95.70
V4	25.73	25.99	9.09	0.25	8.15	43.82	9.35	44.41	9.15	43.71
V5	151.77	152.36	6.98	0.59	138.67	166.05	138.57	166.84	137.64	164.52
V3										
V4	100.45	99.10	9.80	-1.35	79.86	118.33	79.30	116.29	82.47	117.83
V5	133.52	134.38	8.61	0.86	117.48	151.28	120.50	152.60	120.55	152.69
V4										
V5	126.04	126.40	5.09	0.37	116.42	136.39	116.81	137.04	116.25	136.03

Tabla 2.11: Ángulos entre variables para datos simulados.

	V. Obs	Media	SE	Sesgo	t-b inf	t-b sup	perc inf	perc sup	BCa inf	BCa sup
V1										
E.1	33.52	32.55	11.32	-0.97	10.33	54.77	6.75	52.14	5.95	51.53
E.2	56.48	57.45	11.32	0.97	35.23	79.67	37.86	83.25	38.47	84.05
V2										
E.1	28.75	28.15	7.33	-0.61	13.76	42.54	13.67	42.80	15.97	45.58
E.2	61.25	61.85	7.33	0.61	47.46	76.24	47.20	76.33	44.42	74.03
V3										
E.1	45.96	45.12	8.96	-0.84	27.53	62.71	26.31	60.39	26.98	61.00
E.2	44.04	44.88	8.96	0.84	27.29	62.47	29.61	63.69	29.00	63.02
V4										
E.1	54.49	54.11	5.80	-0.38	42.72	65.49	42.39	65.15	43.56	66.26
E.2	35.51	35.89	5.80	0.38	24.51	47.28	24.85	47.61	23.74	46.45
V5										
E.1	0.52	1.00	0.79	0.47	-0.56	2.55	0.05	3.03	0.01	1.85
E.1	89.48	89.00	0.79	-0.47	87.45	90.56	86.97	89.95	88.15	89.99

Tabla 2.12: Ángulos entre variables y ejes para datos simulados.

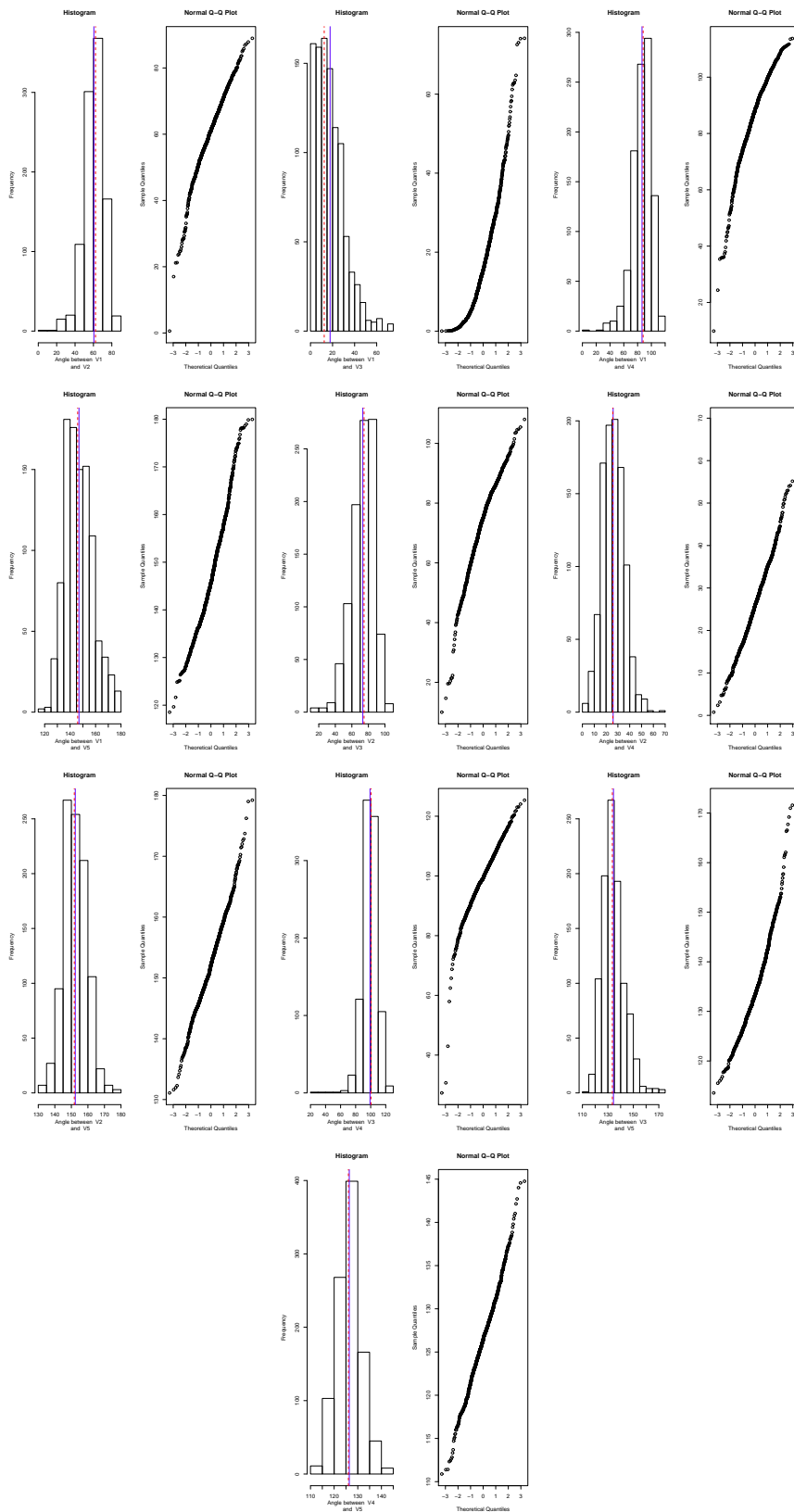


Figura 2.11: Histograma de los ángulos entre variables para datos simulados.

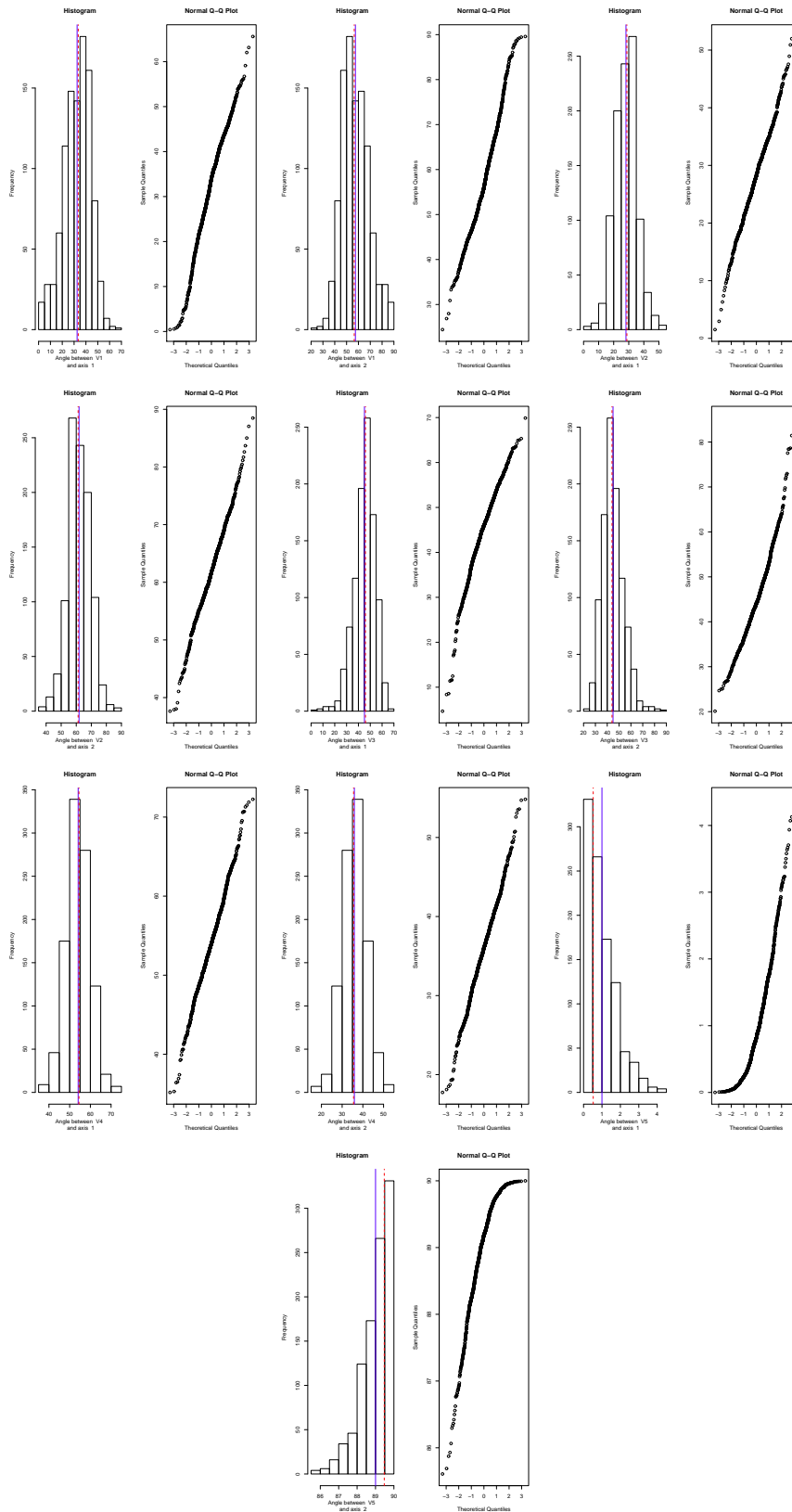


Figura 2.12: Histograma de los ángulos entre variables y ejes para datos simulados.

Si analizamos la longitud de las variables, la tabla 2.13 y figura 2.13 muestran que son estables ya que los sesgos son bajos y las desviaciones estándar también. Para estos parámetros las distribuciones de las réplicas bootstrap son simétricas y se pueden suponer normales.

	V. Obs	Media	SE	Sesgo	t-b inf	t-b sup	perc inf	perc sup	BCa inf	BCa sup
V1	10.32	10.36	1.46	0.04	7.49	13.22	7.72	13.47	7.84	13.69
V2	10.23	10.18	1.07	-0.05	8.08	12.28	8.06	12.20	8.28	12.54
V3	12.18	12.06	1.98	-0.12	8.17	15.95	7.74	15.56	7.94	15.70
V4	12.08	11.89	1.11	-0.19	9.72	14.06	9.64	13.86	9.86	14.02
V5	33.06	32.71	2.79	-0.35	27.24	38.18	27.41	38.36	28.56	39.89

Tabla 2.13: Longitud de las variables para datos simulados.

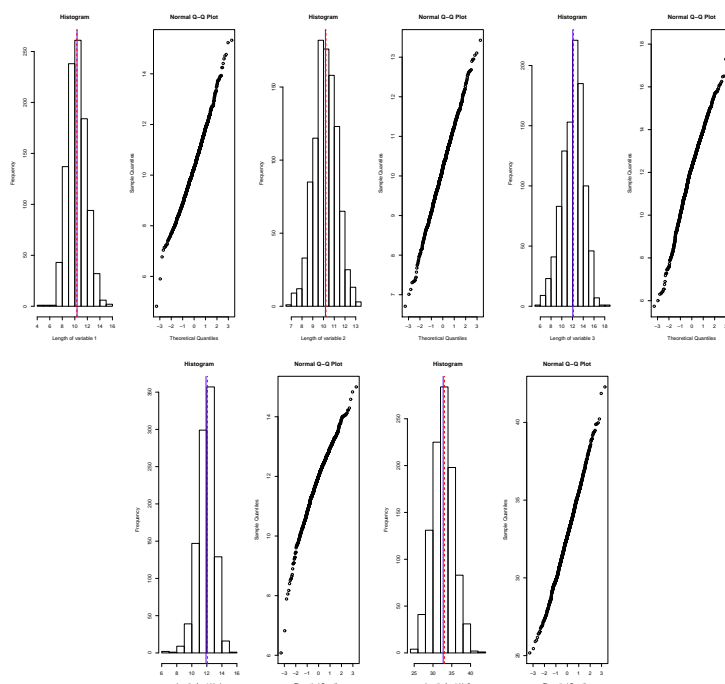


Figura 2.13: Histograma de la longitud de las variables para datos simulados.

Los últimos parámetros analizados para las variables son la contribución relativa a la variabilidad total y las contribuciones relativas del factor al elemento y del elemento al factor (tablas 2.14, 2.15 y 2.16 y figuras 2.14, 2.15 y 2.16). Nuevamente, las diferencias no son altas y se puede observar, como en el caso anterior, que el intervalo t-bootstrap no funciona adecuadamente ya que



presenta límites inferiores negativos. Observando las distribuciones, se encuentran asimetrías en las contribuciones del eje 1 a la variable 5, del eje 2 a las variables 1 y 5 y del eje 3 a las variables 2, 4 y 5. Respecto a las contribuciones de las variables a los ejes se ven asimetrías en la contribución de la variable 1 a los ejes 2 y 3; de la variable 2 al eje 3; de la variable 4 al eje 3 y de la variable 5 a los tres ejes retenidos. Las contribuciones relativas de las variables a la variabilidad total son en general bastante simétricas y siguen una distribución normal.

	V.	Media	SE	Sesgo	t-b inf	t-b sup	perc inf	perc sup	BCa inf	BCa sup
	Obs									
V1	100.00	100.44	14.47	0.44	72.05	128.83	74.35	130.76	74.44	130.77
V2	63.43	64.41	8.62	0.99	47.50	81.32	47.28	80.33	46.45	79.65
V3	118.31	119.14	18.68	0.83	82.48	155.79	85.85	158.57	87.35	160.96
V4	89.78	91.38	14.47	1.60	62.98	119.77	67.17	122.85	64.66	119.41
V5	628.49	624.64	27.30	-3.85	571.07	678.21	569.53	673.50	578.32	678.80

Tabla 2.14: Contribuciones relativas a la variabilidad total de las variables para datos simulados.

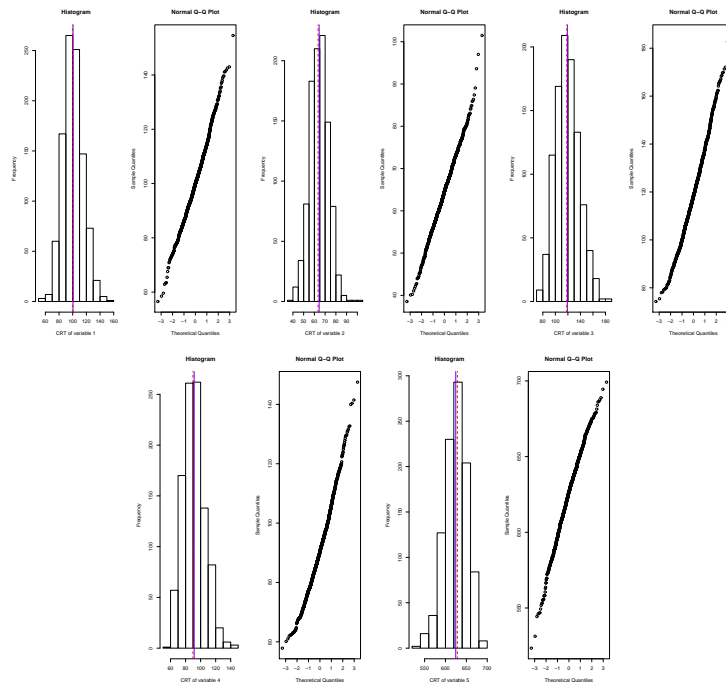


Figura 2.14: Histograma de las contribuciones relativas a la variabilidad total de las variables para datos simulados.

	V. Obs	Media	SE	Sesgo	t-b inf	t-b sup	perc inf	perc sup	BCa inf	BCa sup
E.1										
V1	425.94	427.53	87.15	1.60	256.51	598.56	245.78	585.62	231.05	574.76
V2	728.96	724.12	94.91	-4.85	537.87	910.37	516.33	884.78	498.61	878.52
V3	348.33	346.37	86.49	-1.95	176.65	516.10	173.25	504.84	170.40	504.40
V4	315.60	316.22	84.38	0.62	150.64	481.79	151.99	494.69	147.11	486.66
V5	999.89	999.34	0.80	-0.55	997.76	1000.92	996.99	999.99	999.53	1000.00
E.2										
V1	186.89	204.53	128.09	17.64	-46.83	455.89	6.61	480.04	4.35	469.04
V2	219.47	219.67	100.80	0.19	21.87	417.46	51.92	444.95	67.96	481.53
V3	372.42	373.20	144.71	0.78	89.22	657.18	91.11	641.67	86.51	629.07
V4	619.74	599.70	106.14	-20.04	391.41	807.99	380.67	796.27	418.22	808.41
V5	0.08	0.49	0.75	0.41	-0.98	1.97	0.00	2.79	0.00	0.99
E.3										
V1	387.18	367.94	115.40	-19.24	141.49	594.38	147.53	563.66	162.19	588.92
V2	51.56	56.21	43.08	4.65	-28.32	140.75	1.63	158.28	3.76	174.71
V3	279.25	280.43	136.16	1.17	13.24	547.61	73.52	594.99	94.59	636.13
V4	64.67	84.08	79.80	19.42	-72.50	240.67	0.22	282.47	0.30	286.34
V5	0.03	0.17	0.25	0.14	-0.32	0.66	0.00	0.92	0.00	0.29

Tabla 2.15: Contribuciones relativas de los ejes a las variables para datos simulados.

	V. Obs	Media	SE	Sesgo	t-b inf	t-b sup	perc inf	perc sup	BCa inf	BCa sup
V1										
E.1	54.13	54.94	14.38	0.81	26.71	83.17	28.53	84.81	28.51	84.67
E.2	141.19	155.71	106.89	14.52	-54.05	365.47	4.90	399.36	5.92	405.47
E.3	478.88	458.33	110.76	-20.55	240.99	675.68	216.06	632.15	236.00	640.88
V2										
E.1	58.76	59.57	11.53	0.81	36.94	82.21	36.95	82.75	35.82	81.40
E.2	105.17	102.25	40.91	-2.92	21.97	182.53	29.57	184.04	33.57	195.45
E.3	40.45	45.99	33.75	5.54	-20.25	112.23	1.24	127.93	1.27	128.46
V3										
E.1	52.38	52.83	16.49	0.45	20.48	85.19	24.13	87.70	25.66	89.91
E.2	332.88	334.09	142.01	1.21	55.43	612.76	72.94	600.77	64.75	594.44
E.3	408.66	403.24	147.24	-5.42	114.30	692.17	130.46	679.40	147.33	692.12
V4										
E.1	36.01	37.01	12.28	1.00	12.91	61.11	16.77	64.38	17.32	65.05
E.2	420.36	405.79	81.34	-14.57	246.17	565.41	238.36	543.16	261.81	554.84
E.3	71.81	91.21	76.18	19.39	-58.28	240.69	0.35	266.99	0.13	254.64
V5										
E.1	798.71	795.64	3.55	-3.07	788.68	802.60	786.30	799.48	797.34	799.86
E.2	0.40	2.16	3.04	1.76	-3.81	8.13	0.00	10.82	0.00	4.62
E.3	0.20	1.24	1.71	1.04	-2.12	4.60	0.00	6.09	0.00	2.11

Tabla 2.16: Contribuciones relativas de las variables a los ejes para datos simulados.

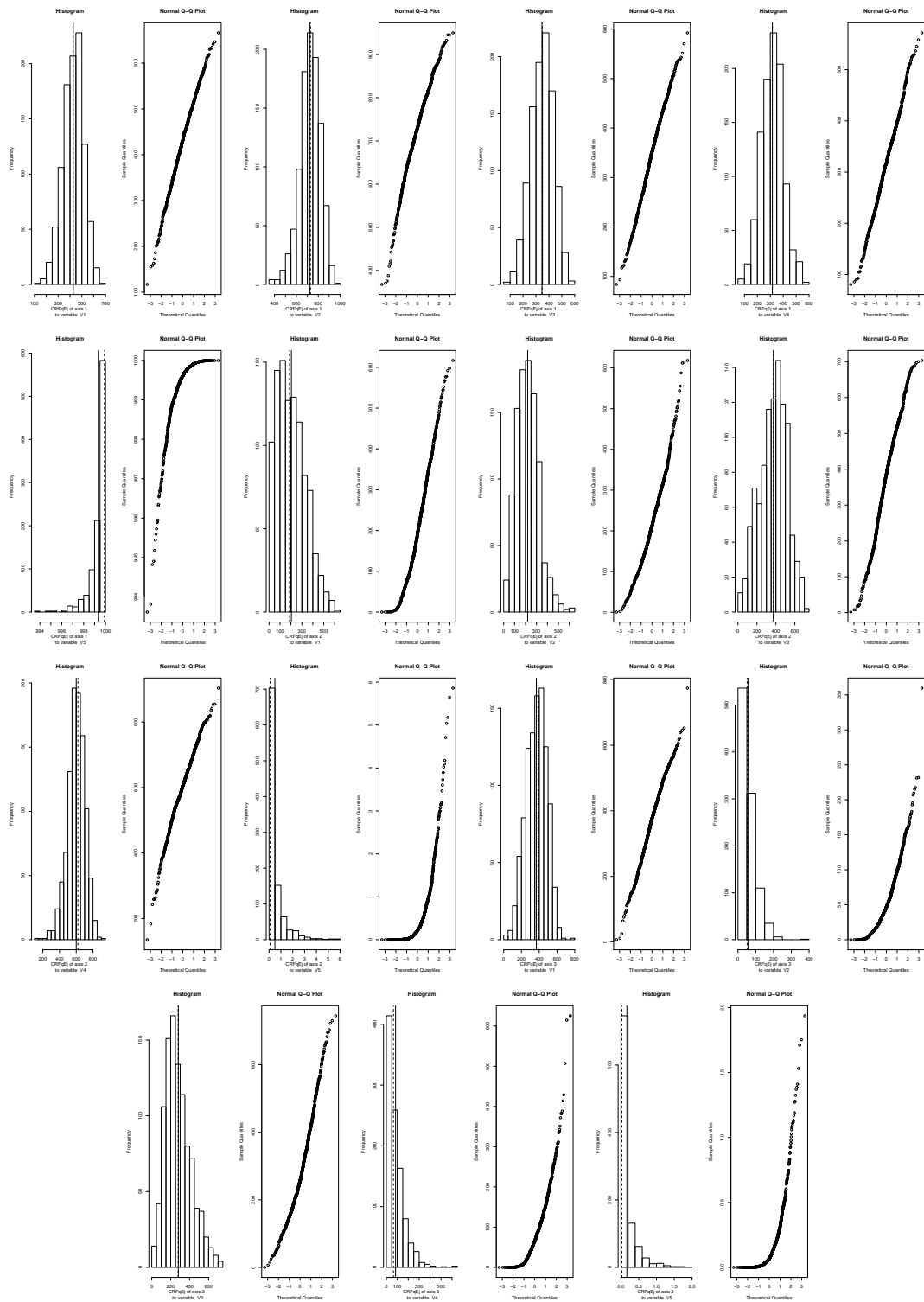


Figura 2.15: Histograma de las contribuciones relativas de los ejes a las variables para datos simulados.

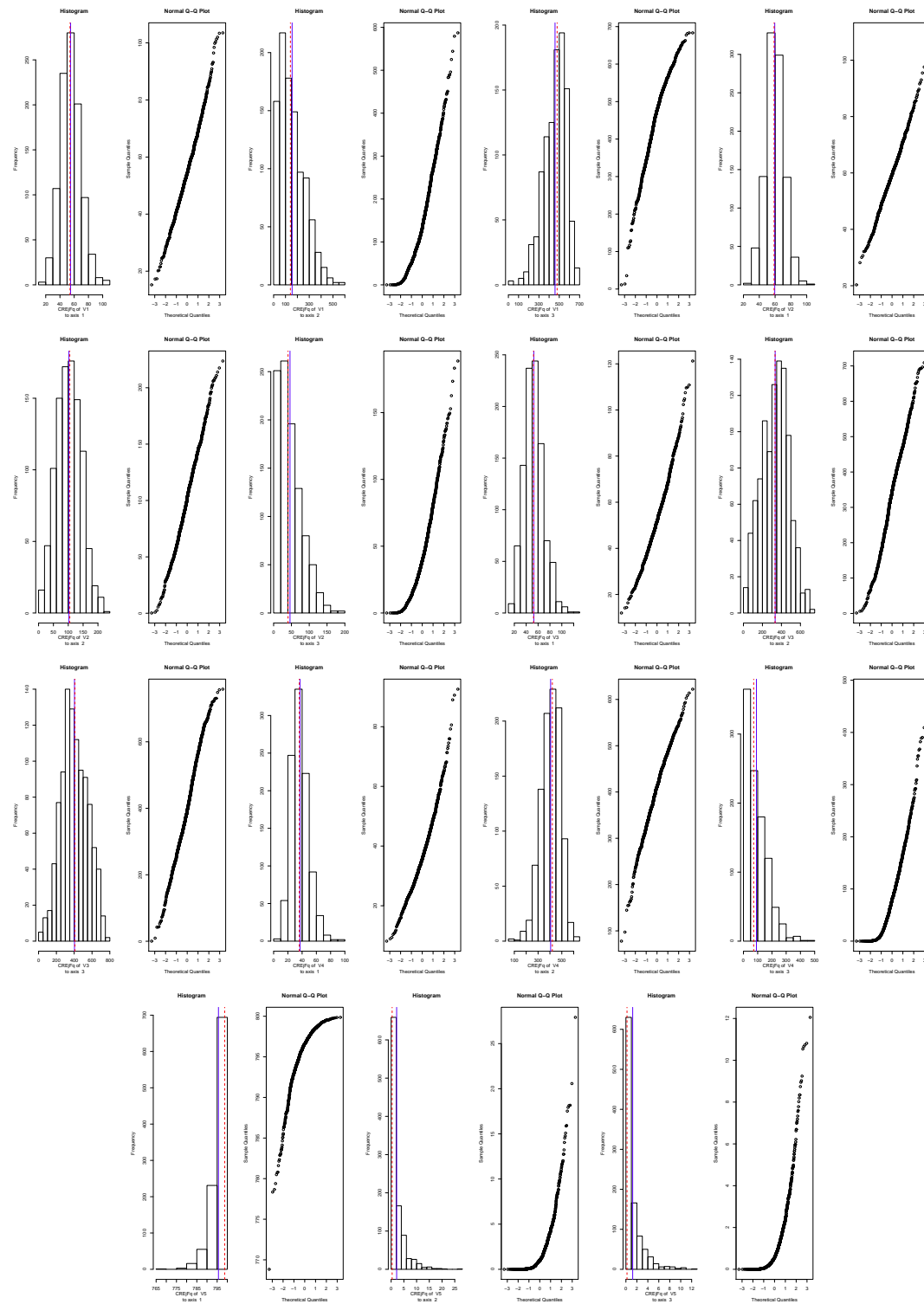


Figura 2.16: Histograma de las contribuciones relativas de las variables a los ejes para datos simulados.

En cuanto a las observaciones se refiere, se ha obtenido el mismo resultado para la calidad de las filas que para las columnas ya que se ha elegido la factorización HJ Biplot. Los resultados obtenidos para la contribución relativa a la variabilidad total de las observaciones y para las contribuciones relativas del factor al elemento y del elemento al factor se muestran en el apéndice A.

Por último, se muestra el gráfico que contiene las coordenadas de las variables calculadas para cada una de las 1000 réplicas bootstrap. Se han utilizado rotaciones Procrustes para que las configuraciones se puedan superponer (figura 2.17). Cada conjunto de coordenadas que representan a la misma variable se ha representado con el mismo color y se ha dibujado un polígono envolvente (convex-hull) que engloba todos sus valores y proporciona visualmente una idea de como de estables son dichas coordenadas.

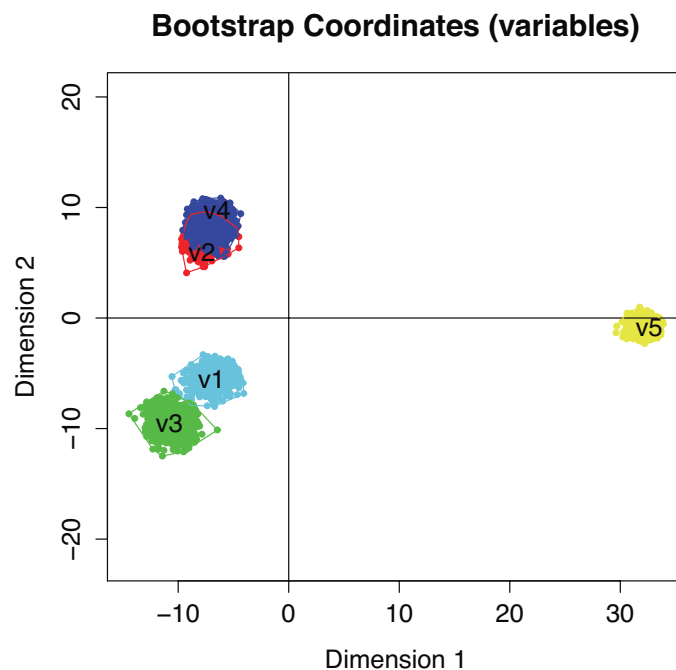


Figura 2.17: Coordenadas bootstrap para las variables datos simulados.

### 2.5.2. Datos Iris

Para el segundo ejemplo se utilizan los datos recogidos por Anderson en 1935 (Anderson, 1935) y que constan de las medidas en centímetros de las variables longitud y anchura del sépalo y longitud y anchura del pétalo de 50 flores provenientes de 3 especies de iris. Se ejecutó el programa, se pidieron 1000 muestras bootstrap y se eligieron todos los parámetros que están disponibles. A continuación se optó por la opción de realizar un HJ Biplot y *Estandarizar por columnas*.

Los resultados obtenidos para el HJ Biplot se detallan a continuación.

En la tabla 2.17 se observa la información proporcionada por los valores propios. En ella se aprecia que el primer eje absorbe la mayor parte de la información (más del 70 %) y con los tres primeros ejes se explica casi el 100 % de la información (99.48 %).

No.	Valor Propio	Variabilidad	Variabilidad Acumulada
1	20.85	72.96	72.96
2	11.67	22.85	95.81
3	4.68	3.67	99.48

Tabla 2.17: Valores propios y variabilidad explicada (%) por cada eje para los datos iris.

La siguiente tabla 2.18 recoge las contribuciones relativas del factor al elemento de las diferentes flores que se han analizado en los tres ejes que se han retenido. Por motivos de espacio sólo se muestran los 10 primeros. La tabla completa se puede consultar en el apéndice A.

En la tabla 2.19 se presentan las contribuciones relativas del factor al elemento columna observándose una buena representación para todas las variables en el plano 1-2.

No.	Eje 1	Eje 2	Eje 3
1	954.10	42.86	3.03
2	894.73	93.90	11.37
3	979.18	20.48	0.34
4	935.39	63.14	1.47
5	931.71	68.25	0.04
6	660.10	339.79	0.11
7	981.14	0.37	18.49
8	988.57	9.87	1.56
9	811.63	185.24	3.13
10	943.75	43.51	12.74

Tabla 2.18: Contribuciones relativas del factor al elemento fila para datos iris.

Variable	Eje 1	Eje 2	Eje 3
Sepal Length	793.52	130.38	76.09
Sepal Width	211.80	779.43	8.77
Petal Length	996.44	0.56	3.00
Petal Width	936.50	4.12	59.38

Tabla 2.19: Contribuciones relativas del factor al elemento columna para datos iris.

La representación HJ Biplot en el primer plano principal se muestra en la figura 2.18. En ella se han resaltado en diferentes colores los tres tipos de iris incluidos en la muestra (Rojo-Setosa, Verde-Versicolor, Naranja-Virginica).

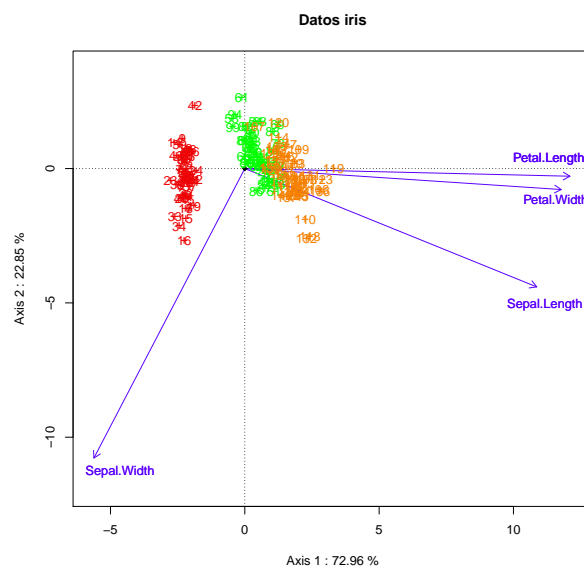


Figura 2.18: Gráfico que muestra la representación HJ Biplot para los datos iris en las dos primeras dimensiones.

La estructura de covariación de las variables pone de manifiesto una alta correlación entre las variables referentes al tamaño y longitud de los pétalos (Petal.Length y Petal.Width) representadas por un ángulo muy pequeño. Ambas variables tienen una correlación alta con la variable Sepal.Length. Sin embargo, no tienen prácticamente relación con la variable Sepal.Width al presentar ángulos próximos a 90 grados.

Se puede observar que el primer eje separa el grupo setosa del resto caracterizándose por longitudes y tamaños de pétalos más pequeños que los otros dos grupos.

Para analizar con más detalle estas relaciones, se muestran los ángulos entre las variables y entre las variables y los ejes en el plano 1-2 (Tablas 2.20 y 2.21).

Variable	Sepal Length	Sepal Width	Petal Length	Petal Width
Sepal Length	0.00	95.47	20.71	18.27
Sepal Width	95.47	0.00	116.18	113.74
Petal Length	20.71	116.18	0.00	2.44
Petal Width	18.27	113.74	2.44	0.00

Tabla 2.20: Ángulos entre variables para datos iris.

Variable	Eje 1	Eje 2
Sepal Length	22.07	67.93
Sepal Width	62.47	27.53
Petal Length	1.35	88.65
Petal Width	3.79	86.21

Tabla 2.21: Ángulos entre variables y ejes para datos iris.

Para estos datos se realizó el remuestreo bootstrap con 1000 réplicas y un nivel de confianza del 95 %.

En primer lugar se puede observar los resultados para la calidad de aproximación para las columnas (tabla 2.22 y figura 2.19). En ella se aprecia que prácticamente no hay diferencia entre el valor observado y el estimado, y que los intervalos de confianza son similares. Esto se traduce en una gran estabilidad



para esta medida.

V.	Media	SE	Sesgo	t-b inf	t-b sup	perc inf	perc sup	BCa inf	BCa sup
Obs									
99.48	99.49	0.08	0.01	99.34	99.64	99.34	99.62	99.27	99.60

Tabla 2.22: Calidades de aproximación de las columnas para datos iris.

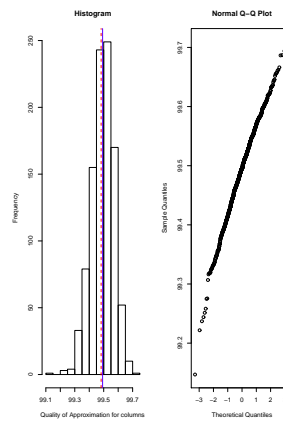


Figura 2.19: Histograma de las calidades de aproximación de las variables para datos iris.

A continuación se muestran los principales resultados para cada uno de los valores propios resultantes de la descomposición (tabla 2.23 y figura 2.20). En ellos se observa de nuevo que hay una diferencia muy pequeña entre los valores observados y los calculados a través de las réplicas bootstrap y que hay concordancia entre ambos tipos de intervalos. También se puede ver que las distribuciones de las réplicas son simétricas y tienden a la normalidad.

	V.	Media	SE	Sesgo	t-b inf	t-b sup	perc inf	perc sup	BCa inf	BCa sup
1	20.85	20.89	0.22	0.03	20.46	21.31	20.50	21.35	20.45	21.27
2	11.67	11.61	0.38	-0.06	10.86	12.36	10.76	12.28	10.97	12.36
3	4.68	4.66	0.31	-0.02	4.05	5.26	4.07	5.30	4.10	5.35
4	1.76	1.74	0.13	-0.02	1.48	1.99	1.50	1.99	1.55	2.09

Tabla 2.23: Valores propios para datos iris.

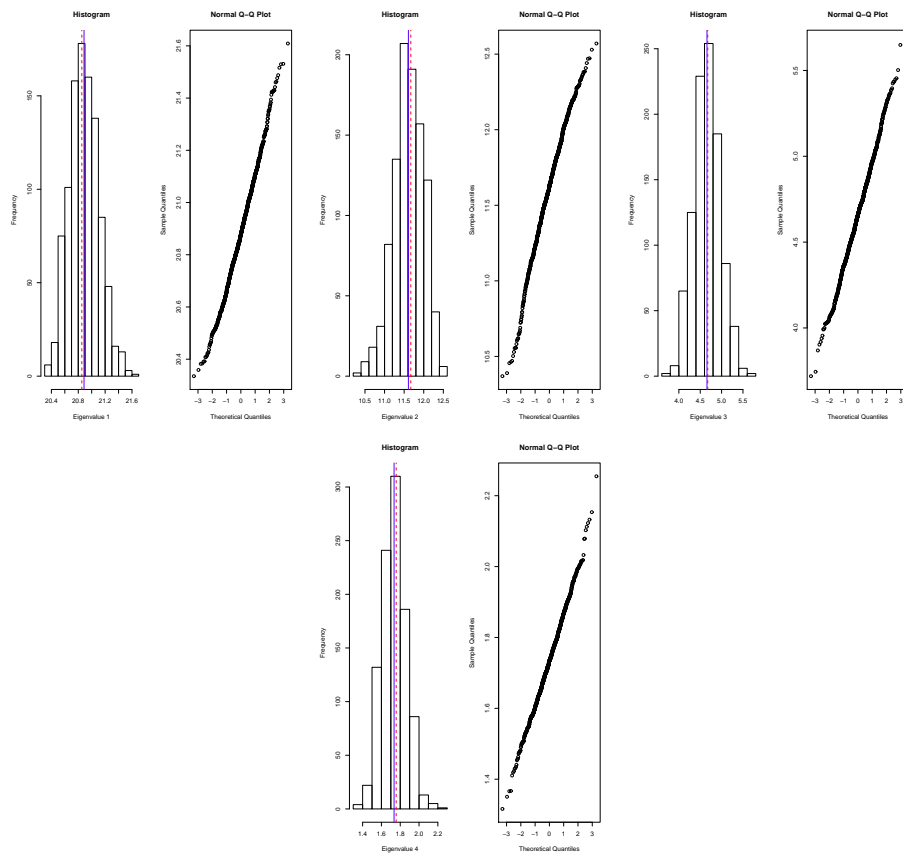


Figura 2.20: Histograma de los valores propios para datos iris.

El siguiente resultado que se muestra es el relativo a los ángulos que forman entre sí las variables (tabla 2.24 y figura 2.21). Se puede observar, al igual que en los casos anteriores que no existen grandes diferencias entre los valores observados y los calculados y que los intervalos tampoco difieren entre sí. En estos parámetros cabe destacar que existe una ligera asimetría en aquellos ángulos próximos a cero. Esto es debido a que se han realizado transformaciones como se explicó en el ejemplo anterior.

	V. Obs	Media	SE	Sesgo	t-b inf	t-b sup	perc inf	perc sup	BCa inf	BCa sup
S.L										
S.W	95.47	95.93	4.40	0.46	87.30	104.56	87.29	104.62	85.61	102.93
P.L	20.71	20.67	2.60	-0.04	15.57	25.77	15.67	25.94	16.27	26.36
P.W	18.27	18.16	3.20	-0.11	11.88	24.43	11.75	24.34	12.15	24.62
S.W										
P.L	116.18	116.60	4.40	0.42	107.97	125.23	107.90	124.67	105.55	123.86
P.W	113.74	114.09	4.41	0.35	105.43	122.75	105.22	122.63	104.03	121.94
P.L										
P.W	2.44	2.54	1.29	0.09	-0.01	5.07	0.16	5.10	0.12	4.94

Tabla 2.24: Ángulos entre variables para datos iris.

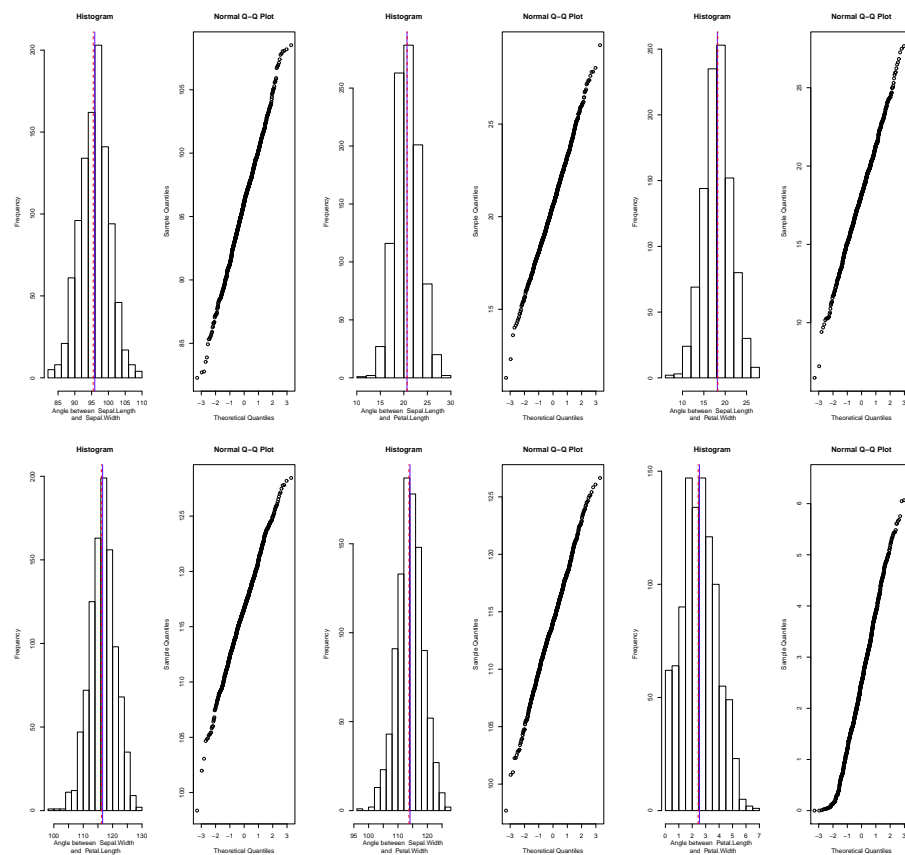


Figura 2.21: Histograma de los ángulos entre variables para datos iris.

El siguiente parámetro al que se le han calculado intervalos de confianza son los ángulos que forman las variables con los dos primeros ejes (tabla 2.25 y figura 2.22). Las apreciaciones en este caso son las mismas que para los ángulos entre variables.

	V. Obs	Media	SE	Sesgo	t-b inf	t-b sup	perc inf	perc sup	BCa inf	BCa sup
S.L										
E.1	22.07	21.95	2.52	-0.11	17.01	26.90	16.84	27.14	17.25	27.47
E.2	67.94	68.05	2.52	0.11	63.10	72.99	62.86	73.17	62.53	72.75
S.W										
E.1	62.47	62.12	5.35	-0.35	51.62	72.62	52.33	72.80	53.83	75.86
E.2	27.53	27.88	5.35	0.35	17.38	38.38	17.20	37.67	14.14	36.17
P.L										
E.1	1.35	1.54	0.92	0.19	-0.26	3.34	0.08	3.41	0.03	3.07
E.2	88.65	88.46	0.92	-0.19	86.66	90.26	86.59	89.92	86.94	89.97
P.W										
E.1	3.79	3.82	1.64	0.03	0.61	7.04	0.48	6.84	0.27	6.61
E.2	86.21	86.18	1.64	-0.03	82.97	89.39	83.16	89.52	83.40	89.73

Tabla 2.25: Ángulos entre variables y los dos primeros ejes para datos iris.

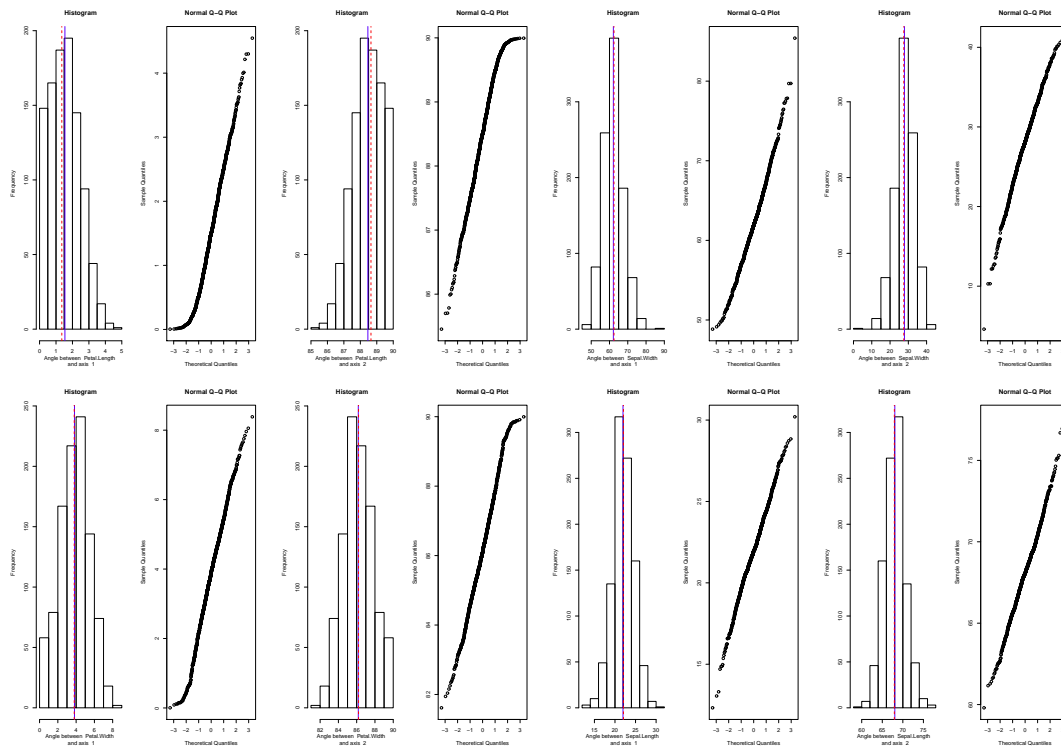


Figura 2.22: Histograma de los ángulos entre variables y ejes para datos iris.

A continuación, se presentan los resultados para la contribución a la variabilidad total de las variables (tabla 2.26 y figura 2.23). Se observa que entre los valores observados y los calculados no hay prácticamente diferencia y que los intervalos de confianza tienen una amplitud muy pequeña lo que nos sugiere una gran estabilidad en estas medidas.

	V. Obs	Media	SE	Sesgo	t-b inf	t-b sup	perc inf	perc sup	BCa inf	BCa sup
S.L	250.95	250.93	0.16	-0.02	250.62	251.24	250.64	251.27	250.71	251.34
S.W	251.22	251.20	0.18	-0.03	250.85	251.54	250.88	251.55	250.96	251.71
P.L	247.96	248.00	0.31	0.04	247.40	248.60	247.39	248.53	247.10	248.42
P.W	249.87	249.88	0.12	0.00	249.64	250.12	249.64	250.12	249.64	250.12

Tabla 2.26: Contribuciones a la variabilidad total de las variables para datos iris.

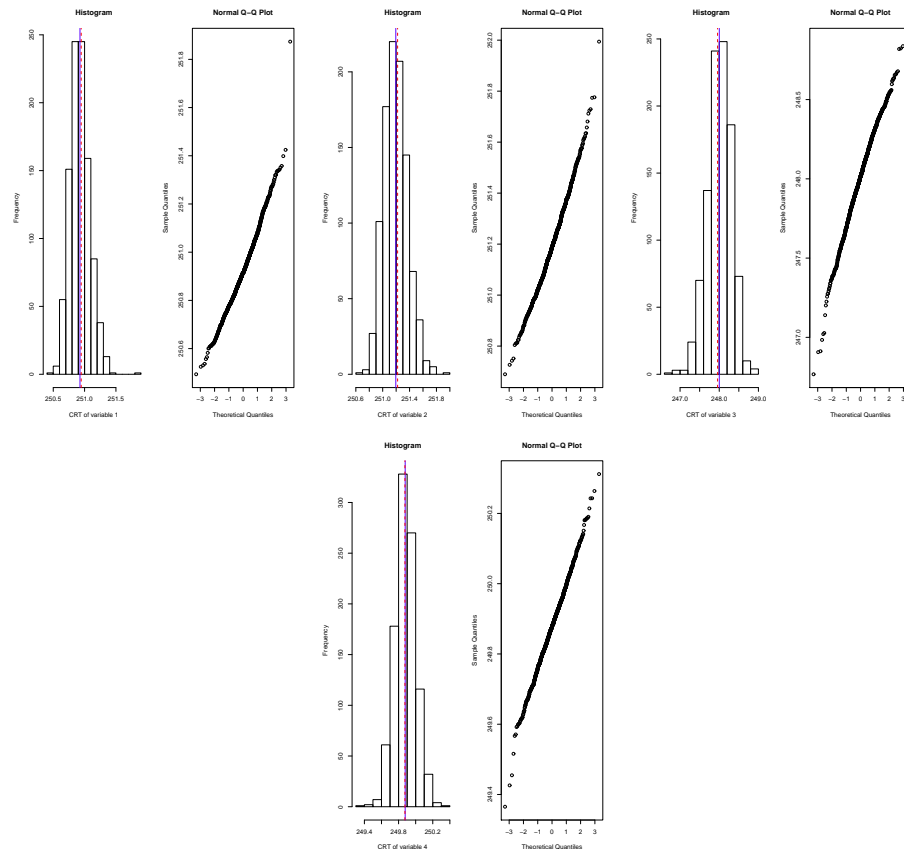


Figura 2.23: Histograma de las contribuciones relativas de las variables a la variabilidad total para datos iris.

También se dispone de medidas de precisión para la longitud de las variables. En la tabla 2.27 y figura 2.24 se muestran dichos resultados. Se observa que los sesgos no son superiores a 0.01, las desviaciones estándar no superan el 0.07 y los intervalos de confianza calculados tienen amplitudes muy pequeñas lo que nos permite asegurar que las longitudes de dichas variables son estables. En cuanto a las distribuciones, presentan una ligera asimetría.

Los últimos parámetros que se han estimado para las variables han sido las

	V. Obs	Media	SE	Sesgo	t-b inf	t-b sup	perc inf	perc sup	BCa inf	BCa sup
S.L	11.73	11.73	0.07	0.00	11.59	11.86	11.58	11.85	11.56	11.84
S.W	12.15	12.15	0.02	0.00	12.11	12.19	12.11	12.18	12.10	12.18
P.L	12.11	12.11	0.01	0.00	12.08	12.14	12.08	12.13	12.07	12.13
P.W	11.81	11.81	0.05	0.00	11.71	11.91	11.70	11.90	11.69	11.89

Tabla 2.27: Longitud de las variables para datos iris.

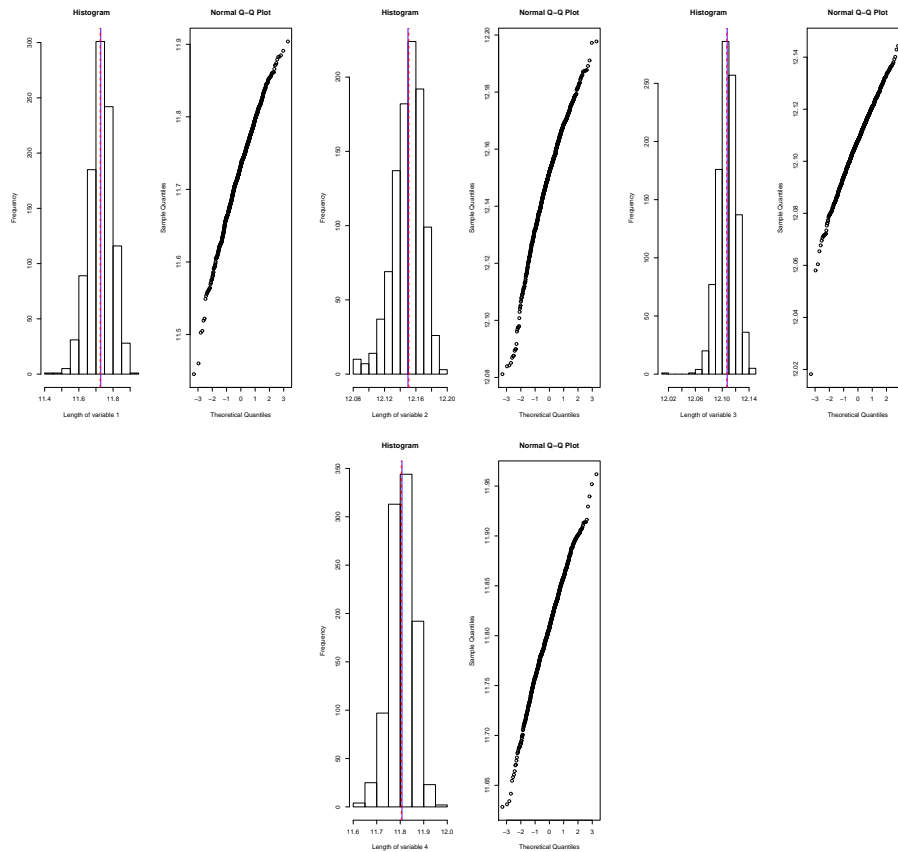


Figura 2.24: Histograma de la longitud de las variables para datos iris.

contribuciones relativas del elemento al factor y las contribuciones relativas del factor al elemento (tablas 2.28 y 2.29 y figuras 2.25 y 2.26). Nuevamente se puede apreciar que existen pequeñas diferencias entre los valores observados y los calculados a partir del remuestreo bootstrap. En estos parámetros hay que destacar que en algunos casos los intervalos t-bootstrap y los basados en percentiles no concuerdan. Además hay extremos inferiores de los intervalos t-bootstrap que son negativos. Esto es debido a que la distribución de los valores

calculados tiene mucha asimetría y los intervalos t-bootstrap no son adecuados en estos casos.

	V. Obs	Media	SE	Sesgo	t-b inf	t-b sup	perc inf	perc sup	BCa inf	BCa sup
S.L										
E.1	271.51	270.92	11.74	-0.60	247.88	293.95	247.74	295.39	249.78	296.68
E.2	142.44	145.14	35.03	2.70	76.40	213.89	79.92	221.08	80.06	221.91
E.3	517.78	515.61	25.06	-2.16	466.44	564.79	462.90	560.47	464.42	564.21
S.W										
E.1	72.55	75.23	23.70	2.68	28.72	121.74	30.38	121.07	20.93	114.06
E.2	852.48	848.10	35.79	-4.38	777.86	918.33	772.05	915.55	774.66	920.22
E.3	59.72	61.01	18.84	1.28	24.04	97.97	28.52	103.82	29.08	105.32
P.L										
E.1	336.88	335.82	7.27	-1.06	321.56	350.08	320.47	348.55	323.60	350.81
E.2	0.60	1.12	1.20	0.52	-1.23	3.47	0.00	4.10	0.00	3.45
E.3	20.20	21.43	8.85	1.23	4.07	38.79	6.86	41.51	5.92	39.08
P.W										
E.1	319.06	318.03	7.58	-1.03	303.17	332.90	302.58	331.88	305.59	334.04
E.2	4.48	5.65	4.41	1.16	-3.01	14.30	0.07	16.90	0.01	14.22
E.3	402.30	401.96	24.94	-0.35	353.02	450.89	354.77	450.73	356.03	451.81

Tabla 2.28: Contribuciones relativas del elemento columna al factor para datos iris.

	V. Obs	Media	SE	Sesgo	t-b inf	t-b sup	perc inf	perc sup	BCa inf	BCa sup
E.1										
S.L	793.52	793.94	28.21	0.41	738.59	849.29	736.90	847.53	730.70	842.04
S.W	211.80	221.62	73.65	9.82	77.10	366.14	87.07	368.65	59.40	345.29
P.L	996.44	995.84	1.84	-0.61	992.23	999.44	991.66	998.72	993.21	999.16
P.W	936.50	936.00	9.30	-0.51	917.75	954.24	915.73	953.23	916.66	953.47
E.2										
S.L	130.38	130.52	28.28	0.14	75.02	186.02	76.69	191.19	80.22	197.56
S.W	779.43	769.44	74.58	-10.00	623.09	915.78	619.11	905.45	645.27	937.14
P.L	0.56	0.98	0.99	0.42	-0.97	2.92	0.00	3.52	0.00	2.84
P.W	4.12	4.94	3.63	0.82	-2.19	12.06	0.07	13.47	0.02	12.40
E.3										
S.L	76.10	75.54	10.86	-0.55	54.24	96.85	55.83	98.66	58.56	101.88
S.W	8.77	8.94	3.06	0.17	2.94	14.94	3.83	15.75	4.26	16.98
P.L	3.00	3.19	1.43	0.19	0.39	5.99	0.93	6.35	0.85	6.20
P.W	59.38	59.07	8.25	-0.31	42.87	75.26	44.12	76.50	45.01	78.46

Tabla 2.29: Contribuciones relativas del factor al elemento columna para datos iris.

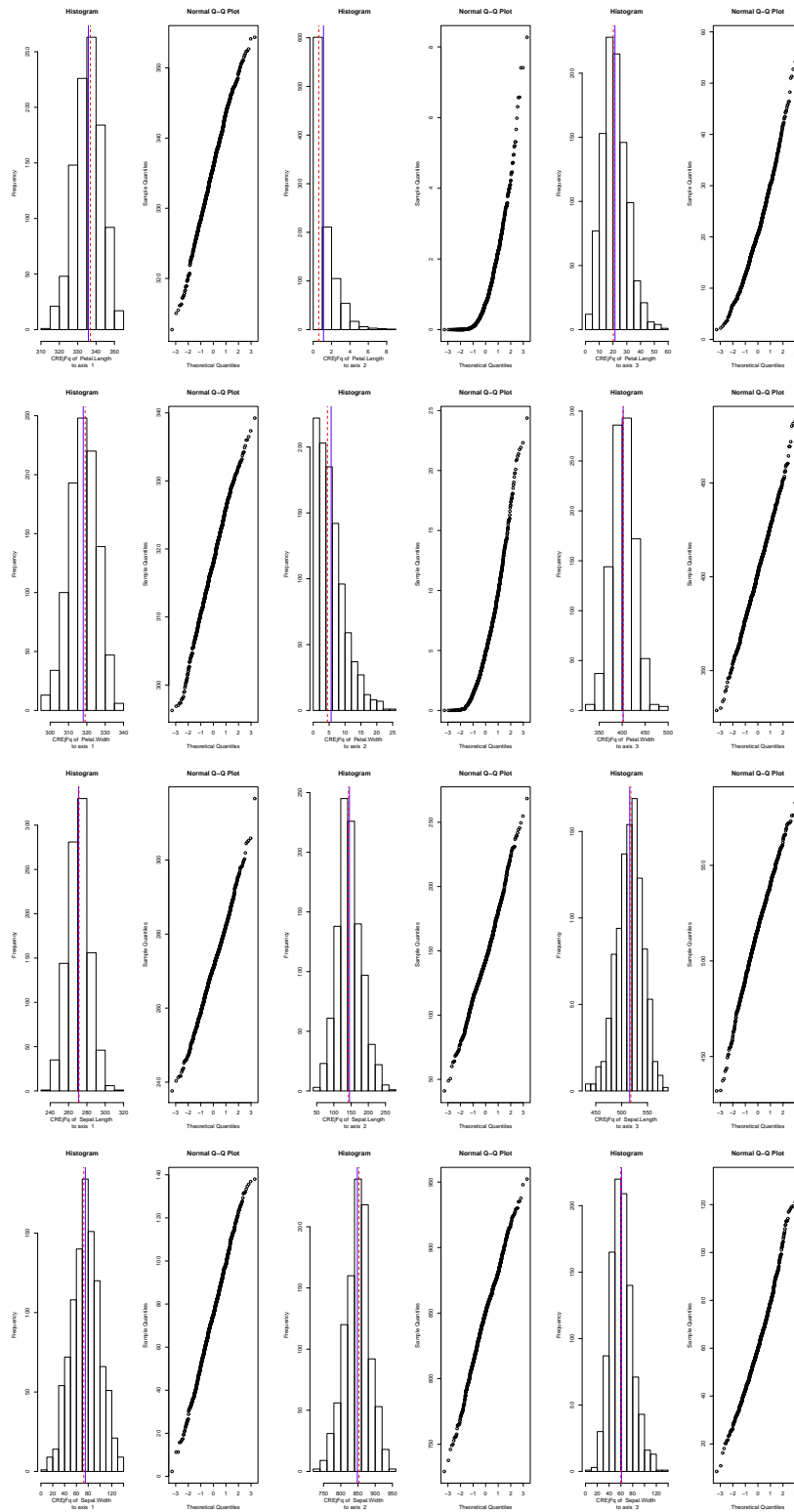


Figura 2.25: Histograma de las contribuciones relativas de las variables a los ejes para datos iris.



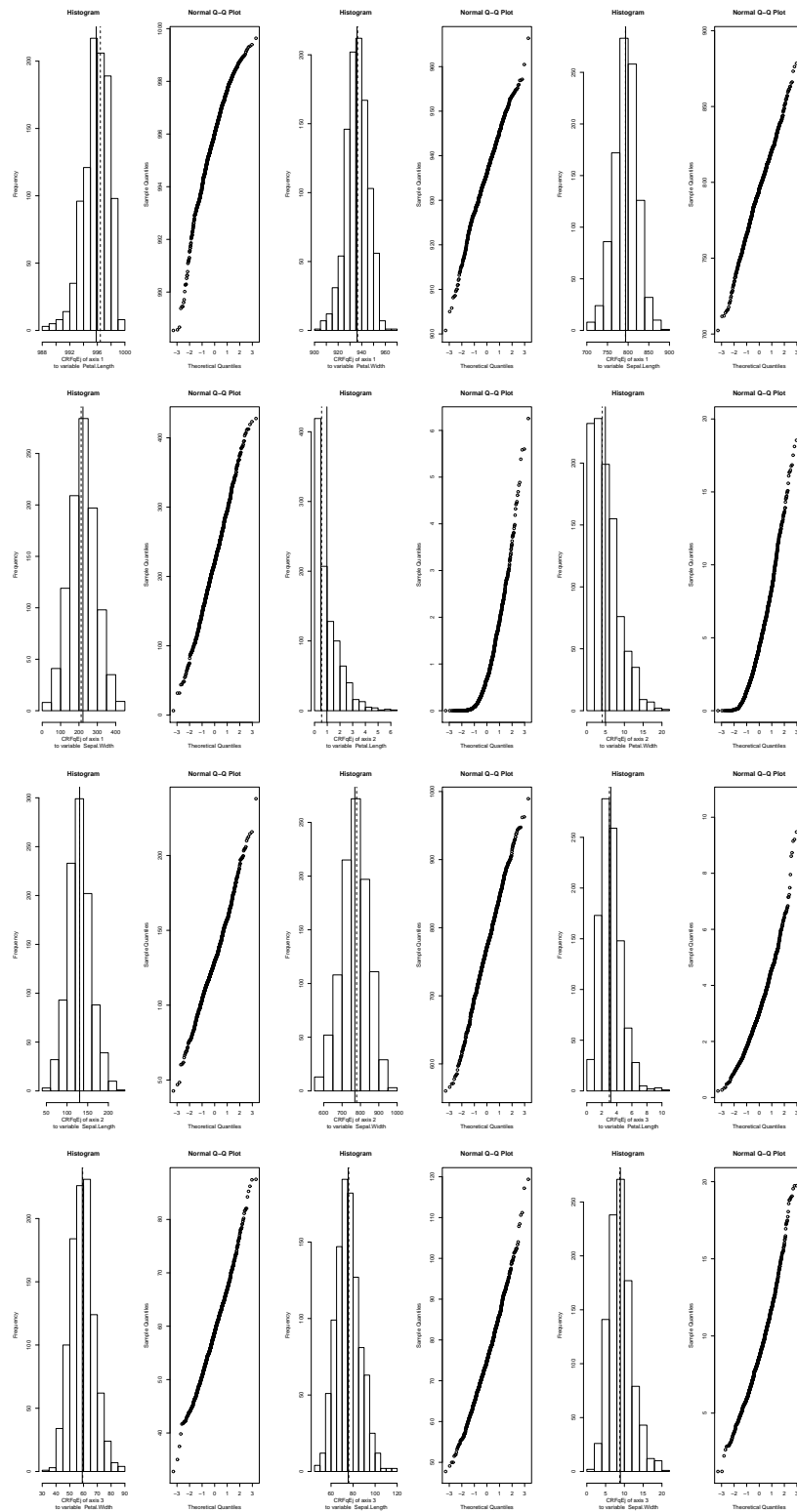


Figura 2.26: Histograma de las contribuciones relativas de los ejes a las variables para datos iris.

También se han calculado medidas de precisión para parámetros referentes a los individuos. La tabla 2.30 y figura 2.27 muestran la calidad de aproximación para las filas. Como se puede apreciar, los resultados son los mismos que para la calidad de las columnas puesto que se ha elegido realizar un HJ Biplot.

V. Obs	Media	SE	Sesgo	t-b inf	t-b sup	perc inf	perc sup	BCa inf	BCa sup
99.48	99.49	0.08	0.01	99.34	99.64	99.34	99.62	99.27	99.60

Tabla 2.30: Calidad de aproximación de las filas para datos iris.

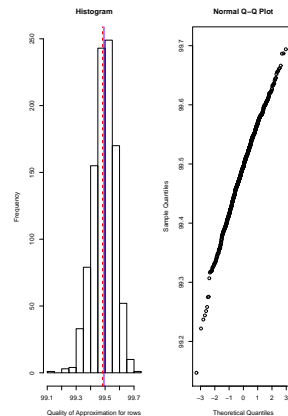


Figura 2.27: Histograma de la calidad de aproximación de los individuos para datos iris.

Se han calculado las contribuciones relativas del elemento al factor y las contribuciones relativas del factor al elemento (apéndice A). Al igual que en el caso de las variables, se puede apreciar que existen pequeñas diferencias entre los valores observados y los calculados a partir del remuestreo bootstrap y que algunos de los intervalos inferiores t-bootstrap son negativos obteniéndose la misma conclusión que en el caso de las variables.

Se presentan también, los resultados para la contribución relativa a la variabilidad total de los individuos (apéndice A). Nuevamente se observan diferencias mínimas entre los valores observados y los calculados a partir del

remuestreo bootstrap, deduciéndose una gran precisión para estos datos.

Por último, se muestra el gráfico que contiene la posición de las variables para las 1000 réplicas bootstrap. Se han utilizado rotaciones Procrustes con el objetivo de eliminar el posible efecto espejo que pudieran presentar dichas posiciones (figura 2.28). En este gráfico se ve que las regiones que ocupan los conjuntos que representan a cada variable son muy compactas lo que nos permite afirmar que las configuraciones son muy parecidas a lo largo de todas las muestras bootstrap.

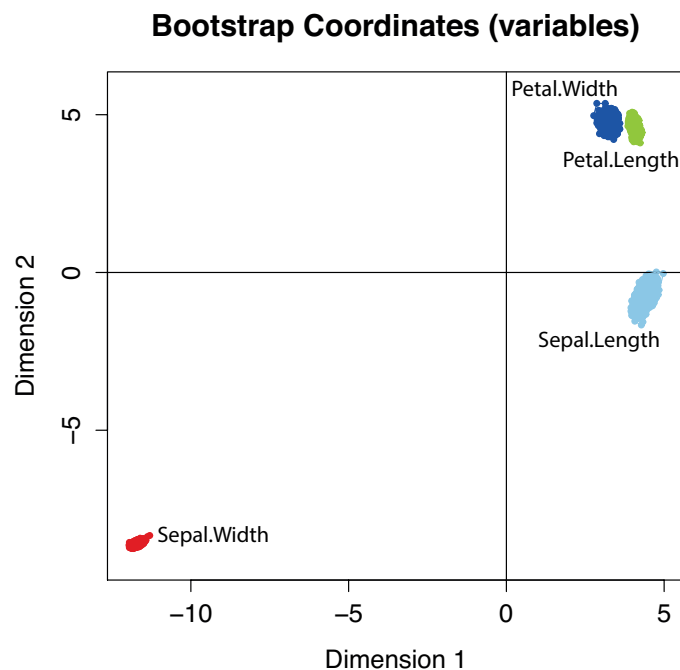


Figura 2.28: Coordenadas bootstrap para las variables datos iris.

En este capítulo, se ha realizado una revisión bibliográfica de los métodos Biplot (Gabriel, 1971; Galindo, 1986) y sus aplicaciones. Al encontrar controversia en cuanto a su validez al haber autores que lo consideran una técnica simplemente exploratoria, se ha propuesto una versión inferencial basada en los métodos Bootstrap. Para facilitar su puesta en práctica se ha desarrollado un nuevo software en el lenguaje R que permite realizar análisis Biplot y complementar sus

resultados con medidas de precisión. Dicho software permite la interacción con el usuario a través de ventanas, botones y menús. Para mostrar el funcionamiento y la interpretación de los resultados que proporciona el software se ha aplicado sobre un conjunto de datos simulados y un conjunto de datos reales.

## Capítulo 3

# ANÁLISIS DE PARES DE MATRICES CON UNA DIMENSIÓN COMÚN



## 3.1. Introducción

En el capítulo anterior se ha mostrado una solución al problema de las estimaciones puntuales de técnicas multivariantes aplicadas a datos de dos vías. En el presente capítulo se da solución al mismo problema en el caso de que la información disponible sean dos matrices que compartan una vía. A continuación, se realiza una revisión bibliográfica sobre dicho contexto.

El análisis del gradiente (Whittaker, 1967) es un conjunto de técnicas que permiten el estudio de las relaciones entre la composición de una comunidad y las variables ambientales. Esta información viene expresada en dos tablas; una contiene la composición de las comunidades en unos lugares determinados y otra que contiene las medidas de las variables ambientales en dichos lugares. Hay dos tipos de análisis del gradiente de acuerdo a la información utilizada: análisis indirecto del gradiente, que analiza datos ecológicos sin tener en cuenta información ambiental que podría estar incluida de un modo pasivo después de obtener los ejes factoriales; y análisis directo del gradiente, que analiza datos ecológicos con información ambiental, ambos de un modo activo. En los métodos incluidos en este último tipo de análisis del gradiente, las puntuaciones de los lugares se restringen para ser combinaciones lineales de las variables ambientales.

Una de las técnicas más utilizadas en el análisis indirecto del gradiente es el análisis de Correspondencias (CA, Benzècri, 1973), también llamado método de las medias recíprocas debido a que el algoritmo utilizado para encontrar la solución calcula repetidas medias de puntuaciones de lugares y especies. Esta técnica tiene la desventaja de otorgar un excesivo peso a las especies raras debido al uso de la métrica Chi-cuadrado. Una modificación del CA, llamada análisis de Correspondencias Detrended (DCA, Hill y Gauch, 1980) ha sido desarrollada para corregir el efecto arco. Dicho efecto es la tendencia que tienen los puntos en

un CA a formar una curva.

Cuando se quiere tomar en cuenta la información ambiental, las técnicas más comunes son el análisis de la Redundancia (RDA - Rao, 1964, Van den Wollenberg, 1977), un método lineal que es una forma restringida del análisis de Componentes Principales y el análisis Canónico de Correspondencias (CCA - Ter Braak, 1986) que es una versión restringida del CA y es apropiado tanto para modelos lineales como de respuesta unimodal. El CCA es la unión del CA y de la regresión múltiple. Una de las mayores ventajas de esta técnica es la visualización gráfica: el triplot. Se pueden mostrar la información de los tres conjuntos de datos estudiados: lugares, especies y variables ambientales. Al igual que en el caso del CA, se puede encontrar el efecto arco en el segundo eje. Para resolver este problema, se ha desarrollado una versión detrended del CCA (Ter Braak, 1986) que utiliza esencialmente el mismo algoritmo que en el DCA. Otra desventaja que posee el CCA es que necesita que el número de lugares en la muestra objeto de estudio debe ser mayor que el número de variables. El análisis de la Coinercia (COIA) (Chessel y Mercier, 1993; Dolédec y Chessel, 1994) es otro método que puede ser utilizado para estudiar este tipo de datos. Esta técnica trata las especies y las variables ambientales de un modo simétrico y analiza la covariación, por lo que no están involucrados ni modelos de regresión ni de predicción. En este método no se requiere la presencia de mayor número de lugares que de variables ambientales como ocurre en el caso del CCA. Como alternativa al análisis de la Coinercia, Ter Braak y Verdonschot, 1995 desarrollaron una versión del CCA basada en mínimos cuadrados parciales (en inglés Partial Least Square, PLS). RLQ (Dolédec et al., 1996) es una extensión del análisis de la Coinercia para analizar relaciones entre patrones de las especies y de las variables ambientales a través de una matriz que contiene especies por lugares. Dray et al., 2002 desarrollaron una adaptación de este método para estudiar



las relaciones entre dos conjuntos de datos provenientes de diferentes esquemas de muestreo con distintos lugares. El análisis de la Coinercia también ha sido adaptado para estudiar las variaciones espacio-temporales de la composición de la comunidad ecológica concanteniendo  $m$  matrices que contienen información de  $m$  conjuntos de variables ambientales medidas sobre el mismo conjunto de lugares de muestreo. Este método se llama análisis de la Coinercia Múltiple (Chessel y Hanafi, 1996). Dray et al., 2003 desarrollaron un método basado en la utilización conjunta del análisis de la Coinercia y el análisis Procrustes denominado Coinercia Procrustea. Para medir las relaciones entre las variables ambientales y varios conjuntos de especies, Lafosse y Hanafi, 1997 propusieron el análisis de la Concordancia. En el contexto del análisis de tres vías, el método de la Coinercia ha sido extendido para analizar pares de tablas: STATICO (STATIS y Coinercia, Simier et al., 1999; Thioulouse et al., 2004 y más recientemente COSTATIS (Coinercia y STATIS, Thioulouse, 2011), son una interesante elección que permiten el estudio de las relaciones entre especies y variables ambientales durante varios períodos de tiempo muestreados. Un método relacionado con ellos es el análisis de la Co-correspondencia (Ter Braak, 2004) que maximiza la covarianza ponderada entre los promedios ponderados de las puntuaciones para las especies de una comunidad y los promedios ponderados para las especies de la otra comunidad. Una aproximación basada en mínimos cuadrados generalizados ha sido desarrollada por Böckenholt y Böckenholt, 1990 para incorporar restricciones lineales en puntuaciones de filas y columnas estandarizadas proporcionadas por el análisis Canónico de una tabla de contingencia. Si lo que se pretende es capturar las interacciones o las diferencias, Mendes, 2011 propone el CO-TUCKER, que combina el análisis de la Coinercia (Dolédec y Chessel, 1994) y el modelo TUCKER3 (Tucker, 1966) para analizar pares de  $k$ -tablas y capturar la parte no estable de la información (interacciones).

El CA es aplicable al estudio de una tabla de contingencia para analizar la asociación entre dos variables categóricas. Cuando se tienen más de dos variables categóricas, Benzècri, 1973 y Benzècri, 1977 propusieron el análisis de Correspondencias Múltiple (MCA). Una alternativa a esta técnica es el análisis de Correspondencias Conjunto (JCA) propuesto por Greenacre, 1988, 1990, 1991. Para obtener la solución a este método se han desarrollado varios algoritmos (Boik, 1996; Greenacre, 1988; Tateneni y Browne, 2000).

Cuando el investigador no está interesado en incluir todas las variables en el estudio, se han desarrollado técnicas parciales donde se permite prescindir de los efectos de una o más variables para realizar el análisis. Estas técnicas incluyen el CA parcial (Yanai, 1987); el CCA parcial (Ter Braak, 1988b); el RDA parcial (Liu, 1997) y el MCA parcial (Yanai y Maeda, 2002) entre otros.

Como alternativa al CA, se ha propuesto el análisis de Correspondencias No Simétrico como técnica para evaluar una tabla de contingencia con una estructura de dependencia (NSCA, Gimaret-Carpentier et al., 1998; Kroonenberg y Lombardo, 1999). Como alternativa no simétrica del CCA Willems y Galindo, 2008 propusieron el análisis Canónico de Correspondencias no Simétrico (CNCA). Cuando hay más de dos variables categóricas, también se ha propuesto una extensión no simétrica llamada análisis Múltiple de Correspondencias No Simétrico (D'Ambra y Lauro, 1989, 1992; Lauro y D'Ambra, 1984). Simonetti y Gallo, 2002 desarrollaron interpretaciones alternativas a este método utilizando mínimos cuadrados parciales. Lombardo et al., 2007 adaptan el NSCA para analizar datos que contengan variables en escala ordinal utilizando polinomios ortogonales. D'Ambra y Lauro, 1989 y Lombardo et al., 1996 lo adaptan en el caso de tener una tabla de contingencia de tres vías, D'Ambra y Lauro, 1993 desarrollaron una versión normalizada y D'Ambra et al., 2006 propusieron una versión para tablas de contingencia de tres vías con información en escala ordinal.

Siciliano et al., 1993 desarrollaron una versión del NSCA basada en distribuciones probabilísticas. Cuando los datos objeto de estudio contienen ruido, Yanai, 1987 propone el NSCA parcial que elimina los efectos de las variables responsables del ruido del análisis de una tabla de contingencia. Esposito, 1997 desarrolló una versión no simétrica del análisis de la Coinercia llamada análisis de la Coinercia No Simétrico. Hay casos en los que las variables predictoras poseen relaciones interesantes entre ellas. Esta información puede ser incluida en el análisis como restricciones lineales (Böckenholt y Böcknholt, 1990; Böckenholt y Takane, 1994; D´Ambra y Beh, 2010; Takane et al., 1991; Yanai, 1986). Para obtener mejores estimadores de los parámetros, Takane y Jung, 2009 propusieron una versión del NSCA en los que se introduce un tipo de regularización ridge. Este tipo de regularización también se ha incluido en el MCA (Takane y Hwang, 2006), en el RDA (Takane y Hwang, 2007) y en el CA parcial y/o restringido (Takane, 2008). Otras versiones restringidas del MCA se presentan en Gifi, 1990; Hwang y Takane, 2002; Nishisato, 1984; Van Buuren y De Leeuw, 1992; Yanai, 1998.

### 3.1.1. Software

Entre los softwares disponibles para la utilización de estas técnicas se encuentran DECORANA (Hill, 1979a) que es un programa desarrollado en `fortran` para poder ejecutar el CA y su versión `detrended`; TWINSPAN, un programa escrito en `fortran` para datos multivariantes organizados en una tabla ordenada para clasificación de individuos y atributos (Hill, 1979b). CANOCO (Ter Braak, 1985a,b, 1988a) un programa en el lenguaje `fortran` para la ejecución del CCA y el DCA y ADE-4 (Thioulouse et al., 1997).

En el entorno R (R-Team, 2014), hay varios paquetes para ejecutar algunas de estas técnicas como `calibrate` (Graffelman, 2012) que proporciona una función

para el RDA; `ade4` (Dray y Dufour, 2007) que contiene técnicas de visualización gráfica y análisis de datos ecológicos, su interfaz gráfica de usuario Tcl/Tk, llamada `ade4TkGUI` (Thioulouse y Dray, 2007) y su versión para analizar datos de microarrays `made4` (Culhane et al., 2005) y `omicade4` (Meng et al., 2014); `anacor` (De Leeuw y Mair, 2009) que permite ejecutar CA simple y canónico sobre una tabla de frecuencias de dos vías (con posibles valores perdidos) utilizando la descomposición en valores singulares. También proporciona diferentes métodos de escalado (standard, centroid, Benzecri, Goodman) así como varios gráficos que incluyen elipsoides de confianza; `vegan` (Oksanen et al., 2013) que permite ejecutar métodos de ordenación, análisis de la Diversidad y otras funciones para el análisis de comunidades ecológicas; `ca` (Greenacre y Nenadic, 2012) y `caGUI` (Markos, 2012) que han sido desarrollados para la computación y visualización de CA simple, múltiple y conjunto; `cabootcrs` (Ringrose, 2012) y (Ringrose, 2013) que realiza un CA sobre una tabla de contingencia de dos vías y produce regiones de confianza elípticas basadas en los métodos Bootstrap alrededor de las coordenadas proyectadas para los puntos que representan las categorías; `CAvariants` (Lombardo y Beh, 2014) que permite la ejecución de seis variantes del CA: simple, individualmente ordenado, doblemente ordenado, no simétrico, individualmente ordenado no simétrico y doblemente ordenado no simétrico; `cocorresp` (Simpson, 2009) que ajusta modelos predictivos y simétricos basados en análisis de Co-correspondencias para relacionar una matriz con otra; `SimultAnR` (Zárraga y Goitisoló, 2011, 2013) que ejecuta CA simple y análisis Simultáneos; `soc.ca` (Larsen, 2014) que es un paquete para realizar CA específicos en el ámbito de las ciencias sociales.

A continuación se exponen con más detalle los principales aspectos teóricos del CCA, el CNCA y el análisis de la Coinercia.

## 3.2. Análisis Canónico de Correspondencias

Como se señalaba en la introducción, cuando se quiere tener en cuenta la información ambiental contenida en una matriz, una de las técnicas disponibles es el CCA (Ter Braak, 1986), que proviene de la combinación del CA con la regresión múltiple y es válido tanto para modelos lineales como de respuesta unimodal. En esta sección se exponen los aspectos teóricos principales de dicha técnica.

Sea la información de  $q$  especies y  $p$  variables ambientales cuantitativas medidas sobre  $n$  lugares, expresada en las siguientes matrices:

- matriz  $\mathbf{Y}$ , de orden  $n \times q$ , cuyo elemento genérico  $y_{ik}$ ,  $i = 1, \dots, n$ ,  $k = 1, \dots, q$ , contiene información de la  $k$ -ésima especie en el  $i$ -ésimo lugar, con matriz de correspondencias asociada  $\mathbf{F} = y^{-1}\mathbf{Y}$ , siendo  $y$  la suma total de la matriz  $\mathbf{Y}$ ;
- matriz  $\mathbf{Z}$ , de orden  $n \times p$ , cuyo elemento genérico  $z_{ij}$ ,  $i = 1, \dots, n$ ,  $j = 1, \dots, p$ , contiene el valor de la  $j$ -ésima variable ambiental cuantitativa en el  $i$ -ésimo lugar.

Sea  $\mathbf{D}_n = \text{diag}(f_{1.}, f_{2.}, \dots, f_{n.})$  la matriz cuya diagonal es el vector  $\mathbf{f}_n = \mathbf{F}\mathbf{1}_q$ , que son los marginales fila (lugares) de la matriz  $\mathbf{F}$ ; con  $\mathbf{1}_q$  vector de  $q$  unos.  $\mathbf{1}_n$  es un vector de  $n$  unos;  $\mathbf{f}_q = \mathbf{F}^\top \mathbf{1}_n$ , es el vector de marginales de las  $q$  columnas (especies)  $f_{.k}$ ,  $k = 1, \dots, q$ .

A partir de las matrices calculadas anteriormente, se define la matriz:

$$\mathbf{W} = \mathbf{D}_q^{-1} \mathbf{F}^\top \mathbf{Z} \quad (3.1)$$

que se denomina matriz de promedios ponderados y es de orden  $q \times p$ . Su cálculo se representa en la figura 3.1.

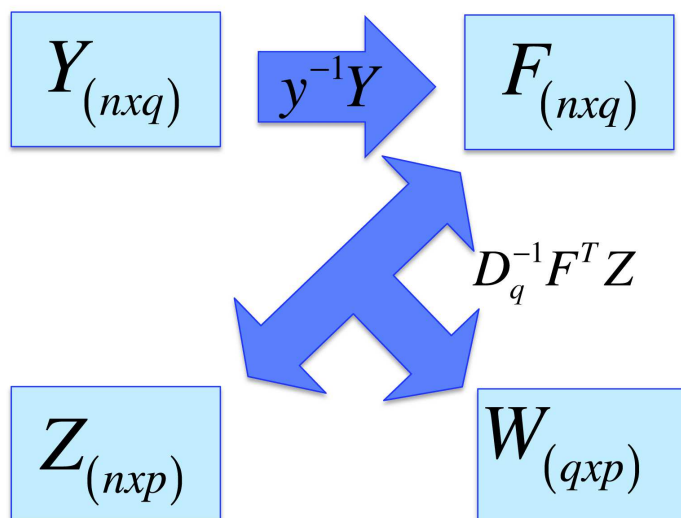


Figura 3.1: Esquema de la obtención de  $\mathbf{W}$ .

Dicha matriz  $\mathbf{W}$  se pondera con  $\mathbf{D}_q$  (matriz que contiene los marginales de las especies) y con  $\mathbf{S} = (\mathbf{Z}^\top \mathbf{D}_n \mathbf{Z})^{-1}$  (inversa de la matriz de covarianzas de las variables ambientales), para que dichas variables ambientales sean invariantes a transformaciones lineales (no singulares) y para disminuir el peso de las especies que tienen baja presencia. Esto equivale a realizar un ACP de la tripleta  $(\mathbf{W}, \mathbf{S}^{-1}, \mathbf{D}_q)$ . Realizando la descomposición en valores singulares:

$$\mathbf{D}_q^{1/2} \mathbf{W} \mathbf{S}^{-1/2} = \mathbf{R} \mathbf{\Lambda}^{1/2} \mathbf{T}^\top \quad (3.2)$$

donde  $\mathbf{R}$  y  $\mathbf{T}$  son matrices cuyas columnas contienen los vectores singulares izquierdos y derechos, respectivamente de la matriz  $\mathbf{D}_q^{1/2} \mathbf{W} \mathbf{S}^{-1/2}$ , tal que,

$$\mathbf{R}^\top \mathbf{R} = \mathbf{I}_\nu \quad (3.3)$$

y

$$\mathbf{T}^\top \mathbf{T} = \mathbf{I}_\nu, \quad (3.4)$$

siendo  $\nu$  el número máximo de ejes retenidos y  $\mathbf{\Lambda}^{1/2}$  una matriz diagonal que contiene los valores singulares asociados a dichos vectores singulares en orden decreciente.

Si se expresa la matriz  $\mathbf{W} = \mathbf{U}\mathbf{B}^{o\top}$ , se obtienen las coordenadas principales para las especies

$$\mathbf{U} = \mathbf{D}_q^{-1/2}\mathbf{R}\mathbf{\Lambda}^{1/2}$$

y las coordenadas estándar para las variables ambientales

$$\mathbf{B}^o = \mathbf{S}^{1/2}\mathbf{T}.$$

Por otro lado,  $\mathbf{C} = \mathbf{S}^{-1/2}\mathbf{T}$  contiene los pesos canónicos que permiten calcular las coordenadas para los lugares:

$$\mathbf{X}^o = \mathbf{Z}\mathbf{C}.$$

La bondad del ajuste en esta técnica se calcula como el porcentaje de la variabilidad total contenida en la matriz de promedios ponderados  $\mathbf{W}$  que es explicada por los ejes retenidos en la solución. Si se supone que  $\nu$  es el número de ejes retenidos y  $\mathbf{Y} = \tilde{\mathbf{R}}\tilde{\mathbf{\Lambda}}^{1/2}\tilde{\mathbf{T}}^\top$  la descomposición en valores singulares en el espacio sin restringir, se tiene:

- Proporción de inercia explicada (Espacio original)

$$IE^{OS} = \frac{\sum_{\alpha=1}^{\nu} \lambda_{u\alpha}}{\sum_{\alpha=1}^{\min(n,q)} \lambda_{u\alpha}}$$

- Proporción de inercia explicada (Espacio proyectado)

$$IE^{PS} = \frac{\sum_{\alpha=1}^{\nu} \lambda_{\alpha}}{\sum_{\alpha=1}^{\min(q,p)} \lambda_{\alpha}}$$

### 3.2.1. Análisis de las Contribuciones de los elementos y los factores en el CCA

Aunque en la práctica común no se suelen comentar las contribuciones relativas de elementos y factores, se ha introducido una breve sección donde se explican la manera de calcularlas en este contexto.

Al igual que se explicó para los métodos Biplot, se pueden calcular las contribuciones relativas de los elementos a los factores y de los factores a los elementos. En el contexto del CCA es posible tener dichas contribuciones tanto para lugares como para especies en el espacio original y en el espacio proyectado.

Las contribuciones relativas del  $m$ -ésimo elemento en el  $\alpha$ -ésimo factor para ambos espacios se calculan como:

- Espacio original:

- Para lugares:

$$CRE_i F_{\alpha} = \frac{f_i x_{ui\alpha}^2}{\lambda_{u\alpha}} \quad i = 1, \dots, n \text{ and } \alpha = 1, \dots, \nu, \quad (3.5)$$

siendo  $x_{ui\alpha}$  la coordenada principal del lugar  $i$ -ésimo en el  $\alpha$ -ésimo eje en el espacio sin restringir.

- Para especies:

$$CRE_k F_{\alpha} = \frac{u_{uk\alpha}^2}{\lambda_{u\alpha}} \quad k = 1, \dots, q \text{ and } \alpha = 1, \dots, \nu, \quad (3.6)$$



siendo  $u_{uk\alpha}$  la coordenada principal de la especie  $k$ -ésima en el  $\alpha$ -ésimo eje en el espacio sin restringir.

■ Espacio proyectado:

• Para lugares:

$$CRE_i F_\alpha = \frac{f_i x_{i\alpha}^2}{\lambda_\alpha} \quad i = 1, \dots, n \text{ and } \alpha = 1, \dots, v, \quad (3.7)$$

siendo  $x_{i\alpha}$  la coordenada principal del lugar  $i$ -ésimo en el  $\alpha$ -ésimo eje.

• Para especies:

$$CRE_k F_\alpha = \frac{u_{k\alpha}^2}{\lambda_\alpha} \quad k = 1, \dots, q \text{ and } \alpha = 1, \dots, v, \quad (3.8)$$

siendo  $u_{k\alpha}$  la coordenada principal de la especie  $k$ -ésima en el  $\alpha$ -ésimo eje.

También es posible calcular las contribuciones relativas del  $\alpha$ -ésimo factor al  $m$ -ésimo elemento para ambos espacios:

■ Espacio original:

• Para lugares:

$$CRF_\alpha E_i = \frac{f_i x_{ui\alpha}^2}{\sum_{r=1}^n f_r x_{ur\alpha}^2} \quad i = 1, \dots, n \text{ and } \alpha = 1, \dots, \nu, \quad (3.9)$$

siendo  $x_{ui\alpha}$  la coordenada principal del lugar  $i$ -ésimo en el  $\alpha$ -ésimo eje en el espacio sin restringir.

- Para especies:

$$CRF_{\alpha}E_k = \frac{u_{uk\alpha}^2}{\sum_{r=1}^q u_{ur\alpha}^2} \quad k = 1, \dots, q \text{ and } \alpha = 1, \dots, \nu, \quad (3.10)$$

siendo  $u_{uk\alpha}$  la coordenada principal de la especie  $k$ -ésima en el  $\alpha$ -ésimo eje en el espacio sin restringir.

- Espacio proyectado:

- Para lugares:

$$CRF_{\alpha}E_i = \frac{f_i x_{i\alpha}^2}{\sum_{r=1}^n f_r x_{r\alpha}^2} \quad i = 1, \dots, n \text{ and } \alpha = 1, \dots, \nu, \quad (3.11)$$

siendo  $x_{i\alpha}$  la coordenada principal del lugar  $i$ -ésimo en el  $\alpha$ -ésimo eje.

- Para especies:

$$CRF_{\alpha}E_k = \frac{u_{k\alpha}^2}{\sum_{r=1}^q u_{r\alpha}^2} \quad k = 1, \dots, q \text{ and } \alpha = 1, \dots, \nu, \quad (3.12)$$

siendo  $u_{k\alpha}$  la coordenada principal de la especie  $k$ -ésima en el  $\alpha$ -ésimo eje.

### 3.2.2. Análisis Canónico de Correspondencias para variables cualitativas

A continuación se presenta teóricamente el CCA para el caso de tener variables ambientales de carácter cualitativo.

En presencia de variables ambientales de carácter mixto (cuantitativas y cualitativas), la matriz  $\mathbf{Z}$  cambia su composición. Se dispone de  $p$  variables, con  $p_Q$  cuantitativas y  $p_A$  cualitativas. Para las cuantitativas, cada columna de

$\mathbf{Z}$  representa una variable. Sin embargo, cada variable cualitativa está definida por una submatriz de  $\mathbf{Z}$ , que se denomina  $\mathbf{Z}_j$ , formada por los  $m_j$  indicadores asociados a las  $m_j$  modalidades de dicha variable  $j$ ,  $j = 1, \dots, p_A$ .

Se define, por tanto, la matriz combinada  $\mathbf{Z}_{mix}$ , compuesta por la submatriz correspondiente a las  $p_Q$  variables cuantitativas  $\mathbf{D}_n$ -estandarizadas ( $\mathbf{Z}_Q$ ) y por los indicadores  $\mathbf{D}_n$ -centrados asociados a cada una de las modalidades del conjunto de variables cualitativas ( $\mathbf{Z}_A$ ):

$$\mathbf{Z}_{mix_{n \times h}} = \left[ \mathbf{Z}_{Q_{n \times p_Q}} \parallel \mathbf{Z}_{A_{n \times m}} \right],$$

siendo  $m = \sum_{j=1}^{p_A} m_j$  el número total de modalidades de las  $p_A$  variables cualitativas, con  $h = p_Q + m$ . Su rango  $\nu$  sería como mucho, el mínimo de  $n - 1$  y  $p_Q + m - p_A$ .

El hecho de elegir los indicadores  $\mathbf{D}_n$ -centrados tiene como objetivo encontrar estimaciones de los marcadores de las modalidades de forma que sean los centroides de los lugares en donde dicha modalidad esté presente.

Para que la matriz de productos cruzados de  $\mathbf{Z}_{mix}$  sea invertible, es necesario trabajar con su inversa generalizada. Realizando la descomposición en valores singulares de la matriz de productos cruzados  $\mathbf{Z}_{mix}^\top \mathbf{D}_n \mathbf{Z}_{mix}$  y operando sobre la matriz de valores singulares, se obtiene la inversa generalizada de Penrose.

Es decir, si:

$$\mathbf{Z}_{mix}^\top \mathbf{D}_n \mathbf{Z}_{mix} = \mathbf{\Gamma} \mathbf{\Lambda} \mathbf{\Gamma}^\top,$$

tal que  $\mathbf{\Gamma}^\top \mathbf{\Gamma} = \mathbf{I}_\nu$ , con  $\mathbf{\Lambda}$  matriz diagonal conteniendo los autovalores en orden decreciente, siendo  $\nu$  el rango de la matriz a descomponer. Y sea  $\mathbf{\Lambda}^-$  la matriz diagonal con los elementos no nulos iguales a las inversas de los respectivos elementos no nulos de  $\mathbf{\Lambda}$ . Por tanto, la inversa generalizada de Penrose de la

matriz de productos cruzados es:

$$(\mathbf{Z}_{mix}^\top \mathbf{D}_n \mathbf{Z}_{mix})^- = \mathbf{\Gamma} \mathbf{\Lambda}^- \mathbf{\Gamma}^\top \quad (3.13)$$

A partir de 3.13 se obtiene el valor de:

$$\left[ (\mathbf{Z}_{mix}^\top \mathbf{D}_n \mathbf{Z}_{mix})^- \right]^{1/2} = \mathbf{\Gamma} [\mathbf{\Lambda}^-]^{1/2}$$

necesario para el cálculo de  $\mathbf{W}$ .

Una vez definida la inversa generalizada de la matriz de productos cruzados, el CCA se basa en la descomposición de

$$(\mathbf{Z}_{mix}^\top \mathbf{D}_n \mathbf{Z}_{mix})^- \mathbf{W}^\top \mathbf{D}_q \mathbf{W}.$$

Para obtener los marcadores de las variables ambientales se debe premultiplicar la descomposición espectral de  $\mathbf{W}$  por una matriz diagonal  $\mathbf{N}^{-1}$  formada por dos bloques:

$$\mathbf{N}^{-1} = \begin{bmatrix} \mathbf{I}_Q & \mathbf{0} \\ \mathbf{0} & \mathbf{N}_A^{-1} \end{bmatrix}_{h \times h},$$

tal que,

$$\mathbf{N}_A^{-1}{}_{m \times m} = [\text{diag}(\mathbf{Z}_A^\top \mathbf{D}_{nn \times n} \mathbf{1}_{n \times 1})]^{-1}.$$

con  $\mathbf{Z}_A$  matriz que contiene los indicadores sin centrar.

Las expresiones para los marcadores de las variables ambientales en coordenadas estándar se calculan:

$$\mathbf{B}^o = \mathbf{N}^{-1}(\mathbf{Z}_{mix}^T \mathbf{D}_n \mathbf{Z}_{mix})^{-1/2} \mathbf{T}.$$

### 3.2.3. Interpretación

Las coordenadas resultantes del CCA se pueden representar en un mismo gráfico. Los lugares, las especies y las variables ambientales de carácter cualitativo están representadas por puntos y las variables ambientales cuantitativas por vectores.

Las coordenadas de los lugares son los valores de los lugares en los gradientes obtenidos correspondientes a los ejes que se estén representando. Cada punto especie en el gráfico está situado en el centroide (promedio ponderado) de los puntos lugares en los que está presente dicha especie, es decir, el centroide de su nicho. Dado que es un promedio ponderado, los lugares próximos al punto especie tienden a presentar mayores abundancias de la especie en cuestión que aquellos posicionados lejos de dicho punto.

La distancia Chi-cuadrado es una medida de la disimilaridad entre el perfil de abundancia de una especie y el de otra. Así, diferencias en la abundancia total entre especies no incrementan necesariamente su disimilaridad medida ésta con la distancia Chi-cuadrado. Por lo tanto, especies que están próximas en el gráfico se espera que sean similares en cuanto a su distribución a lo largo de los lugares mientras que especies distantes tienden a ser más diferentes en ese aspecto. Hay que tener en cuenta que, puntos especies situados cerca pueden presentar diferencias considerables si la bondad del ajuste es mala, ya que pueden estar situados a más distancia en otros ejes diferentes a los presentados en el gráfico.

La interpretación para las distancias entre los puntos lugares es la misma que para el caso de los puntos especies.

Las variables ambientales cuantitativas se muestran según sus correlaciones con los ejes. Las puntas de las flechas muestran la dirección de máximo cambio en los valores asociados a las variables y la longitud es proporcional a la máxima tasa de cambio. En la dirección perpendicular la variable no cambia su valor. También es posible estudiar las correlaciones entre las variables ya que pueden ser estimadas a partir de la proyección de las puntas de las flechas sobre las otras variables. El orden de dichas proyecciones sobre una determinada variable ofrece un orden en cuanto a la correlación de dicha variable con el resto.

Los puntos especies son los promedios ponderados de los lugares donde están presentes no sólo en el gráfico sino también cuando se proyectan sobre una determinada variable ambiental. Esto es debido a que las proyecciones de los puntos que representan a lugares y especies sobre una determinada variable aproximan los valores de los lugares y los promedios ponderados de las especies para dicha variable.

Las variables de carácter cualitativo están representadas por los centroides de sus categorías. Una variable cualitativa está formada por un número de categorías que conforman una partición de los lugares. Cada categoría se representa como un punto en el gráfico que es el centroide de los puntos que pertenecen a dicha categoría. (El centroide es el promedio ponderado, siendo la ponderación la abundancia total de un lugar).

### 3.3. Análisis Canónico No Simétrico de Correspondencias

Como se explicó en la introducción, cuando el objetivo es evaluar una tabla de contingencia en la que ambas variables no juegan el mismo papel, es decir, cuando existe una estructura de dependencia, Gimaret-Carpentier et al., 1998; Kroonenberg y Lombardo, 1999 proponen una alternativa al CA denominada NSCA (análisis de Correspondencias no Simétrico). En este sentido, cuando se dispone de más de dos variables categóricas, (D'Ambra y Lauro, 1989, 1992; Lauro y D'Ambra, 1984) proponen el análisis Múltiple de Correspondencias no Simétrico. También existe una versión no simétrica del COIA (Esposito, 1997). Cuando además de la estructura de dependencia se dispone de una matriz de variables externa, Willems y Galindo, 2008 proponen el CNCA (análisis Canónico de Correspondencias no Simétrico). A continuación se presentan los aspectos teóricos más relevantes de esta técnica. El objetivo es estudiar la relación entre los lugares muestreados y las especies estudiadas teniendo en cuenta la información externa disponible relativa a las variables pero considerando las especies dependientes de los lugares.

Sea la información de  $q$  especies y  $p$  variables ambientales cuantitativas medidas sobre  $n$  lugares, expresada en las siguientes matrices:

- matriz  $\mathbf{Y}$ , de orden  $n \times q$ , cuyo elemento genérico  $y_{ik}$ ,  $i = 1, \dots, n$ ,  $k = 1, \dots, q$ , contiene información de la  $k$ -ésima especie en el  $i$ -ésimo lugar, con matriz de correspondencias asociada  $\mathbf{F} = y^{-1}\mathbf{Y}$ , siendo  $y$  la suma total de la matriz  $\mathbf{Y}$ ;
- matriz  $\mathbf{Z}$ , de orden  $n \times p$ , cuyo elemento genérico  $z_{ij}$ ,  $i = 1, \dots, n$ ,  $j = 1, \dots, p$ , contiene el valor de la  $j$ -ésima variable ambiental cuantitativa en el  $i$ -ésimo

lugar.

Sea  $\mathbf{D}_n = \text{diag}(f_{1.}, f_{2.}, \dots, f_{n.})$  la matriz cuya diagonal es el vector  $\mathbf{f}_n = \mathbf{F}\mathbf{1}_q$ , que son los marginales fila (lugares) de la matriz  $\mathbf{F}$ ; con  $\mathbf{1}_q$  vector de  $q$  unos.  $\mathbf{1}_n$  es un vector de  $n$  unos;  $\mathbf{f}_q = \mathbf{F}^\top \mathbf{1}_n$ , es el vector de marginales de las  $q$  columnas (especies)  $f_{.k}$ ,  $k = 1, \dots, q$ .

A partir de las definiciones dadas anteriormente, se define la matriz the perfiles-fila como:

$$\mathbf{P} = \mathbf{D}_n^{-1} \mathbf{F} \quad (3.14)$$

cuya expresión centrada es:

$$\tilde{\mathbf{P}} = \mathbf{D}_n^{-1} (\mathbf{F} - \mathbf{f}_n \mathbf{f}_q^\top) \quad (3.15)$$

Para cada especie, la matriz centrada  $\tilde{\mathbf{P}}$  muestra la magnitud de la diferencia entre su participación en ese perfil-lugar y el perfil media, indicando una mayor o menor abundancia relativa al lugar dado.

Proyectando  $\tilde{\mathbf{P}}$  a través de la matriz

$$\mathbf{\Pi} = \mathbf{Z}(\mathbf{Z}^\top \mathbf{D}_n \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{D}_n \quad (3.16)$$

se obtiene:

$$\tilde{\mathbf{P}}^* = \mathbf{\Pi} \tilde{\mathbf{P}} \quad (3.17)$$

Utilizando el concepto de tripletas de Escoufier, 1987, el CNCA se basa en el análisis de la tripleta  $(\tilde{\mathbf{P}}^*, \mathbf{I}_q, \mathbf{D}_n)$ , lo que implica la siguiente descomposición en valores singulares:



$$\tilde{\mathbf{P}}^* = \mathbf{R}\mathbf{\Lambda}^{1/2}\mathbf{T}^\top \quad (3.18)$$

tal que,

$$\mathbf{R}^\top \mathbf{D}_n \mathbf{R} = \mathbf{I}_\nu \quad (3.19)$$

y

$$\mathbf{T}^\top \mathbf{T} = \mathbf{I}_\nu, \quad (3.20)$$

donde  $\mathbf{R}$  y  $\mathbf{T}$  son matrices que contienen en sus columnas los vectores singulares izquierdos y derechos, respectivamente, de la matriz  $\tilde{\mathbf{P}}^*$ , siendo  $\nu$  el número máximo de ejes retenidos (rango de la matriz  $\mathbf{Z}$ ), con  $\mathbf{\Lambda}^{1/2}$  matriz diagonal cuyos términos son los respectivos valores singulares, en orden decreciente ( $\lambda_1^{1/2} \geq \lambda_2^{1/2} \geq \dots \geq \lambda_\nu^{1/2}$ ).

De 3.18 se obtiene la expresión para las puntuaciones de lugares y especies.

Si se escribe  $\tilde{\mathbf{P}}^* = \mathbf{X}\mathbf{U}^\top$  tal que,

$$\mathbf{X} = \mathbf{R}\mathbf{\Lambda}^{1/2} = \tilde{\mathbf{P}}^* \mathbf{T}, \quad (3.21)$$

$$\mathbf{U}^\circ = \mathbf{T}. \quad (3.22)$$

La matriz  $\mathbf{X}$  contiene en su  $\alpha$ -ésima columna las coordenadas principales para los lugares correspondientes al eje factorial  $\alpha$ -ésimo, con matriz de covarianzas  $\mathbf{\Lambda}$ , siendo sus coordenadas estándar  $\mathbf{X}^\circ = \mathbf{R}$ .

En las columnas de  $\mathbf{U}^\circ$  se encuentran las coordenadas estándar para las especies, con varianza unitaria, siendo sus coordenadas principales  $\mathbf{U} = \mathbf{T}\mathbf{\Lambda}^{1/2}$ .

Para calcular las coordenadas para las variables ambientales, se define la intertabla  $\mathbf{L} = \mathbf{F}_I^\top \mathbf{Z}$  donde  $\mathbf{F}_I = \mathbf{F} - \mathbf{f}_n \mathbf{f}_q^\top$ .

Para conseguir que las estimaciones de las puntuaciones sean invariantes por

transformaciones lineales no singulares de las variables ambientales, la matriz  $\mathbf{L}$  se pondera por la inversa de la matriz de covarianzas de  $\mathbf{Z}$ . Entonces, el CNCA se basa en la descomposición en valores singulares de la tripleta  $\mathbf{L}_p = (\mathbf{Z}^\top \mathbf{D}_n \mathbf{Z})^{-1/2} \mathbf{L}^\top$ :

$$\mathbf{L}_p = \mathbf{K} \mathbf{\Lambda}^{1/2} \mathbf{T}^\top \quad (3.23)$$

tal que,

$$\mathbf{K}^\top \mathbf{K} = \mathbf{I}_\nu \quad (3.24)$$

y

$$\mathbf{T}^\top \mathbf{T} = \mathbf{I}_\nu, \quad (3.25)$$

donde la matriz  $\mathbf{K}$ , de orden  $p \times \nu$ , y la matriz  $\mathbf{T}$ , de orden  $q \times \nu$ , contienen los vectores singulares izquierdos y derechos, respectivamente, de la matriz  $\mathbf{L}_p$ , con  $\mathbf{\Lambda}^{1/2}$  matriz diagonal de orden  $\nu$ , cuyos elementos son los respectivos valores singulares, en orden decreciente ( $\lambda_1^{1/2} \geq \lambda_2^{1/2} \geq \dots \geq \lambda_\nu^{1/2}$ ). Las dos últimas matrices ( $\mathbf{T}$  y  $\mathbf{\Lambda}^{1/2}$ ) son equivalentes a las obtenidas en 3.18.

Si escribimos  $\mathbf{L}^\top$  como:

$$\mathbf{L}^\top = (\mathbf{Z}^\top \mathbf{D}_n \mathbf{Z})^{1/2} \mathbf{K} \mathbf{\Lambda}^{1/2} \mathbf{T}^\top, \quad (3.26)$$

se obtienen las puntuaciones para las variables ambientales en coordenadas principales:

$$\mathbf{B} = (\mathbf{Z}^\top \mathbf{D}_n \mathbf{Z})^{1/2} \mathbf{K} \mathbf{\Lambda}^{1/2}, \quad (3.27)$$

siendo las coordenadas estándar,

$$\mathbf{B}^o = (\mathbf{Z}^\top \mathbf{D}_n \mathbf{Z})^{1/2} \mathbf{K}. \quad (3.28)$$

El cálculo de las matrices  $\tilde{\mathbf{P}}^*$  y  $\mathbf{L}$  se muestra en la figura 3.2.

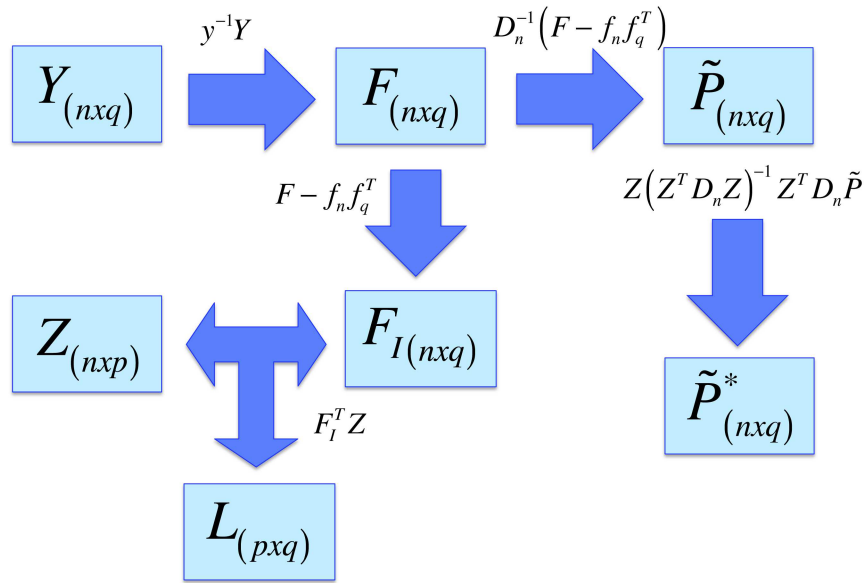


Figura 3.2: Esquema de la obtención de  $\tilde{\mathbf{P}}^*$  y  $\mathbf{L}$ .

### 3.3.1. Análisis de las Contribuciones de los elementos y los factores en el CNCA

En el contexto del CCA se explicó la manera de calcular contribuciones relativas de los elementos y de los factores. En el caso del CNCA también se pueden calcular para lugares y especies en los espacios original y proyectado.

Las expresiones para las contribuciones del  $m$ -ésimo elemento en el  $\alpha$ -ésimo factor  $(CRE_m F_\alpha)$ , vienen dadas por:

- Para lugares:

$$CRE_i F_\alpha = \frac{f_i \cdot x_{i\alpha}^2}{\lambda_\alpha} \quad i = 1, \dots, n \text{ and } \alpha = 1, \dots, v, \quad (3.29)$$

siendo  $x_{i\alpha}$  la coordenada principal del lugar  $i$ -ésimo en el  $\alpha$ -ésimo eje.

- Para especies:

$$CRE_k F_\alpha = \frac{u_{k\alpha}^2}{\lambda_\alpha} \quad k = 1, \dots, q \text{ and } \alpha = 1, \dots, v, \quad (3.30)$$

siendo  $u_{k\alpha}$  la coordenada principal de la especie  $k$ -ésima en el  $\alpha$ -ésimo eje.

La tabla 3.2 presenta las medidas de calidad para ejes factoriales y elementos. Para su cálculo se parte de las distancias al cuadrado (tabla 3.1) de especies y lugares respecto a dos diferentes espacios:

- El espacio proyectado, es decir, la inercia total de los valores ajustados (indicado con superíndice  $PS$ ).
- El espacio original, es decir, la inercia total de los valores observados (indicado con superíndice  $OS$ ).

Los indicadores denominados  $Q_{\alpha,m}$  son una medida de la calidad de representación del  $m$ -ésimo elemento en el eje  $\alpha$ -ésimo. Toman valores entre cero y uno (uno indica representación exacta). El estadístico  $Q_{\alpha,k}^{(OS)}$  indica como de bien dichas variables ambientales explican cada una de las especies. Para especies, la varianza de los valores ajustados es menor, o igual, a la varianza de los valores originales (debido a las propiedades de la regresión). Este no es el caso de los lugares, el cuadrado de la distancia entre el respectivo centroide y los valores proyectados pueden ser mayores que la propia distancia en el espacio original.

Así, el estadístico para los lugares relativo al espacio original se denomina  $D_{\alpha,i}$ , para diferenciarlo de aquellas que son medidas de calidad, ya que no es un coseno cuadrado.

	Especies	Lugares
Espacio Original	$d_{O,k}^2 = \sum_{i=1}^n \frac{1}{f_i} (f_{ik} - f_i \cdot f_{.j})^2$	$d_{O,i}^2 = \sum_{k=1}^q \left( \frac{f_{ik}}{f_i} - f_{.k} \right)^2$
Espacio de Proyección	$d_{Pr,k}^2 = \sum_{i=1}^n f_i \cdot (\tilde{p}_{ik}^*)^2$	$d_{Pr,i}^2 = \sum_{k=1}^q (\tilde{p}_{ik}^*)^2$

Tabla 3.1: Distancias al cuadrado (especies y lugares), con respecto al espacio proyectado ( $PS$ ) y al espacio original ( $OS$ ).

Relativo a ejes factoriales	Respecto del espacio proyectado (superíndice PS)	Respecto del espacio original (superíndice OS)
	$I_{\alpha}^{(PS)} = \frac{\lambda_{\alpha}^2}{\sum_{\alpha=1}^v \lambda_{\alpha}^2}$	$I_{\alpha}^{(OS)} = \frac{\lambda_{\alpha}^2}{TI(NSCA)}$
	Proporción de inercia explicada por el eje $\alpha$ -ésimo, con respecto a la inercia total de los datos proyectados.	Proporción de inercia explicada por el eje $\alpha$ -ésimo, con respecto a la inercia total de los datos originales.
Relativo a Lugares elementos	$Q_{\alpha,i}^{(PS)} = \frac{x_{i\alpha}^2}{d_{Pr,i}^2}$	$D_{\alpha,i} = \frac{x_{i\alpha}^2}{d_{O,i}^2}$
Especies	$Q_{\alpha,k}^{(PS)} = \frac{u_{k\alpha}^2}{d_{Pr,k}^2}$	$Q_{\alpha,k}^{(OS)} = \frac{u_{k\alpha}^2}{d_{O,k}^2}$

$D_{\alpha,i}$  puede ser mayor que uno.

Tabla 3.2: Calidades de representación para las puntuaciones de ejes factoriales y elementos (especies y lugares), con respecto al espacio proyectado ( $PS$ ) y al espacio original ( $OS$ ).  $\alpha = 1, \dots, v$ .

### 3.3.2. Análisis Canónico No Simétrico de Correspondencias para variables cualitativas

A continuación, se detalla el desarrollo teórico del CNCA para el caso de tener un conjunto de variables ambientales que no sean cuantitativas.

En presencia de variables ambientales de carácter mixto (cuantitativas y cualitativas), la matriz  $\mathbf{Z}$  cambia su composición. Se dispone de  $p$  variables, con  $p_Q$  cuantitativas y  $p_A$  cualitativas. Para las cuantitativas, cada columna de  $\mathbf{Z}$  representa una variable. Sin embargo, cada variable cualitativa está definida por una submatriz de  $\mathbf{Z}$ , que se denomina  $\mathbf{Z}_j$ , formada por los  $m_j$  indicadores asociados a las  $m_j$  modalidades de dicha variable  $j$ ,  $j = 1, \dots, p_A$ . Luego esta matriz  $\mathbf{Z}$  no es de rango completo, lo cual es necesario para que su matriz de productos cruzados sea invertible y se pueda calcular la matriz  $\mathbf{L}_p$ .

Se define, por tanto, la matriz combinada  $\mathbf{Z}_{mix}$ , compuesta por la submatriz correspondiente a las  $p_Q$  variables cuantitativas  $\mathbf{D}_n$ -estandarizadas ( $\mathbf{Z}_Q$ ) y por los indicadores  $\mathbf{D}_n$ -centrados asociados a cada una de las modalidades del conjunto de variables cualitativas ( $\mathbf{Z}_A$ ):

$$\mathbf{Z}_{mix_{n \times h}} = \left[ \mathbf{Z}_{Q_{n \times p_Q}} \parallel \mathbf{Z}_{A_{n \times m}} \right],$$

siendo  $m = \sum_{j=1}^{p_A} m_j$  el número total de modalidades de las  $p_A$  variables cualitativas, con  $h = p_Q + m$ . Su rango  $\nu$  sería como mucho, el mínimo de  $n - 1$  y  $p_Q + m - p_A$ .

El hecho de elegir los indicadores  $\mathbf{D}_n$ -centrados tiene como objetivo encontrara estimaciones de los marcadores de las modalidades de forma que sean los centroides de los lugares en donde dicha modalidad esté presente.

Los elementos  $l_{kj}$  de la matriz  $\mathbf{L}$ ,  $k = 1, \dots, q$ ,  $j = p_Q + 1, \dots, h$ , asociados a las modalidades de las variables cualitativas, contienen la diferencia entre los

totales de los elementos positivos y negativos de cada columna (especie) de  $\mathbf{F}_I$  considerando sólo los lugares donde la modalidad esté presente.

Para que la matriz de productos cruzados de  $\mathbf{Z}_{mix}$  sea invertible, es necesario trabajar con su inversa generalizada. Realizando la descomposición en valores singulares de la matriz de productos cruzados  $\mathbf{Z}_{mix}^\top \mathbf{D}_n \mathbf{Z}_{mix}$  y operando sobre la matriz de valores singulares, se obtiene la inversa generalizada de Penrose.

Es decir, si:

$$\mathbf{Z}_{mix}^\top \mathbf{D}_n \mathbf{Z}_{mix} = \mathbf{\Gamma} \mathbf{\Lambda} \mathbf{\Gamma}^\top,$$

tal que  $\mathbf{\Gamma}^\top \mathbf{\Gamma} = \mathbf{I}_\nu$ , con  $\mathbf{\Lambda}$  matriz diagonal conteniendo los autovalores en orden decreciente, siendo  $\nu$  el rango de la matriz a descomponer. Y sea  $\mathbf{\Lambda}^-$  la matriz diagonal con los elementos no nulos iguales a las inversas de los respectivos elementos no nulos de  $\mathbf{\Lambda}$ . Por tanto, la inversa generalizada de Penrose de la matriz de productos cruzados es:

$$\left(\mathbf{Z}_{mix}^\top \mathbf{D}_n \mathbf{Z}_{mix}\right)^- = \mathbf{\Gamma} \mathbf{\Lambda}^- \mathbf{\Gamma}^\top \quad (3.31)$$

A partir de 3.31 se obtiene el valor de:

$$\left[\left(\mathbf{Z}_{mix}^\top \mathbf{D}_n \mathbf{Z}_{mix}\right)^-\right]^{1/2} = \mathbf{\Gamma} \left[\mathbf{\Lambda}^-\right]^{1/2}$$

necesario para el cálculo de  $\mathbf{L}_p$ .

Una vez definida la inversa generalizada de la matriz de productos cruzados y sustituyendo en 3.16  $\left(\mathbf{Z}_{mix}^\top \mathbf{D}_n \mathbf{Z}_{mix}\right)^-$  por su equivalente definido en 3.31, y  $\mathbf{Z}$  por  $\mathbf{Z}_{mix}$ , se obtienen los marcadores de lugares y especies.

Para obtener los marcadores de las variables ambientales se debe premultiplicar la descomposición espectral de  $\mathbf{L}_p = \mathbf{K} \mathbf{\Lambda}^{1/2} \mathbf{T}^\top$  por una matriz diagonal  $\mathbf{N}^{-1}$

formada por dos bloques:

$$\mathbf{N}^{-1} = \begin{bmatrix} \mathbf{I}_Q & \mathbf{0} \\ \mathbf{0} & \mathbf{N}_A^{-1} \end{bmatrix}_{h \times h},$$

tal que,

$$\mathbf{N}_A^{-1}{}_{m \times m} = [\text{diag}(\mathbf{Z}_A^\top \mathbf{D}_{nn \times n} \mathbf{1}_{n \times 1})]^{-1},$$

con  $\mathbf{Z}_A$  matriz que contiene los indicadores sin centrar.

Las expresiones para los marcadores de las variables ambientales se calculan:

$$\mathbf{B} = \mathbf{N}^{-1}(\mathbf{Z}_{mix}^\top \mathbf{D}_n \mathbf{Z}_{mix})^{1/2} \mathbf{K} \mathbf{\Lambda}^{1/2},$$

siendo las coordenadas estándar,

$$\mathbf{B}^o = \mathbf{N}^{-1}(\mathbf{Z}_{mix}^\top \mathbf{D}_n \mathbf{Z}_{mix})^{1/2} \mathbf{K}.$$

### 3.3.3. Interpretación

La interpretación conjunta de especies-lugares y especies-variables ambientales se realiza a través de las proyecciones de los marcadores de una de las nubes de puntos sobre la que se pretende relacionar. Esto se debe a que los conjuntos de marcadores forman dos representaciones biplot superpuestas. Como se trabaja sobre perfiles-lugares, la representación biplot que se ajusta al objetivo de analizar la distribución de los lugares según su población de especies, es la que representa a los lugares y a las variables ambientales en coordenadas principales y a las especies en coordenadas estándar.

El producto interno de los marcadores de especies y lugares permite recuperar



los valores de los perfiles-lugares centrados ( $\tilde{p}_{ik}^*$ ). Este valor es solamente una aproximación si sólo se retienen los primeros ejes. La proyección de los marcadores de especies sobre la recta que une el origen con cada uno de los marcadores lugares, es proporcional a dicha aproximación.

Por lo tanto, las especies proyectadas en el extremo positivo del eje correspondiente a un determinado lugar, son especies con mayores incrementos en probabilidad de tener en dicho lugar un valor de composición relativa centrada ajustada de importancia. Por el contrario, si esta proyección se sitúa en el extremo negativo, la probabilidad de tener en ese lugar tales valores es escasa o nula.

Dado que las relaciones entre ambos conjuntos de puntos se obtiene a través de productos escalares, una especie puede estar situada lejos de un lugar en términos de distancia Euclídea pero, debido a su proyección, ser de importancia relativa en ese lugar. Lugares próximos entre sí, bien representados, indican que su composición, en cuanto a especies se refiere y ajustadas según las variables ambientales, es similar.

Por otro lado, el producto interno entre los marcadores de especies y variables ambientales permiten recuperar los valores de la matriz  $\mathbf{L}$ . Si se retienen sólo los primeros ejes, esta afirmación es aproximada, siendo este valor proporcional a las proyecciones de los marcadores de especies (en coordenadas estándar) sobre la recta determinada por el vector que representa una variable ambiental (en coordenadas principales). Por tanto, para cada especie se obtiene una cuantificación de la magnitud de las diferencias que existen en la variable ambiental de referencia, entre los lugares con coberturas relativas  $f_{ik}$  superiores a la esperada si la distribución de tal especie fuera al azar, de aquellas inferiores a tal valor.

En este escalamiento condicional-lugar, los marcadores de lugares y variables ambientales no conforman un biplot. Aún así, la orientación de cada vector que

representa a una variable ambiental permite identificar las zonas (lugares) hacia donde crecen los valores de dicha variable.

### 3.4. Análisis de Coinercia

Tanto el CCA como el CNCA tienen la desventaja de que necesitan que el número de lugares de muestreo sea mayor que el número de variables analizadas. Como alternativa a ambos métodos se presenta el COIA (Dolédec y Chessel, 1994). Esta técnica no tiene limitaciones en cuanto al número de variables y lugares de muestreo.

Diferentes estadísticos como la covarianza o el coeficiente de correlación de Pearson miden la relación entre dos variables. El COIA (Dolédec y Chessel, 1994) se basa en la definición de un estadístico para medir dos o más grupos de variables.

Sea la información de  $q$  especies y  $p$  variables ambientales cuantitativas medidas sobre  $n$  lugares, expresada en las siguientes matrices:

- matriz  $\mathbf{Y}$ , de orden  $n \times q$ , cuyo elemento genérico  $y_{ik}$ ,  $i = 1, \dots, n$ ,  $k = 1, \dots, q$ , contiene información de la  $k$ -ésima especie en el  $i$ -ésimo lugar. Cada lugar puede ser representado como un punto en el espacio  $q$ -dimensional de las especies;
- matriz  $\mathbf{Z}$ , de orden  $n \times p$ , cuyo elemento genérico  $z_{ij}$ ,  $i = 1, \dots, n$ ,  $j = 1, \dots, p$ , contiene el valor de la  $j$ -ésima variable ambiental cuantitativa en el  $i$ -ésimo lugar. Cada lugar puede ser representado como un punto en el espacio de  $p$  dimensiones donde los ejes representan a cada una de las  $p$  variables ambientales.

Sea  $\mathbf{D}_n = \text{diag}(f_{1.}, f_{2.}, \dots, f_{n.})$  la matriz cuya diagonal contiene los pesos de los lugares (se suponen iguales para ambas matrices),  $\mathbf{D}_q$  de orden  $q \times q$  una métrica

en el espacio de las especies y  $\mathbf{D}_p$  de orden  $p \times p$  una métrica en el espacio de las variables, entonces la inercia de la nube de lugares alrededor del punto de referencia (en el espacio de las variables) o es:

$$I_o = \sum_{i=1}^n f_i \cdot \|\mathbf{Z}_i\|_{D_p}^2 = \text{traza}(\mathbf{Z}\mathbf{D}_p\mathbf{Z}^\top\mathbf{D}_n).$$

Los lugares se pueden proyectar sobre un vector  $D_p$ -normado  $\mathbf{u}$  y la inercia proyectada se puede expresar como:

$$I(u) = \mathbf{u}^\top \mathbf{D}_p \mathbf{Z}^\top \mathbf{D}_n \mathbf{Z} \mathbf{D}_p \mathbf{u}$$

Consecuentemente, la inercia total se puede descomponer sobre un conjunto de vectores  $D_p$ -normados ortogonales  $\mathbf{u}_k$ :

$$I_o = \sum_{k=1}^p I(u_k) = \sum_{k=1}^p \mathbf{u}_k^\top \mathbf{D}_p \mathbf{Z}^\top \mathbf{D}_n \mathbf{Z} \mathbf{D}_p \mathbf{u}_k = \sum_{k=1}^p \|\mathbf{Z} \mathbf{D}_p \mathbf{u}_k\|_{D_n}^2.$$

Si se calcula la inercia respecto del espacio de las especies, se tiene:

$$J_o = \text{traza}(\mathbf{Y} \mathbf{D}_q \mathbf{Y}^\top \mathbf{D}_n)$$

y puede ser descompuesta sobre un conjunto de vectores  $D_q$ -normados ortogonales  $\mathbf{v}_k$ .

La co-inercia es una medida global de la co-estructura de los lugares sobre los espacios de las variables ambientales y las especies: es alto cuando las dos estructuras varían simultáneamente y bajo cuando las dos estructuras varían independientemente o cuando no varían. Su expresión es:

$$Co - I = \sum_{k=1}^p \sum_{j=1^q} (\mathbf{u}_k^\top \mathbf{D}_p \mathbf{Z}^\top \mathbf{D}_n \mathbf{Y} \mathbf{D}_q \mathbf{v}_j)^2 = \text{traza} (\mathbf{Z}^\top \mathbf{D}_p \mathbf{Z}^\top \mathbf{D}_n \mathbf{Y} \mathbf{D}_q \mathbf{Y}^\top \mathbf{D}_n).$$

Si las nubes de puntos están centradas, la inercia es la suma de las varianzas y la co-inercia la suma de los cuadrados de las covarianzas.

Basándose en el concepto de co-inercia, Chessel y Mercier, 1993; Dolédec y Chessel, 1994 desarrollaron el análisis de la Coinercia. Es una técnica que descompone la co-inercia sobre un conjunto de vectores ortogonales. Se define como el estudio de la tripleta  $(\mathbf{Y}^\top \mathbf{D}_n \mathbf{Z}, \mathbf{D}_p, \mathbf{D}_q)$ . El objetivo es encontrar un vector  $\mathbf{v}_1$  en el espacio de las especies y un vector  $\mathbf{u}_1$  en el espacio de las variables que tengan máxima co-inercia. Si las matrices  $\mathbf{Y}$  y  $\mathbf{Z}$  están centradas, entonces el COIA maximiza la covarianza al cuadrado entre la proyección de  $\mathbf{Y}$  sobre  $\mathbf{v}_1$  y la proyección de  $\mathbf{Z}$  sobre  $\mathbf{u}_1$ :

$$\text{cov}^2(\mathbf{Z} \mathbf{D}_p \mathbf{u}_1, \mathbf{Y} \mathbf{D}_q \mathbf{v}_1) = \text{corr}^2(\mathbf{Z} \mathbf{D}_p \mathbf{u}_1, \mathbf{Y} \mathbf{D}_q \mathbf{v}_1) \times \text{var}(\mathbf{Z} \mathbf{D}_p \mathbf{u}_1) \times \text{var}(\mathbf{Y} \mathbf{D}_q \mathbf{v}_1).$$

Para el estudio de la tripleta  $(\mathbf{Y}^\top \mathbf{D}_n \mathbf{Z}, \mathbf{D}_p, \mathbf{D}_q)$  se realiza la siguiente descomposición en valores singulares:

$$\mathbf{D}_p^{(1/2)} \mathbf{Z}^\top \mathbf{D}_n \mathbf{Y} \mathbf{D}_q \mathbf{Y}^\top \mathbf{D}_n \mathbf{Z} \mathbf{D}_p^{(1/2)} \approx \mathbf{K}_r \mathbf{\Lambda}_r \mathbf{K}_r^\top$$

donde  $\mathbf{K}_r$  de orden  $p \times r$  cumple  $\mathbf{K}_r^\top \mathbf{K}_r = \mathbf{I}_r$ .

A partir de esta expresión se obtienen los ejes de co-inercia  $D_p$ -normalizados en el espacio de las variables ambientales:

$$\mathbf{v}_r^Z = \mathbf{D}_p^{-(1/2)} \mathbf{K}_r.$$

Los ejes de co-inercia  $D_q$ -normalizados en el espacio de las especies:

$$\mathbf{v}_r^Y = \mathbf{Y}^\top \mathbf{D}_n \mathbf{Z} \mathbf{D}_p^{(1/2)} \mathbf{K}_r \mathbf{\Lambda}_r^{-(1/2)}.$$

Las coordenadas para los lugares en cada uno de los espacios se obtienen:

- En el espacio de las variables:

$$\mathbf{X}_{COIA}^Z = \mathbf{Z} \mathbf{D}_p \mathbf{v}_r^Z$$

- En el espacio de las especies:

$$\mathbf{X}_{COIA}^Y = \mathbf{Y} \mathbf{D}_q \mathbf{v}_r^Y.$$

En la figura 3.3 se muestra el esquema de este método.

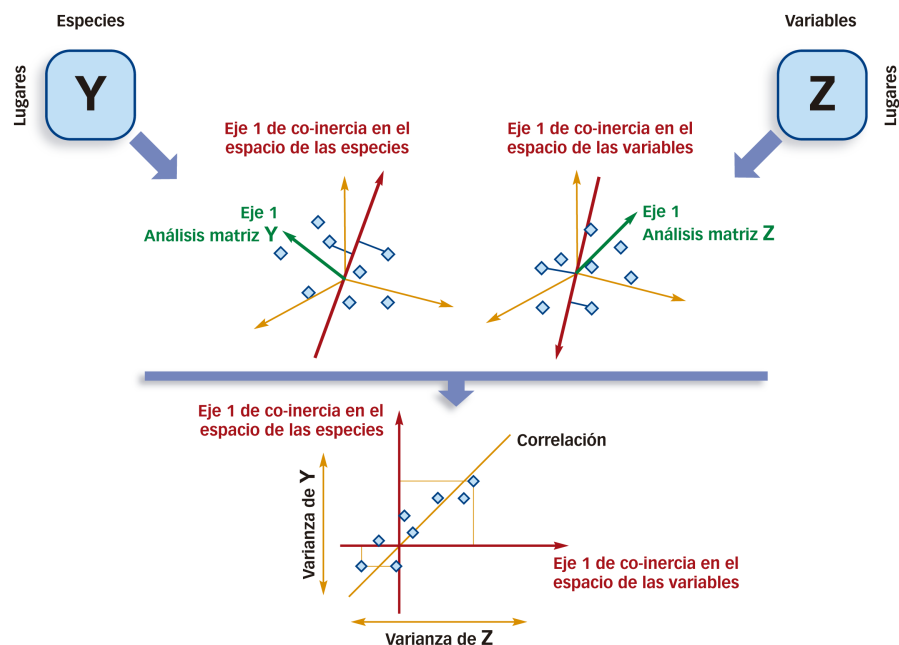


Figura 3.3: Esquema del análisis de la Coinercia.

Esta técnica suele ser utilizada como alternativa al CCA o al CNCA en el caso

de tener datos cuyo número de variables excede al número de lugares analizados. Estos datos cada vez son más frecuentes sobretodo en campos como la genética y la proteómica donde, sobre un conjunto muy pequeño de pacientes se han medido cantidades elevadas de variables.

### 3.5. Bootstrap sobre Análisis Canónico de Correspondencias Simétrico y no Simétrico

Todos los métodos explicados anteriormente presentan los resultados obtenidos de una forma meramente descriptiva, sin proporcionar medidas que permitan decidir sobre la precisión de dichos resultados. Para ello, algunos autores proponen la utilización de la metodología Bootstrap (Efron, 1979, 1987; Efron y Tibshirani, 1993) explicada en la sección 2.3 para estudiar dicha precisión. De este modo, Meulman, 1982 propone volver a ejecutar en cada muestra bootstrap un CA y poner todas las soluciones juntas. Greenacre, 1984 propuso usar las réplicas como elementos suplementarios en el análisis de la matriz original. Knox y Peet, 1989 aplican los métodos Bootstrap sobre el DCA. La estabilidad del NSCA se ha estudiado también utilizando los métodos Bootstrap (Balbi, 1992). Markus y Visser, 1992 aplican dichos métodos para generar regiones de confianza en el MCA. Reiczigel, 1996 desarrollaron un test bootstrap para el CA que incluye la construcción de intervalos de confianza. Lombardo y Ringrose, 2012; Lombardo et al., 2012 proponen la construcción de regiones de confianza tanto para el NSCA como para el CA clásico.

Con el fin de evaluar la precisión (expresada en términos de sesgo, error estándar e intervalos de confianza) de la información extraída de matrices de datos utilizando el CCA en su versiones simétrica y no simétrica, se utilizaron los

métodos Bootstrap (Efron, 1979, 1987; Efron y Tibshirani, 1993).

Para llevar a cabo este análisis se parte de la matriz  $\mathbf{Y}$  de composición de especies. Hay que tener en cuenta que las unidades de remuestreo no son los lugares sino cada individuo de cada especie que se ha contabilizado en cada uno de los lugares de muestreo. Por lo tanto, el remuestreo se va a realizar manteniendo fijo los marginales columna de la muestra original. Con los resultados obtenidos a partir del remuestreo bootstrap, es posible construir intervalos de confianza para cada parámetro de interés. Se utilizan los intervalos bootstrap, basado en percentiles y BCa (resumidos en la sección 2.3).

### 3.6. Programa *cncaGUI*

Dado que el análisis Canónico de Correspondencias no Simétrico no tiene un software específico para poder utilizarse se ha desarrollado en R (R-Team, 2014) una nueva interfaz gráfica de usuario que permita el uso de dicha técnica y además se incorpore la posibilidad de hacer inferencia de los resultados obtenidos mediante los métodos Bootstrap descritos anteriormente. En la interfaz se han incorporado también el CCA con su correspondiente versión inferencial y el COIA como alternativa a ambos en el caso de tener más variables que individuos aunque, esta última únicamente se incorpora de un modo descriptivo para complementar la utilidad de dicha interfaz.

El nuevo paquete implementado recibe el nombre de *cncaGUI*. Este paquete está disponible en <http://cran.r-project.org/web/packages/cncaGUI>. El nombre se debe a las iniciales del *análisis Canónico de Correspondencias no Simétrico* (**cnca**) e *Interfaz Gráfica de Usuario* (**GUI**).

Al igual que en el caso del paquete descrito en el capítulo anterior, es necesario bajar e instalar R desde la web [cran.r-project.org](http://cran.r-project.org). A continuación hay que

descargar el paquete `cncaGUI` y sus dependencias que son los paquetes: `rgl`, `tcltk`, `tcltk2`, `tkrplot` y `shapes`; (Adler y Murdoch, 2012; Dryden, 2014; Grosjean, 2012; Tierney, 2012).

Para poder utilizar el paquete `cncaGUI` desde R, se debe introducir el comando `library(cncaGUI)` en la consola de R. Una vez realizadas estas acciones, se deben cargar los datos que se quiere analizar. Los datos necesarios para que la interfaz funcione son:

- `especies`, matriz de datos que contiene la información de  $n$  lugares sobre los que se han medido la abundancia de  $q$  especies.
- `variables`, matriz de datos que contiene la información de los  $n$  lugares (los mismos que la matriz de especies) sobre los que se han evaluado  $p$  variables.

Una vez cargados los datos, para que se muestre la ventana principal (figura 3.4) se debe introducir el comando `cnca(especies,variables)` en la consola.

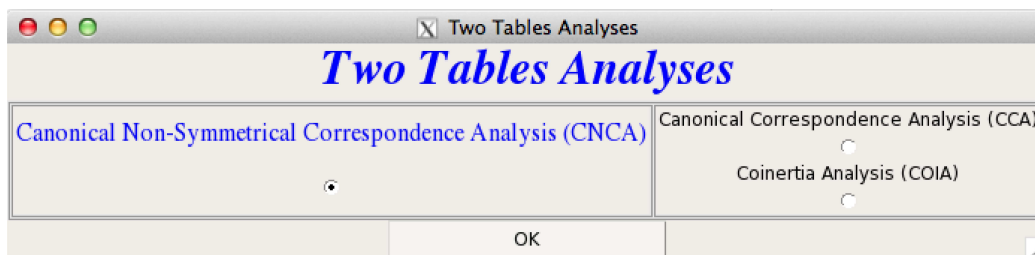


Figura 3.4: Ventana principal.

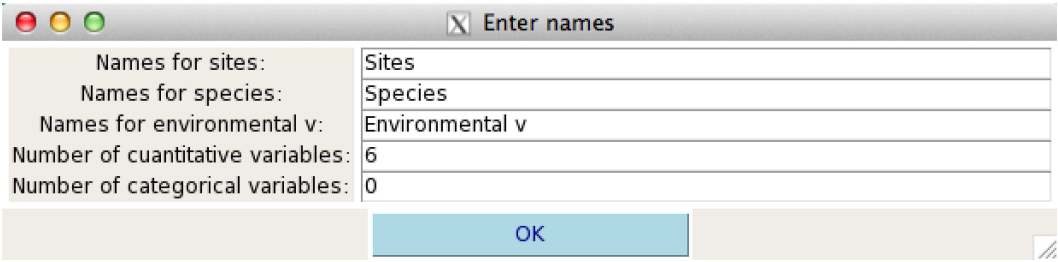
Esta ventana presenta tres radiobuttons que permiten elegir el análisis que se va a utilizar sobre los datos introducidos: análisis Canónico de Correspondencias no Simétrico, análisis Canónico de Correspondencias Simétrico y análisis de la Coinercia.

Cuando se ha elegido el análisis que se va a aplicar sobre los datos el programa realiza las siguientes comprobaciones:



- El usuario ha introducido dos matrices para analizar.
- El número de filas es igual para ambas matrices.
- En el caso de haber elegido CCA o CNCA, se comprueba que el número de filas es superior al número de columnas para ambas matrices. En caso contrario se sugiere la utilización del COIA.

Dado que estas técnicas se pueden utilizar en otros contextos diferentes del ecológico, a continuación se presenta una ventana (figura 3.5) que permite cambiar la denominación de *lugares*, *especies* y *variables* por las unidades que se estén analizando ya que dichas etiquetas se utilizarán en los gráficos y tablas que resulten del análisis. También es posible analizar datos que contengan variables cualitativas. Para ello hay que indicar al programa si los datos contienen variables cualitativas y en caso de tenerlas, cuantas variables son cuantitativas y cuantas cualitativas. El programa elige por defecto las  $p_A$  primeras columnas como cuantitativas y el resto como cualitativas.



Names for sites:	Sites
Names for species:	Species
Names for environmental v:	Environmental v
Number of quantitative variables:	6
Number of categorical variables:	0

OK

Figura 3.5: Ventana para cambio de denominación.

Una vez que el usuario pulsa el botón OK aparece la ventana de opciones (figura 3.6).

En ella es posible:

- Elegir una transformación previa a realizar sobre los datos originales:

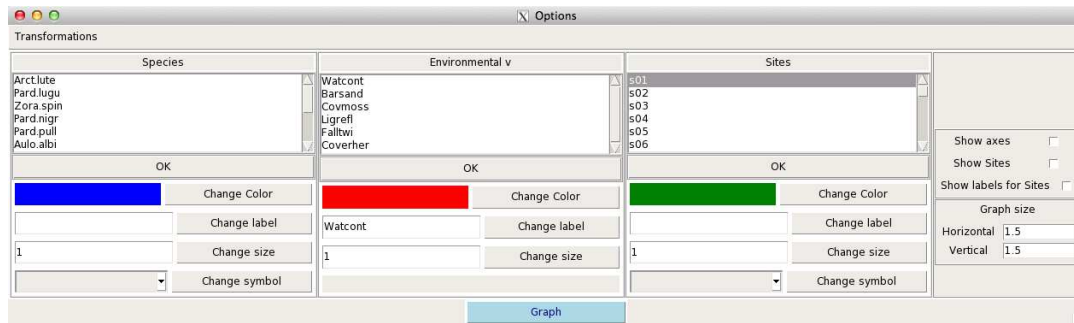


Figura 3.6: Ventana de opciones.

- Restar la media global
  - Centrar por columnas
  - Estandarizar por columnas
  - Centrar por filas
  - Estandarizar por filas
  - Doble centrado
  - No realizar ninguna transformación
- Cambiar color, tamaño, etiqueta y símbolo que van a representar a los lugares en los gráficos.
  - Cambiar color, tamaño, etiqueta y símbolo que van a representar a las especies en los gráficos.
  - Cambiar color, tamaño y etiqueta que van a representar a las variables en los gráficos.
  - Mostrar los ejes de coordenadas en los gráficos.
  - Representar en el gráfico las coordenadas para los lugares.
  - Mostrar en el gráfico las etiquetas correspondientes a los lugares en el caso de que se haya elegido la opción de representar los lugares.

- Cambiar el tamaño de las ventanas para que el programa se pueda utilizar en diferentes tamaños de pantalla.

Una vez que se pulsa el botón **Graph** emerge una ventana (figura 3.7) que contiene un diagrama de barras con el porcentaje de información que explica cada eje.

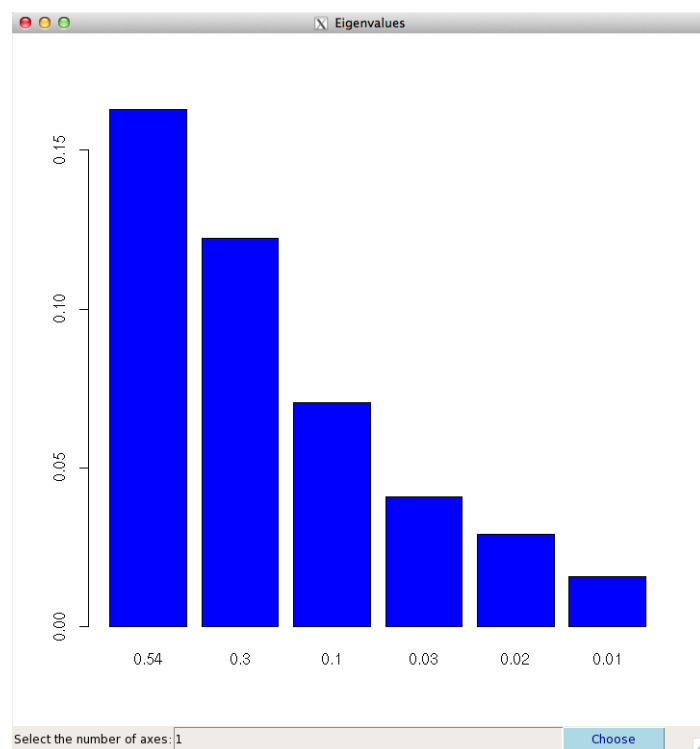


Figura 3.7: Ventana de barplot.

Una vez que se han introducido el número de ejes necesarios, se pulsa el botón **Choose** y aparece la ventana (figura 3.8) que contiene el gráfico con las coordenadas para especies y variables ambientales sobre las dos primeras dimensiones si se ha elegido CCA o CNCA o las coordenadas de los lugares en el espacio de las variables y de las especies superpuestos sobre el mismo gráfico en el que cada par de coordenadas está unido por una flecha.

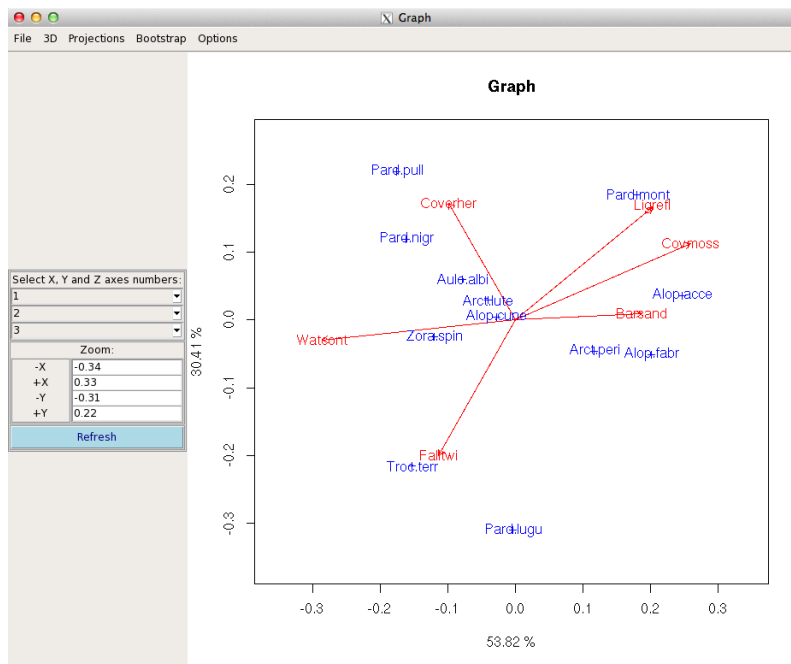


Figura 3.8: Ventana que muestra la representación de las coordenadas obtenidas mediante el método elegido en dos dimensiones.

En esta ventana se pueden ver dos partes y varios menús que cambian según el análisis elegido en la primera ventana. En la parte de la derecha se puede observar la representación gráfica de las coordenadas para especies y variables ambientales. En este gráfico es posible mover las etiquetas de los puntos con el botón izquierdo del ratón y se pueden cambiar las características gráficas de los puntos con el botón derecho del ratón. En la parte izquierda se dispone de tres listbox que sirven para elegir la dimensión que se quiere ver en cada gráfico (2 y 3 dimensiones); cuatro textbox que sirven para elegir los límites de los ejes  $x$  e  $y$  para ver una zona del gráfico con más detalle o menos y un botón **Refresh** que es necesario pulsar para que sea efectivo el cambio de los anteriores elementos.

Los menús de la parte superior dependen del análisis elegido. Si los métodos que se están utilizando son el CCA o el CNCA los menús y sus respectivos submenús son los siguientes:

- File
  - Copy image
  - Save image
    - PDF file
    - Eps file
    - Png file
    - Jpg/Jpeg file
  - Exit
- 3D
  - 3D
- Projections
  - Species
  - Sites
  - Back to original data
- Bootstrap
  - Bootstrap
- Options
  - Change title
  - Show/Hide axes
  - Show/Hide Sites
  - Show/Hide labels for Sites

El primero de ellos permite copiar la imagen al portapapeles, guardarla en varios tipos de formato y salir del programa. El siguiente es el menú que permite mostrar el gráfico en 3 dimensiones. En el caso en que se hayan retenido más de tres ejes, se mostrarán los ejes que se hayan elegido en los listbox de la ventana de gráfico. En este gráfico es posible rotar la imagen con el botón izquierdo del ratón y ampliar o disminuir la imagen con el botón derecho del ratón. El tercer menú que se encuentra disponible sirve para proyectar los puntos que representan a los lugares o a las especies sobre una determinada variable. Si se elige el submenú **Species** se presenta una ventana con el listado de las variables que están siendo analizadas, al elegir una de ellas y pulsar el botón **OK**, se muestran en el gráfico las proyecciones de las especies sobre dicha variable. Si se elige la opción **Sites** aparece la misma ventana pero se proyectan los lugares sobre la variable seleccionada. Esta opción sólo funciona si se tiene habilitada la opción de ver los lugares en el gráfico de dos dimensiones. Si lo que se desea es volver al gráfico anterior se elige el submenú **Back to original graph**. El siguiente menú es el relativo al análisis inferencial a través de la metodología Bootstrap. Cuando se pulsa sobre el submenú **Bootstrap** emerge una ventana (figura 3.9) para elegir los parámetros para los que se quieren medidas de precisión, el número de réplicas bootstrap para generar dichos resultados y el nivel de confianza con los que se van a calcular los intervalos de confianza que se presentan. Esta opción genera gráficos que contienen los histogramas y los gráficos de normalidad de cada uno de los parámetros elegidos. Se guardan en formato .eps y en .pdf y en la ventana es posible elegir si se generan en blanco y negro o en color.

En el caso de haber optado por el COIA los menús disponibles son los siguientes:

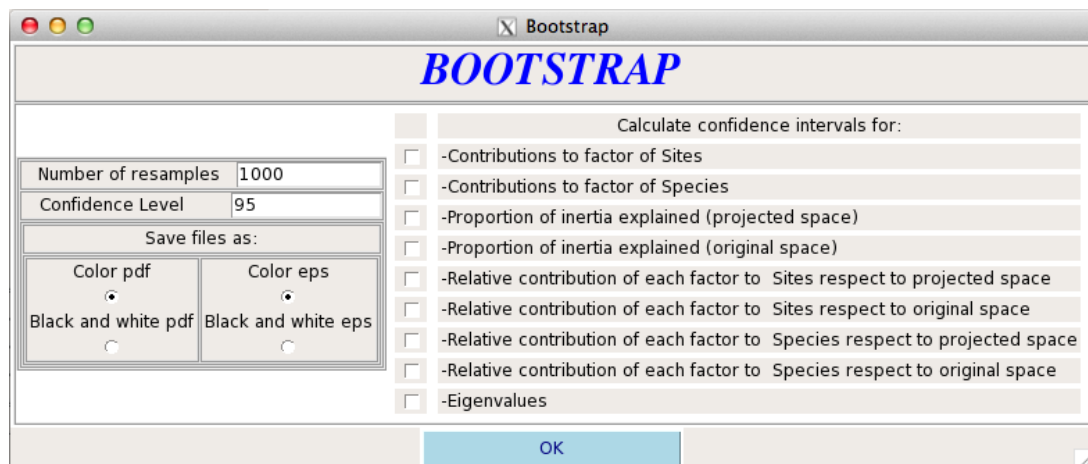


Figura 3.9: Ventana para el análisis Bootstrap.

- File
  - Copy image
  - Save image
    - PDF file
    - Eps file
    - Png file
    - Jpg/Jpeg file
  - Exit
- 3D
  - 3D
- COIA
  - Show all graphs
  - Sites graph
  - Species graph

- Environmental variables graph
- Species axes
- Enviromental variables axes
  
- Options
  - Change title
  - Show/Hide axes

Los distintos submenús del menú COIA muestran:

- Sites graph. Se muestra el gráfico con la superposición de las coordenadas para los lugares en el espacio de las especies y en el de las variables.
- Species graph. Se muestra el gráfico con las coordenadas para las especies.
- Environmental variables graph. Se muestra el gráfico con las coordenadas para las variables.
- Species axes. Se muestra el círculo unidad con la correlación de las especies con los ejes.
- Enviromental variables axes. Se muestra el círculo unidad con la correlación de las variables con los ejes.
- Show all graphs. Aparece una ventana nueva en la que se muestran todos los gráficos anteriores.

Junto a esta ventana emerge un archivo de texto con los resultados del análisis elegido (coordenadas, bondad de ajuste, contribuciones de los elementos a los factores y de los factores a los elementos, inercia absorbida por cada eje y valores propios).



En cuanto a los resultados bootstrap, aparece otro archivo que muestra para cada parámetro elegido el valor observado, la media calculada a partir de los valores del parámetro en cada remuestreo, la desviación estándar, el sesgo y los extremos inferior y superior de los intervalos de confianza t-bootstrap, percentiles y BCa descritos en la sección 2.3.2.

Estos dos archivos de texto se guardan automáticamente en el directorio en el que se encuentre el usuario así como todos las figuras que contienen histogramas y gráficos de normalidad de los parámetros seleccionados en la pantalla principal.

Además, el programa muestra tres gráficos donde aparecen representadas todas las coordenadas de las variables, de las especies y de los lugares que se han calculado a partir de las muestras bootstrap. Para ello se han realizado rotaciones Procrustes con el objetivo de poder evaluar las similitudes y diferencias de las configuraciones obtenidas para variables, especies y lugares. Cada grupo de coordenadas que representan a la misma variable se muestra en el mismo color y se encuentra encerrado bajo una línea poligonal cerrada (convex hull).

### 3.7. Aplicación a Datos

Para ilustrar el manejo del programa y la interpretación de los resultados se va a realizar un ejemplo con unos datos reales. Los datos utilizados contienen información de 12 especies de arañas recogidas en trampas en una zona dunar holandesa (Ter Braak, 1986). Las especies son:

- Alop acce. *Alopecosa accentuata*.
- Alop cune. *Alopecosa cuneata*.
- Alop fabr. *Alopecosa fabrilis*.
- Arct lute. *Arctosa lutetiana*.

- Arct peri. *Arctosa perita*.
- Aulo albi. *Aulonia albimana*.
- Pard lugu. *Pardosa lugubris*.
- Pard mont. *Pardosa monticola*.
- Pard nigr. *Pardosa nigriceps*.
- Pard pull. *Pardosa pullata*.
- Troc terr. *Trochosa terricola*.
- Zora spin. *Zora spinimana*.

Las muestras han sido recogidas sobre 28 puntos de muestreo. Sobre estos mismos puntos se han medido las siguientes variables:

- Water content. Porcentaje de masa seca del suelo.
- Bare sand. Porcentaje de cobertura de arena.
- Fallen twigs. Porcentaje de cobertura de hojas y ramas caídas.
- Cover moss. Porcentaje de cobertura de la capa de musgo.
- Cover herbs. Porcentaje de cobertura de la capa de hierba.
- Light refl. Reflejo de la superficie con cielo despejado.

Por lo tanto, se dispone de dos matrices que comparten la dimensión relativa a los lugares de muestreo que tienen dimensiones  $28 \times 12$  y  $28 \times 6$  respectivamente.

Como ilustración de la utilización del programa y por ser una técnica menos conocida, se va a aplicar un CNCA a los datos. Se estandarizan las variables por

columnas y se retienen tres ejes. En la tabla 3.3 se muestran los valores propios, la variabilidad explicada por cada eje y la acumulada para el espacio proyectado. En ella se puede observar que los tres primeros ejes explican el 95,65% de la variabilidad total.

No.	Valor Propio	Variabilidad	Variabilidad Acumulada
1	0.186	60.76	60.76
2	0.120	25.06	85.82
3	0.075	9.83	95.65

Tabla 3.3: Valores propios y variabilidad explicada (%) y acumulada para los tres primeros ejes retenidos. Espacio Projectado.

Esta información también se encuentra disponible para el espacio original (tabla 3.4) observándose como la variabilidad explicada es mucho mayor en el espacio proyectado ya que en el original los tres primeros ejes sólo explican un 63,06%.

No.	Variabilidad	Variabilidad Acumulada
1	40.06	40.06
2	16.52	56.58
3	6.48	63.06

Tabla 3.4: Variabilidad explicada (%) y acumulada para los tres primeros ejes retenidos. Espacio Original.

La interfaz también nos ofrece las contribuciones de los lugares y de las especies a la conformación de los ejes factoriales (expresadas en tanto por mil). Estas contribuciones sirven para saber cuales son los elementos que tienen mayor influencia en la dirección de los ejes factoriales. Dichas cantidades se encuentran en el apéndice A. Como se puede observar para los lugares, en general, dichas contribuciones están repartidas entre todos los lugares. Esto significa que todos los lugares contribuyen de una manera similar a la conformación de los tres ejes factoriales considerados a excepción de los casos s14 y s07 para el segundo eje y s25 para el tercero ya que sus contribuciones son más elevadas que las del resto.

Respecto a las especies, se aprecia que la que más contribuye a la orientación del primer eje es *Trochosa terricola* con una contribución del 237,08. También contribuyen a dicho eje las especies *Alopecosa accentuata*, *Alopecosa fabrilis* y *Pardosa monticola*. Las especies que más influencia tienen en la conformación del segundo eje factorial son *Pardosa lugubris* (304,44) y *Pardosa pullata* (281,14). Para el tercer eje, las especies más importantes en su dirección son *Pardosa Monticola*, *Alopecosa fabrilis* y *Pardosa nigriceps* con contribuciones de 465,27, 165,77 y 114,62 respectivamente.

También es posible analizar las contribuciones de los factores retenidos a cada uno de los elementos (lugares y especies) tanto en el espacio proyectado como en el original. Como en el caso anterior, el programa las proporciona expresadas en tanto por mil.

En el apéndice A se presentan las contribuciones de los factores a los lugares en el espacio proyectado. Las contribuciones de los tres ejes factoriales considerados a cada uno de los lugares es superior a 900 excepto para s06 y s13 que tienen contribuciones de 690,43 y 847,91 respectivamente, lo que nos permite afirmar que las contribuciones son muy buenas. Se observa además, que la mayoría de los lugares tienen su mayor contribución proveniente del primer eje; sólo un grupo muy pequeño de ellos recibe su mayor aporte del segundo eje y un único caso del tercer eje.

En el apéndice A también se pueden consultar las contribuciones de los factores a los lugares si se considera el espacio original. Se puede ver que las contribuciones de los ejes a cada lugar son buenas aunque menos que en el espacio proyectado ya que únicamente el 39 % de los lugares tienen una contribución total de los tres ejes superior a 900. En este caso, se sigue observando que la mayor parte de la contribución proviene del primer eje retenido.

Las contribuciones de los tres primeros ejes a las especies en el espacio proyectado se pueden observar en la tabla 3.5. Analizando la suma de las contribuciones de los tres primeros factores a la representación de cada especie, se puede ver que dos terceras partes tienen una contribución total de más de 900. El resto, aunque no llegan a ser tan altas, se pueden considerar buenas al no bajar ninguna de 500. Si se observa eje a eje, la mitad de las especies tiene su mayor contribución procedente del primer eje; un grupo muy reducido del segundo eje ninguna de ellas del tercero, con lo cual se deduce que las especies están bien representadas en el plano factorial 1 – 2.

Elemento	Eje 1	Eje 2	Eje 3
Arct.lute	157.71	497.94	173.94
Pard.lugu	402.81	535.43	44.16
Zora.spin	603.27	44.16	97.75
Pard.nigr	255.91	488.68	225.45
Pard.pull	130.98	831.45	0.76
Aulo.albi	256.55	405.54	33.12
Troc.terr	875.60	92.20	17.14
Alop.cune	310.63	6.58	202.05
Pard.mont	568.65	105.95	301.66
Alop.acce	924.11	38.05	6.24
Alop.fabr	722.47	148.72	117.65
Arct.peri	604.10	236.06	135.84

Tabla 3.5: Contribuciones relativas de los primeros tres factores a las especies. Espacio Proyectado.

En cuanto a las contribuciones de los factores a las especies en el espacio original, presentadas en la tabla 3.6, cabe destacar que las contribuciones totales de los tres ejes es mucho más baja que cuando se trata del espacio proyectado. La mayoría de las especies presenta una mayor contribución procedente del primer eje, al igual que se observó en el espacio proyectado.

Elemento	Eje 1	Eje 2	Eje 3
Arct.lute	99.18	265.30	108.32
Pard.lugu	228.68	489.46	114.79
Zora.spin	569.66	19.71	52.65
Pard.nigr	213.36	443.31	137.96
Pard.pull	179.39	625.09	84.67
Aulo.albi	184.14	312.91	99.34
Troc.terr	647.11	162.46	70.38
Alop.cune	110.94	46.80	136.56
Pard.mont	424.42	132.55	429.33
Alop.acce	834.49	43.02	5.21
Alop.fabr	570.13	208.43	143.98
Arct.peri	294.79	193.47	230.52

Tabla 3.6: Contribuciones relativas de los primeros tres factores a las especies. Espacio Original.

A continuación se muestra el gráfico donde se representan los lugares y las especies mediante puntos y las variables ambientales mediante vectores (Figura 3.10). Observando dicho gráfico y las contribuciones de las especies a la conformación de los ejes factoriales se puede ver que el eje 1 está caracterizado por la presencia de *Pardosa monticola*, *Alopecosa accentuata* y *Alopecosa frabilis* en el lado positivo y por *Trochosa terricola* en su lado negativo. En cuanto al eje 2, las especies *Pardosa lugubris* y *Pardosa pullata* caracterizan la parte positiva y negativa respectivamente. Si se tiene en cuenta las variables ambientales representadas, se puede considerar que el eje 1 diferencia entre zonas húmedas con alto porcentaje de arena y musgo y que tienen mejor reflejo con cielo despejado. El eje 2 diferencia zonas con alta cobertura de hierbas. Respecto a las variables, se observa un gradiente respecto del eje 1 que de izquierda a derecha, va de lugares secos, con alto porcentaje de follaje y bajo porcentaje de arena y musgo a lugares más húmedos con mayor cobertura de arena, musgo y mejor reflejo con cielo despejado.

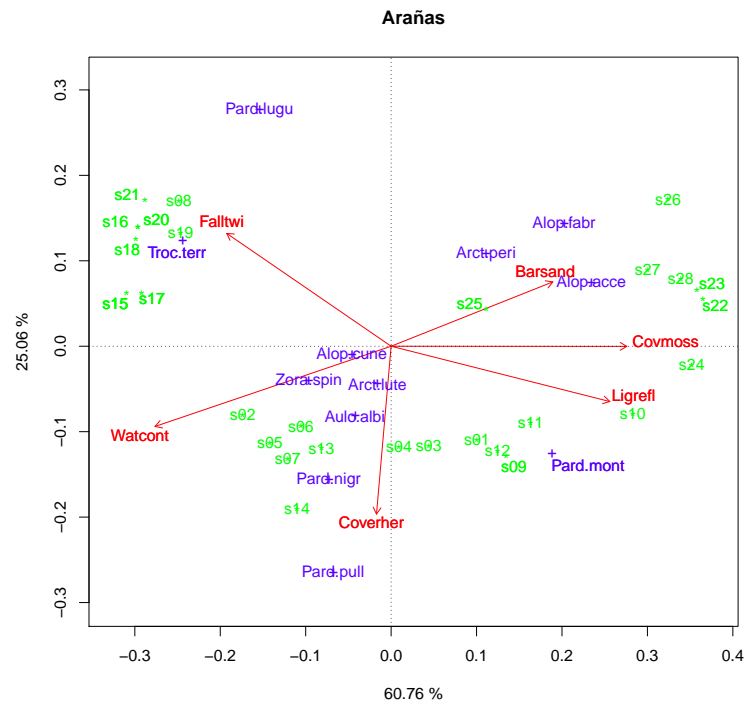


Figura 3.10: Gráfico que muestra la representación de los datos en las dos primeras dimensiones.

Una vez analizadas las estimaciones puntuales de los parámetros calculados para el CNCA, se analizan los resultados provenientes de la versión inferencial que se ha desarrollado en el presente capítulo para dicho análisis. Dicha versión, realizada a partir del remuestreo Bootstrap, ofrece, para cada uno de los parámetros explicados anteriormente, información acerca del valor puntual obtenido mediante la muestra original para el parámetro considerado, la media de todos los valores obtenidos en cada remuestreo, el sesgo entre el valor observado y dicha media, la desviación estándar y los intervalos de confianza t-bootstrap, percentiles y BCa calculados según se explicó en la sección 2.3.2. Dichos cálculos se realizaron a partir de 1000 muestras y con un nivel de confianza del 95 %.

En primer lugar, se van a analizar los resultados obtenidos para los valores propios (tabla 3.7). Se observa que las medias de los 1000 valores obtenidos a partir de las muestras extraídas y los valores observados muestran unas diferencias muy pequeñas. El error estándar para los 6 valores propios es menor que 0,01 en todos los casos y los sesgos son inferiores a 0,02. Al fijarse en la amplitud de los tres tipos de intervalos de confianza, se puede ver que la máxima se alcanza para el valor propio 1 en los intervalos t-bootstrap y percentiles con un valor de 0,05, dicha amplitud también se observa para el valor propio 3 en los mismos tipos de intervalos. También se puede apreciar que, para cada valor propio, el intervalo con menor amplitud es el BCa.

	V. Obs	Media	SE	Sesgo	t-b inf	t-b sup	perc inf	perc sup	BCa inf	BCa sup
1	0.19	0.19	0.01	0.00	0.17	0.21	0.17	0.21	0.16	0.20
2	0.12	0.13	0.01	0.01	0.11	0.14	0.11	0.14	0.10	0.13
3	0.08	0.08	0.01	0.01	0.06	0.11	0.06	0.11	0.06	0.09
4	0.04	0.06	0.01	0.02	0.04	0.07	0.04	0.07	0.03	0.04
5	0.03	0.04	0.01	0.01	0.02	0.05	0.02	0.05	0.02	0.03
6	0.02	0.02	0.01	0.01	0.01	0.04	0.01	0.04	0.01	0.02

Tabla 3.7: Resultados bootstrap para valores propios.

Continuando con el análisis inferencial de los valores propios, se muestran en la figura 3.11 los histogramas y los gráficos de normalidad. Dichos gráficos se construyen a partir de los valores calculados utilizando las 1000 submuestras extraídas mediante remuestreo Bootstrap. En los histogramas se muestran, a su vez, los valores observados y la media calculada a partir de las réplicas bootstrap mediante una línea roja discontinua y una línea azul continua respectivamente. Según se observa en los gráficos, las líneas descritas anteriormente se encuentran próximas para los seis valores propios y los gráficos de normalidad muestran que los valores calculados a partir del remuestreo siguen una distribución normal. Esto indica que los valores propios se mantienen estables para estos datos.



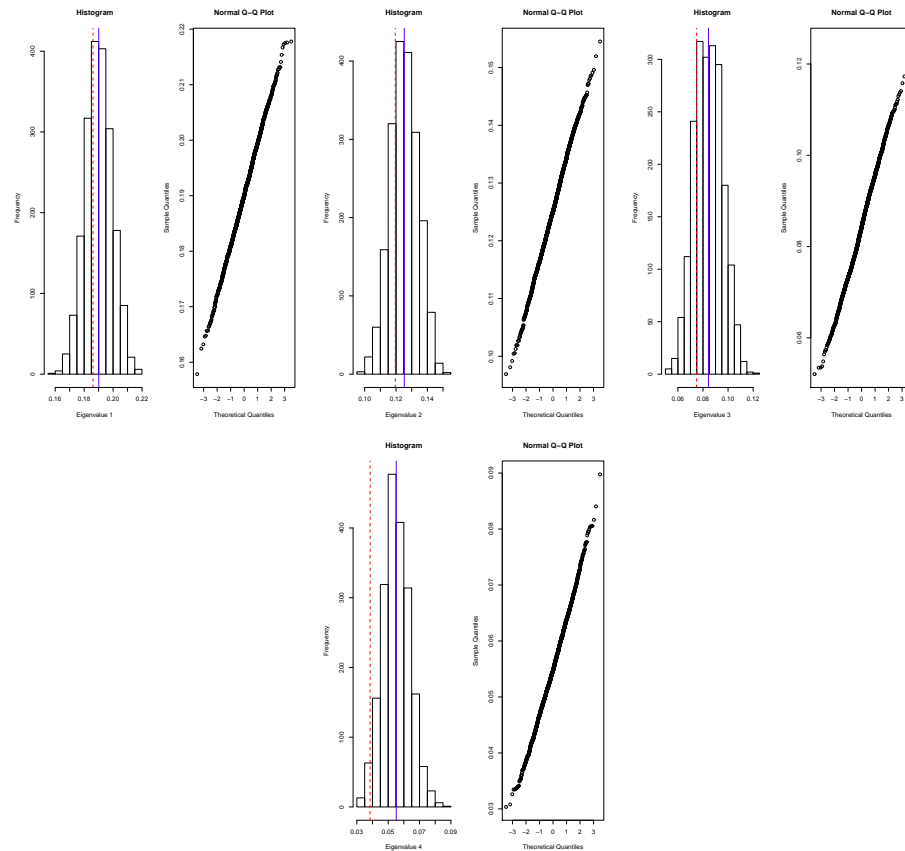


Figura 3.11: Histogramas y gráficos de normalidad para los valores propios.

En la tabla 3.8, se muestran los resultados inferenciales para la proporción de inercia absorbida por cada uno de los tres ejes retenidos referidos al espacio proyectado. Para estos parámetros se tiene que las medias de las cantidades obtenidas mediante el remuestreo difieren un poco de los valores observados. Aunque es necesario tener en cuenta que estas cantidades están multiplicadas por mil, hecho por el cual dichas diferencias resultan mucho mayores a primera vista. Por lo tanto, los sesgos obtenidos son pequeños así como los errores estándar. Si se observa la amplitud de los intervalos de confianza pero teniendo en cuenta el valor original, es decir, dividiendo cada cantidad por 1000, se puede apreciar que todas son inferiores a 0,15 lo que se traduce en una gran estabilidad para dichos parámetros. Al igual que sucedía en el caso de los valores propios, el tipo

de intervalo que presenta una menor amplitud para cada uno de los ejes es el BCa. En la figura 3.12 se muestran los histogramas y los gráficos de normalidad para dichas proporciones.

	V. Obs	Media	SE	Sesgo	t-b inf	t-b sup	perc inf	perc sup	BCa inf	BCa sup
E.1	607.62	562.92	35.39	-44.70	493.46	632.37	497.82	631.85	583.41	706.95
E.2	250.55	244.00	27.51	-6.55	190.01	297.99	192.88	300.89	205.96	313.12
E.3	98.28	111.66	26.26	13.38	60.14	163.18	65.68	167.60	51.32	139.03

Tabla 3.8: Resultados bootstrap para la inercia absorbida por cada eje en el espacio proyectado.

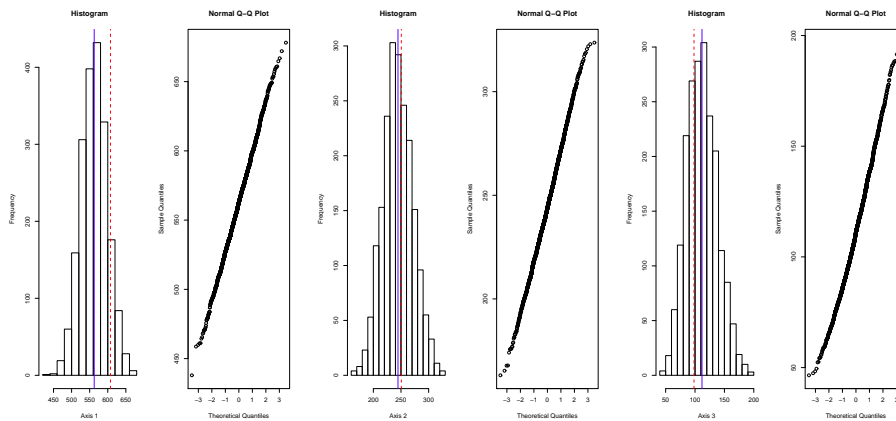


Figura 3.12: Histogramas y gráficos de normalidad para la proporción de inercia explicada por cada eje en el espacio proyectado.

En el apéndice A se muestran los resultados para las contribuciones de las especies y los lugares a la conformación de los ejes factoriales, así como la contribución de los ejes factoriales a cada uno de los elementos (lugares y especies) tanto en el espacio proyectado como en el original. Teniendo en cuenta, como en el caso anterior, que las cantidades están multiplicadas por 1000, los sesgos entre los valores observados y las medias obtenidas a partir de las muestras bootstrap son muy pequeños para las contribuciones tanto de los lugares como de las variables a la conformación de los ejes factoriales. Sólomente en 26 parámetros se supera el 0,01. Sin embargo, aunque en media no difieran mucho, se puede observar que la amplitud de los intervalos de confianza es notable en algunos casos. Este

hecho indica que las contribuciones de los elementos a la dirección de los ejes no es tan estable como los parámetros estudiados anteriormente. En cuanto a las contribuciones de los ejes factoriales a los elementos (lugares y especies) tanto en el espacio proyectado como en el original, se observa también que los sesgos no superan el 0,1 en el 84 % de los parámetros analizados. Aunque, como sucedía con las contribuciones de los elementos a la conformación de los ejes factoriales, la diferencia entre los extremos superior e inferior de los intervalos de confianza indican una estabilidad leve. Cabe destacar, que la mayoría de los extremos inferiores calculados mediante el método t-bootstrap son negativos para las contribuciones. Este suceso hace pensar que este método no es adecuado para estos parámetros ya que las contribuciones nunca son negativas y por esto, dicho método nos proporciona unos resultados irreales. Por último, un aspecto importante a tener en cuenta es cuando el programa no proporciona intervalos de confianza para el tipo BCa. En vez de disponer de los extremos inferior y superior, aparece *NA*. Esto es debido a que los valores de dicho parámetro para cada remuestreo son todos menores que el valor observado y no se puede calcular el intervalo BCa bajo esas características. En estos casos, se deduce que el valor observado para la muestra original no es muy representativo y la estabilidad para dicho parámetro sería mala. Por motivos de espacio no se incluyen los histogramas y los gráficos de normalidad para las contribuciones.

Por último, se presentan tres gráficos (figura 3.13). En cada uno de ellos se representan el conjunto de variables ambientales, el conjunto de especies y el de lugares. Para cada uno de ellos, se muestran las 1000 coordenadas que se han calculado para cada submuestra extraída mediante el método Bootstrap. Cada grupo de coordenadas referidas al mismo elemento se representa con el mismo color y delimitados por un polígono envolvente. Como se puede observar, las variables tienen una zona de presencia muy compacta que indica que este conjunto

de elementos es estable. Respecto a especies y lugares, las zonas de cada uno son más extensas y se solapan pero se sigue apreciando estabilidad. El solapamiento se debe principalmente a la cercanía de los puntos en sí que se puede observar en el gráfico resultante del análisis de la muestra original.

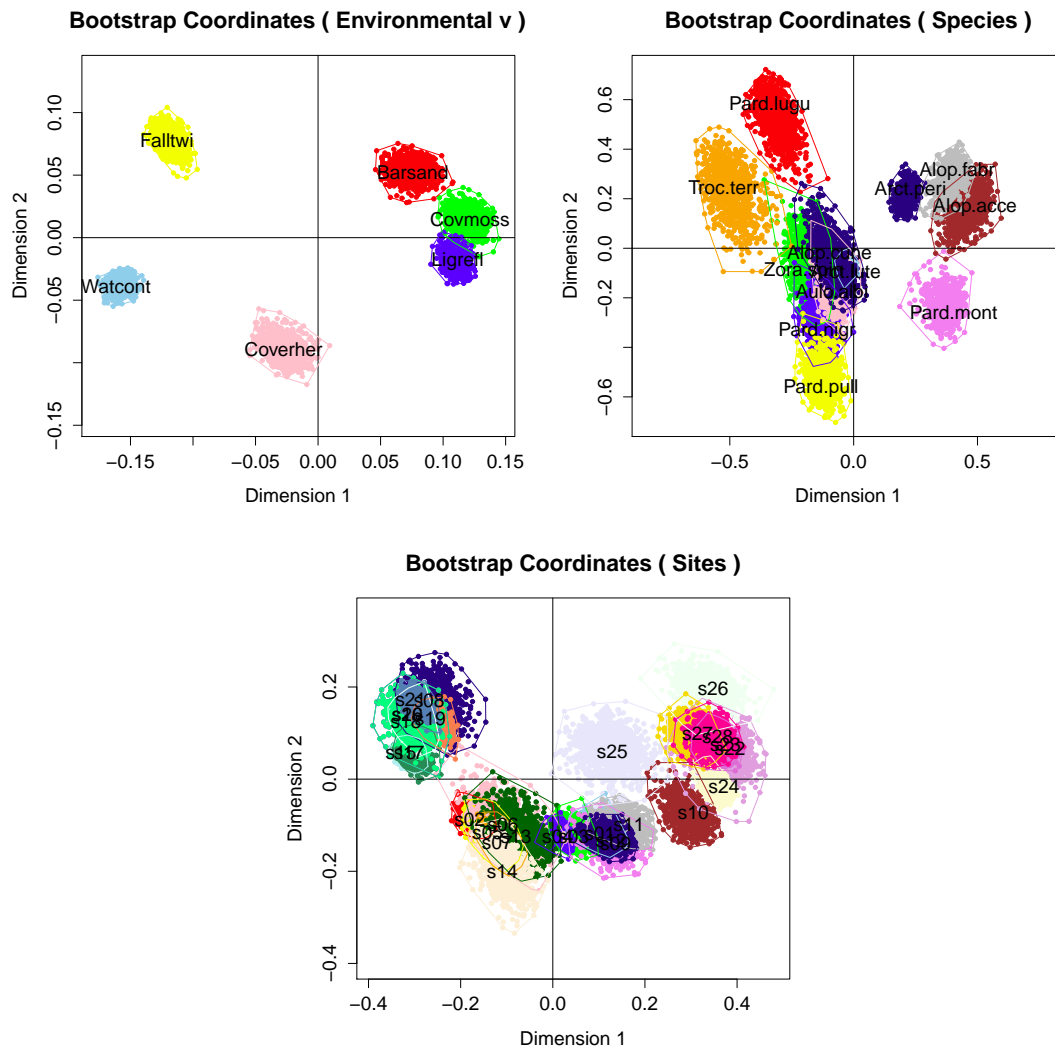


Figura 3.13: Coordenadas de variables, especies y lugares para las 1000 submuestras seleccionadas mediante Bootstrap.

Resumiendo, en el presente capítulo se ha realizado una revisión bibliográfica de las técnicas desarrolladas para el análisis de dos matrices de datos que tienen una vía común; se han explicado los principales aspectos teóricos del CCA, CNCA

---

y COIA con objeto de presentar posteriormente un software que permita la utilización de dichos métodos; se ha presentado una versión inferencial basada en la metodología Bootstrap del CCA y del CNCA que puede ser utilizada desde el propio software y por último se ha ejecutado el programa con unos datos de prueba para poder explicar la interpretación de los resultados obtenidos. El siguiente paso que cabe esperar es cómo se desarrolla una versión inferencial en el contexto de tres vías. Este tema se abordará en el siguiente capítulo.



## Capítulo 4

# BIPLOT MÚLTIPLES





## 4.1. Introducción

En los capítulos anteriores se han presentado versiones inferenciales de métodos para analizar matrices de dos vías y pares de matrices que comparten una vía. El paso siguiente es poder proporcionar versiones inferenciales en el contexto del análisis de tres vías. En primer lugar se realiza una revisión de los principales métodos que permiten analizar este tipo de datos.

Los métodos clásicos de análisis multivariante como el análisis de Componentes Principales o el análisis Factorial nos permiten estudiar las interrelaciones entre un conjunto de individuos y un conjunto de variables, es decir, trabajan con datos de dos vías. Cuando estas relaciones están capturadas en diferentes condiciones experimentales o en diferentes períodos de tiempo, para cada condición o cada período tenemos una matriz de dos vías, con lo cual, la información se muestra en varias matrices que contienen medidas de un conjunto de variables sobre un conjunto de individuos. Es decir, cada observación se crea a partir de tres vías: individuos, variables y condiciones. Los datos de conjuntos múltiples (Kiers, 1988, 1991) son un caso particular de los datos de tres vías, en ellos se tiene información en los que una de las vías, combina varios conjuntos, es decir, cuando una de las vías no está totalmente cruzada. En presencia de este tipo de información, el investigador se hace varias preguntas:

- ¿Es posible identificar una estructura espacial o temporal común para todos los datos?
- ¿Es dicha estructura homogénea en todas las condiciones o a lo largo del tiempo?
- ¿Existe un comportamiento diferente de los individuos?

- ¿Dicho comportamiento se mantiene en las diferentes condiciones o a lo largo del tiempo?

Para responder a estas preguntas se han desarrollado algunos métodos que analizan este tipo de datos. Entre ellos se encuentran: la Structuration de Tableaux À Trois Indices de la Statistique (STATIS) y STATIS dual (L'Hermier des Plantes, 1976); el análisis Factorial Múltiple (MFA) (Escoufier y Pagès, 1984) y (Escoufier y Pagès, 1990) y el MFA dual (Lê y Pagès, 2007) entre otros. El objetivo de estos métodos es encontrar un espacio de representación común (consenso o compromiso) para los elementos (filas y columnas) de las diferentes matrices, y por lo tanto, intentan capturar la parte estable de las relaciones entre ellos.

Los desarrollos en la línea del STATIS se pueden clasificar de cuatro formas según se expuso en Vicente-Galindo, 2013:

1. De acuerdo a los datos de partida: Cuando cada matriz de datos contiene información del mismo conjunto de individuos medidos sobre el mismo conjunto de variables en diferentes condiciones o en diferentes momentos del tiempo, se utiliza el X-STATIS o análisis Triádico Parcial (PTA) (Jaffrenou, 1978). En este caso se trabaja con matrices de datos en lugar de operadores. Si los operadores que se quieren integrar son matrices de covarianzas sobre los mismos individuos, el método es conocido como COVSTATIS (Thioulouse, 2011). Si lo que se pretende es integrar varias matrices de distancias definidas sobre el mismo conjunto de individuos, la técnica utilizada se denomina DISTATIS (Abdi et al., 2007) y puede ser vista como la versión para tres vías del Multidimensional Scaling. Si lo que se quiere analizar son datos de tipo intervalo, el método propuesto es el INTERSTATIS (Corrales y Rodríguez, 2014).

2. De acuerdo a los pesos asignados a cada matriz cuando se construye la matriz consenso o compromiso: Benasseni y Bennani-Dosse, 2012 denominan Power-Statist a una versión del STATIS cuando el peso asignado a cada matriz difiere del original definido por L'Hermier des Plantes, 1976; ANISOSTATIS, extiende el STATIS para otorgar pesos específicos a cada variables en lugar de para cada matriz de dos vías completa (Abdi et al., 2012).
3. Si se tiene en cuenta información externa:  $K + 1$  STATIS (Sauzay et al., 2006) analiza las relaciones de  $K$  tablas con otra que contiene la información externa. Sabatier y Vivien, 2008 extienden el  $K+1$  STATIS y proponen el STATIS-4. En el caso de que la información de partida tenga una estructura predefinida de grupo, Vallejo-Arboleda et al., 2006 desarrollaron el CANOSTATIS.
4. Si disponemos de matrices pareadas en cada situación o condición: el STATICO (STATIS y Coinercia, Simier et al., 1999; Thioulouse et al., 2004) y más recientemente el COSTATIS (Coinercia y STATIS, Thioulouse, 2011), son una opción interesante que permiten estudiar las co-estructuras existentes en los datos.

Analizando el software disponible para poder utilizar dichos métodos, haciendo especial incapié en lo referente al entorno R encontramos: En el paquete `ade4` (Chessel et al., 2004); (Dray y Dufour, 2007); (Dray et al., 2007); (Thioulouse y Dray, 2007) y (Chessel et al., 2013), se pueden ejecutar los métodos STATIS, STATIS dual, PTA, STATICO y COSTATIS. En el paquete `DistatisR` (Beaton et al., 2013) el investigador puede ejecutar los métodos COVSTATIS y DISTATIS. STATIS, STATIS dual, PTA,  $K+1$  STATIS, CANOSTATIS y ANISOSTATIS se pueden encontrar en el paquete `MExPosition` (Chin Fatt

et al., 2013). Vicente-Villardón, 2003 ha implementado en el software comercial MatLab ([www.mathworks.com/products/matlab](http://www.mathworks.com/products/matlab)) un programa para realizar análisis multivariante llamado `multBiplot` en el cual es posible utilizar el método CANOSTATIS.

Desarrollos recientes en la línea del MFA son el HMFA (análisis Factorial Múltiple Jerárquico), propuesto como una extensión del MFA en el que las variables están estructuradas de acuerdo a una jerarquía (Le Dien y Pagès, 2003); cuando un conjunto de individuos es descrito por un conjunto de variables que pueden ser continuas y/o categóricas, el análisis propuesto es un caso particular del MFA llamado análisis Factorial de datos mixtos (Bécue-Bertaut y Pagès, 2008); si lo que se pretende analizar son tablas de contingencia múltiples, el método propuesto se denomina MFACT (Bécue-Bertaut y Pagès, 2004).

Estos métodos pueden ser utilizados a través del paquete `FactomineR` (Husson et al., 2012; Lê et al., 2008) entre otros.

Sin embargo, estos métodos no obtienen un espacio de representación común tanto para individuos como para variables. Para resolver este problema, estas técnicas eligen la vía que tenga la representación óptima y las otras vías se muestran como elementos suplementarios. Aunque el propósito de estos métodos es el mismo, los resultados difieren en la ponderación utilizada en cada uno de ellos. Los pesos en el MFA se calculan para hacer que el primer eje sea la unidad permitiendo relativizar la inercia del resto de ejes respecto de uno. Esto tiene como resultado que las ponderaciones en este método mantienen la multidimensionalidad presente en los datos. En el STATIS, los pesos están ligados al coeficiente RV, por lo tanto, si un grupo de variables tiene una baja correlación con el resto de grupos, esas variables estarán penalizadas al conformar el espacio compromiso. El resto de diferencias entre STATIS y MFA están explicados en Pagès, 1996.

Otro método utilizado para analizar este tipo de datos es el Metabiplot (Martín-Rodríguez et al., 2002). Este método presenta la ventaja respecto del MFA de que, aunque las direcciones principales no se muestren en el mismo orden, si son similares se pueden capturar en la solución.

Sin embargo, estas técnicas sólo muestran la parte estable de las relaciones entre las matrices, sin capturar las interacciones o las diferencias. Los métodos basados en los modelos TUCKER (Tucker, 1966) capturan esta información a través de la "Core matrix" y los Biplots conjuntos relacionados. Recientemente, se ha desarrollado un nuevo método que combina el análisis de la Coinercia (Dolédec y Chessel, 1994) y el modelo TUCKER3 para analizar pares de k-tablas y capturar la parte no estable de la información (interacciones). Esta propuesta se ha denominado CO-TUCKER (Mendes, 2011).

Teniendo en cuenta las ventajas y desventajas de los métodos mencionados anteriormente, se desarrollaron los métodos Biplot Múltiples (Baccalá, 2004) basados en las ponderaciones del MFA y las ventajas de las representaciones Biplot (Gabriel, 1971; Galindo, 1986). A continuación se presentan los aspectos teóricos más relevantes de dichos métodos.

## 4.2. Biplot Múltiples

Los Biplot Múltiples fueron introducidos por Nora Baccalá en su tesis "Contribuciones al análisis de matrices de datos multivía: Tipología de las variables" (Baccalá, 2004). Pueden ser utilizados en las dos situaciones en que se pueden presentar los datos de conjuntos múltiples:

- Varios conjuntos de individuos sobre los que se observa un único conjunto de variables.



Siendo:

$\mathbf{X}$ : tabla completa de orden  $(I \times J)$ , de término general  $x_{ij}$ , con  $i = 1, \dots, I$ ,  $j = 1, \dots, J$ .

$\mathbf{X}_t$ :  $t$ -ésima tabla, de término general  $x_{ij}^t$ , con  $i = 1, \dots, I_t$ ;  $j = 1, \dots, J$ . Es decir:  $x_{ij}^t$  valor de la  $j$ -ésima variable para el  $i$ -ésimo individuo en la tabla  $\mathbf{X}_t$ .

$I_t$ : conjunto de individuos de la  $t$ -ésima tabla ó  $t$ -ésimo grupo de individuos.

$I$ : conjunto de individuos o el número total de individuos, por tanto

$$I = \sum_{t=1}^T I_t$$

$J$ : conjunto de variables.

$I, I_t, T, J$  representan conjuntos y son cardinales.

El método se desarrolla en cinco etapas, que se describen a continuación.

### 4.3.1. Primera Etapa: Estandarización por columnas de la tabla $\mathbf{X}$ .

Se estandarizan los valores de cada variable de la tabla  $\mathbf{X}$  utilizando:

$\bar{x}_j$  Media de la variable  $j$  en la tabla  $t$ , y

$SD(j) = \frac{\sum_{i=1}^I (x_{ij} - \bar{x}_j)^2}{I-1}$  la desviación típica de la variable  $j$  en la tabla  $\mathbf{X}$ .

Restando a cada elemento su media  $\bar{x}_j$  y dividiendo por su desviación típica  $SD(j)$  se obtiene la tabla  $\mathbf{X}$  estandarizada.

Esta estandarización acentúa la dispersión interna de cada tabla pero en la matriz  $\mathbf{X}$  (concatenada), todas las variables tendrán la misma media y la misma desviación obteniéndose la matriz  $\mathbf{X}$  normada.

### 4.3.2. Segunda Etapa: Análisis Individuales

Con el objetivo de eliminar la influencia de las tablas con mayor dispersión se realiza un PCA de cada una de las  $T$  tablas con los datos estandarizados descritos en el paso anterior y en cada uno se selecciona el primer valor propio. Todos los valores de la  $t$ -ésima tabla se ponderan por  $\frac{1}{\lambda_1^t}$ , para todo  $t = 1, \dots, T$ ; siendo  $\lambda_1^t$ : primer valor propio de la  $t$ -ésima tabla.

Se obtiene así una matriz, que denominaremos  $\tilde{\mathbf{X}}$  de orden  $(IxJ)$ .

$$\tilde{\mathbf{X}} = \begin{pmatrix} \frac{1}{\lambda_1^1} \mathbf{X}_1 \\ \vdots \\ \frac{1}{\lambda_1^T} \mathbf{X}_T \end{pmatrix}$$

Por lo tanto  $\tilde{\mathbf{X}} = \mathbf{N}_\lambda \mathbf{X}$

Siendo  $\mathbf{N}_\lambda$  una matriz de orden  $I \times I$ , y su  $t$ -ésimo elemento una submatriz diagonal de orden  $(I_t \times I_t)$ .

$$\tilde{\mathbf{X}} = \left[ \begin{array}{ccc} \begin{pmatrix} \frac{1}{\lambda_1^1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \frac{1}{\lambda_1^1} \end{pmatrix} & \dots & \begin{pmatrix} 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{pmatrix} & \dots & \begin{pmatrix} 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{pmatrix} \\ \vdots & \ddots & \vdots & & \vdots \\ \begin{pmatrix} 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{pmatrix} & \dots & \begin{pmatrix} \frac{1}{\lambda_1^t} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \frac{1}{\lambda_1^t} \end{pmatrix} & \dots & \begin{pmatrix} 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{pmatrix} \\ \vdots & & \vdots & \ddots & \vdots \\ \begin{pmatrix} 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{pmatrix} & \dots & \begin{pmatrix} 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{pmatrix} & \dots & \begin{pmatrix} \frac{1}{\lambda_1^T} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \frac{1}{\lambda_1^T} \end{pmatrix} \end{array} \right] \mathbf{X}$$



### 4.3.3. Tercera Etapa: Biplot ponderado de la matriz $\tilde{\mathbf{X}}$

En esta etapa se realiza un Biplot de la matriz ponderada  $\tilde{\mathbf{X}}$ , calculada en el paso anterior.

Se proponen dos alternativas diferentes, de acuerdo al objetivo perseguido (1 ó 2) especificados anteriormente. Para ello se presentan dos factorizaciones Biplot y en ambos casos se definen medidas de calidad de la representación.

- **Factorización Biplot cuando el objetivo es aproximar los valores de  $\mathbf{X}$**

En este caso la factorización Biplot es:

$$\text{JK (Gabriel, 1971): } \mathbf{A}_q \approx (\mathbf{U}\mathbf{\Lambda})_q \text{ y } \mathbf{B}_q \approx (\mathbf{V})_q$$

Siendo:

$q$ : subíndice que indica las  $q$  primeras columnas de las matrices respectivas,  $q \leq R$ , si  $R$  es el rango de la matriz  $\tilde{\mathbf{X}}$ .

$\mathbf{A}_{(I \times q)}$  matriz que contiene los  $q$  marcadores para las filas, de elemento genérico  $a_i^t$ .

$\mathbf{B}_{(J \times q)}$  matriz que contiene los  $q$  marcadores para las columnas, de elemento genérico  $b_j$ .

En este Biplot se impone la métrica  $\mathbf{B}^\top \mathbf{B} = \mathbf{I}$ , en el espacio de las filas de la matriz  $\tilde{\mathbf{X}}_{(I \times J)}$ .

Luego:

$$\tilde{x}_{ij}^t \simeq a_i^{t\top} b_j$$

### Características de los marcadores obtenidos con el JK Biplot.

1. Los productos escalares (utilizando la métrica identidad) de los individuos de la matriz  $\tilde{\mathbf{X}}$ , coinciden en el espacio completo con los productos escalares de los marcadores fila contenidos en  $\mathbf{A}$ .

Es decir:

$$\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top = \mathbf{A}\mathbf{B}^\top(\mathbf{A}\mathbf{B}^\top)^\top = \mathbf{A}\mathbf{B}^\top\mathbf{B}\mathbf{A}^\top = \mathbf{A}\mathbf{A}^\top$$

Por lo tanto, la aproximación en dimensión reducida de dichos productos escalares es óptima en el sentido de los mínimos cuadrados.

2. La distancia Euclídea entre dos individuos de la matriz  $\tilde{\mathbf{X}}$  coincide en el espacio completo con la distancia entre marcadores fila. Dado que las filas de la matriz  $\tilde{\mathbf{X}}$  se pueden aproximar por  $\tilde{x}_i^t = \mathbf{X}a_i^t$ , se tiene:

$$\begin{aligned}(\tilde{x}_i^t - \tilde{x}_j^t)^\top(\tilde{x}_i^t - \tilde{x}_j^t) &= (\mathbf{B}a_i^t - \mathbf{B}a_j^t)^\top(\mathbf{B}a_i^t - \mathbf{B}a_j^t) = \\ &= (a_i^t - a_j^t)^\top\mathbf{B}^\top\mathbf{B}(a_i^t - a_j^t) = (a_i^t - a_j^t)^\top(a_i^t - a_j^t)\end{aligned}$$

3. Los marcadores fila coinciden con las coordenadas de las filas en el espacio de las componentes principales.

$$\tilde{\mathbf{X}}\mathbf{V} = (\mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top)\mathbf{V} = \mathbf{U}\mathbf{\Lambda} = \mathbf{A}$$

4. Las coordenadas para las columnas de la matriz  $\tilde{\mathbf{X}}$  son las proyecciones de los ejes originales en el espacio de las componentes principales.

Esta propiedad permite interpretar las coordenadas como correlaciones entre las variables originales y los ejes.

Las coordenadas para las columnas marcan la unidad para las escalas de predicción. Dado que cada una de las variables originales están asociadas con un eje en el espacio J-dimensional, las proyecciones de los vectores directores correspondientes se han denominado "Ejes Biplot", cada uno de los cuales identifica a la variable asociada (Gower y Harding, 1988). La proyección de cada uno de los marcadores fila sobre estos ejes, es una aproximación de los valores que toman los individuos en las variables correspondientes.

5. La similitud entre las columnas se aproxima mediante la inversa de la matriz de dispersión entre individuos.

$$(\tilde{x}_i - \tilde{x}_j)^\top (\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top)^{-1}(\tilde{x}_i - \tilde{x}_j) \approx (b_i - b_j)^\top (b_i - b_j)$$

6. La calidad de representación es mejor para las filas que para las columnas.

- **Factorización Biplot cuando el objetivo es obtener una representación conjunta óptima de los individuos y variables.**

En este caso se propone la siguiente factorización Biplot:

$$\text{HJ (Galindo, 1986): } \mathbf{A}_q \approx (\mathbf{U}\mathbf{\Lambda})_q \text{ y } \mathbf{B}_q \approx (\mathbf{V}\mathbf{\Lambda})_q$$

#### **Características de los marcadores obtenidos con el HJ Biplot.**

En esta factorización, los marcadores fila tienen las mismas propiedades que en el caso del JK Biplot pero además presenta otras propiedades respecto a la representación de las columnas y a la representación simultánea de filas y columnas.

1. El producto escalar de filas y columnas de la matriz  $\tilde{\mathbf{X}}$  coincide con el producto escalar de marcadores fila y marcadores columna en el espacio completo, respectivamente. La aproximación de dichos productos escalares es óptima en el sentido de los mínimos cuadrados.

$$\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top = \mathbf{A}\mathbf{A}^\top \text{ y } \tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} = \mathbf{B}\mathbf{B}^\top$$

2. La longitud al cuadrado de los vectores que representan los marcadores columna aproximan la covarianza entre las respectivas variables.

$$b_i^\top b_j = Cov(\tilde{x}_i, \tilde{x}_j)$$

3. La longitud al cuadrado del vector que representa un marcador columna aproxima la varianza de la correspondiente variable así como la longitud aproxima su desviación estándar.

$$\|b_j\| = \|\tilde{x}_j\| = \sqrt{Var(\tilde{x}_j)}$$

4. El coseno del ángulo que forman los vectores que representan a dos marcadores columna aproxima la correlación entre las respectivas variables.

$$\cos(b_i, b_j) = Corr(\tilde{x}_i, \tilde{x}_j)$$

5. La distancia Euclídea entre dos vectores fila ó columna coincide con la distancia entre los correspondientes marcadores fila ó columna.

$$d^2(\tilde{x}_i^t, \tilde{x}_j^t) = d^2(a_i^t, a_j^t)$$

$$d^2(\tilde{x}_i, \tilde{x}_j) = d^2(b_i, b_j)$$

6. Las coordenadas de los marcadores columna son equivalentes a la importancia de las variables a lo largo de los ejes principales.

$$\tilde{\mathbf{X}}\mathbf{V} = \mathbf{A} \text{ y } \tilde{\mathbf{X}}^\top \mathbf{U} = \mathbf{B}$$

7. Los marcadores fila y columna se pueden representar en el mismo sistema de referencia, con óptima calidad de representación.

### Interpretación geométrica de las representaciones Biplot

En las representaciones Biplot, las filas de la matriz de marcadores fila y las columnas de la matriz de marcadores columna son las coordenadas de los puntos en el espacio de representación común. Los marcadores columna  $b_j$  se representan como vectores y los marcadores fila  $a_i^t$  se representan como puntos.

Las relaciones individuo-variable se estudian a través de las proyecciones de los puntos que representan individuos sobre los vectores que representan variables. Es decir,

$$\tilde{x}_{ij}^t \approx (a_i^t)^\top b_j \Rightarrow \tilde{x}_{ij}^t \approx \|\text{proy } a_i^t/b_j\| \text{signo } b_j \|b_j\|$$

La dirección de los vectores columna  $b_j$ , representa la dirección en la que aumentan los valores de la variable a la que representa, y las proyecciones de todos los puntos fila  $a_i^t$  sobre un vector columna, reproducen aproximadamente los elementos de la columna  $j$ -ésima en la matriz original, permitiendo al mismo tiempo una ordenación aproximada de los individuos (filas) respecto a dicha variable.

Una vez explicada la forma de representación de marcadores fila y columna

podemos decir que:

1. La distancia entre individuos se interpreta como disimilaridades entre los mismos de tal forma que una distancia menor entre marcadores fila en el espacio de representación común indica una menor disimilaridad entre los individuos, especialmente si están bien representados.
2. Las longitudes y los ángulos de los vectores que representan a las variables se interpretan como variabilidad y covariabilidad respectivamente.
3. Las relaciones entre individuos y variables se interpretan a través de las proyecciones de los puntos que representan a los individuos sobre los vectores que representan a las variables.
4. La ordenación de los individuos respecto a una variable determinada se interpreta a través del orden en el que se proyectan los marcadores fila sobre el marcador columna correspondiente a dicha variable.

#### 4.3.4. Cuarta Etapa: Medidas de la calidad de representación Biplot para $\tilde{X}$

##### Contribuciones

Las contribuciones se calculan con la finalidad de conocer qué variables están más directamente relacionadas con cada eje y por tanto qué variables contribuyen a la colocación de los individuos sobre el espacio de representación común.

Como los ejes, por construcción, son independientes, se pueden calcular las contribuciones de hiperplanos sin más que sumar las contribuciones de los ejes que lo forman.

En este apartado se detallan las contribuciones que se calculan en ambas alternativas Biplot.

La suma de cuadrados tanto de los marcadores fila como de los marcadores columna en cada eje es igual al cuadrado del valor propio correspondiente de la matriz de productos escalares. Entonces, se puede considerar la suma de cuadrados de las coordenadas como la contribución absoluta a la variabilidad total.

$$\sum_{i=1}^I (a_{iq}^t)^2 = \lambda_q^2 \text{ y } \sum_{j=1}^J (b_{jq})^2 = \lambda_q^2$$

Por lo tanto, se puede definir:

- Contribución relativa a la variabilidad total del elemento fila  $i^t$ :  $CRT_i = \frac{\sum_{q=1}^R (a_{iq}^t)^2}{\sum_{q=1}^R \lambda_q^2}$
- Contribución relativa a la variabilidad total del elemento columna  $j$ :  $CRT_j = \frac{\sum_{q=1}^R (b_{jq})^2}{\sum_{q=1}^R \lambda_q^2}$
- Contribución relativa a la variabilidad total del grupo  $t$ :  $CRT_t = \frac{\sum_{q=1}^R \sum_{i=1}^{I_t} (a_{iq}^t)^2}{\sum_{q=1}^R \lambda_q^2}$
- Contribución Relativa del Elemento fila  $i^t$  al  $q$ -ésimo Factor:  $CRE_i F_q = \frac{(a_{iq}^t)^2}{\lambda_q^2}$
- Contribución Relativa del Elemento columna  $j$  al  $q$ -ésimo Factor:  $CRE_j F_q = \frac{(b_{jq})^2}{\lambda_q^2}$

También es posible considerar la Contribución Relativa del Grupo  $t$ -ésimo al Factor  $CRG_t F_q = \frac{\sum_{i=1}^{I_t} (a_{iq}^t)^2}{\lambda_q^2}$

- Contribución relativa del factor  $q$ -ésimo al elemento fila  $i_t$ :  $CRF_q E_i = \frac{(a_{iq}^t)^2}{\sum_{q=1}^R (a_{iq}^t)^2}$
- Contribución relativa del factor  $q$ -ésimo al elemento columna  $j$ :  $CRF_q E_j = \frac{(b_{jq})^2}{\sum_{q=1}^R (b_{jq})^2}$

- Contribución relativa del factor  $q$ -ésimo al grupo  $t$   $CRF_q G_t = \frac{\sum_{i=1}^{I_t} (a_{iq}^t)^2}{\sum_{q=1}^R \sum_{i=1}^{I_t} (a_{iq}^t)^2}$

### Medidas de Bondad de Ajuste para $\tilde{\mathbf{X}}$

La Bondad del Ajuste para  $\tilde{\mathbf{X}}$ , sólo se calcula si utilizamos el JK Biplot ya que el HJ Biplot no tiene como objetivo la aproximación de la matriz sino lograr una óptima calidad de representación tanto para las filas como para las columnas de la matriz original.

Así, se determina la calidad de la aproximación o Bondad del Ajuste, de la matriz  $\tilde{\mathbf{X}}$  por la matriz  $\tilde{\mathbf{X}}_q$  en el JK Biplot.

La variabilidad total de la matriz  $\tilde{\mathbf{X}}$  se calcula como la suma de sus elementos al cuadrado, que es igual a la traza de  $\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top$ .

$$\text{traza}(\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top) = \sum_{r=1}^R \lambda_r^2$$

Del mismo modo, la variabilidad total de la matriz  $\tilde{\mathbf{X}}_q$  que representa la variabilidad explicada por la representación Biplot viene dada por

$$\sum_{r=1}^q \lambda_r^2$$

Por tanto, la variabilidad residual se calcula:

$$\tilde{\mathbf{X}} - \tilde{\mathbf{X}}_q = \sum_{r=q+1}^R \lambda_r^2$$

Es decir,

$$SCTotal = SCExplicada + SCResidual \left( \sum_{r=1}^R \lambda_r^2 = \sum_{r=1}^q \lambda_r^2 + \sum_{r=q+1}^R \lambda_r^2 \right)$$



Luego:

Una medida de la calidad de la aproximación viene dada por el porcentaje de la suma de cuadrados de  $\tilde{\mathbf{X}}$  que se consigue explicar con la aproximación Biplot:

$$\frac{\sum_{r=1}^q \lambda_r^2}{\sum_{r=1}^R \lambda_r^2}$$

En ambas representaciones Biplot es posible calcular la calidad de aproximación de las filas y las columnas.

La calidad de aproximación para las filas de la matriz  $\tilde{\mathbf{X}}$  por los marcadores  $\mathbf{A}_q$  viene dada por:

$$\frac{\sum_{r=1}^q \lambda_r^4}{\sum_{r=1}^R \lambda_r^4}$$

La calidad de aproximación para las columnas de la matriz  $\tilde{\mathbf{X}}$  por los marcadores  $\mathbf{B}_q$  difiere de acuerdo al Biplot utilizado:

En el JK Biplot es:

$$\frac{q}{R}$$

En el HJ Biplot:

$$\frac{\sum_{r=1}^q \lambda_r^4}{\sum_{r=1}^R \lambda_r^4}$$

#### 4.3.5. **Bondad del Ajuste para la matriz $\mathbf{X}$ .**

En la sección 4.3.4 se calcula la calidad de representación de la matriz  $\tilde{\mathbf{X}}$ , pero el objetivo es la matriz  $\mathbf{X}$ , es decir, se quiere medir qué porcentaje de variación total de la matriz  $\mathbf{X}$ , es explicado por dicho método.

En primer lugar se calcula la estimación de  $\mathbf{X}$ .

Se ha calculado  $\tilde{\mathbf{X}}$  como:

$$\tilde{\mathbf{X}} = \mathbf{N}_\lambda \mathbf{X}$$

Si se multiplican ambos términos por  $\mathbf{N}_\lambda^{-1}$ , se obtiene:

$$\mathbf{N}_\lambda^{-1}\mathbf{N}_\lambda\mathbf{X} = \mathbf{X} = \mathbf{N}_\lambda^{-1}\tilde{\mathbf{X}}$$

Como la estimación de  $\tilde{\mathbf{X}}$  viene dada por:

$$\hat{\tilde{\mathbf{X}}} \approx \mathbf{A}_q\mathbf{B}_q^\top$$

Se obtiene que:

$$\hat{\mathbf{X}} = \mathbf{N}_\lambda^{-1}\hat{\tilde{\mathbf{X}}}$$

Mediante la suma de cuadrados explicada por la estimación se puede calcular el porcentaje de variación total explicada por el método.

$$SCT = \text{traza}(\mathbf{X}^\top\mathbf{X}) \text{ y}$$

$$SCExplicada = \text{traza}(\hat{\tilde{\mathbf{X}}}^\top\hat{\tilde{\mathbf{X}}}).$$

Luego:

$$\text{Porcentaje de Variación Explicada} = \frac{SCExplicada}{SCT}.$$

## 4.4. Varios conjuntos de variables observados sobre un único conjunto de individuos

Sea  $\mathbf{X}$  la matriz que contiene la información de las  $T$  tablas originales que se quieren analizar, concatenadas por filas (4.2).

$$\mathbf{X} = \begin{array}{c} \\ \\ \\ I \end{array} \begin{array}{c} \text{Grupo 1} \\ 1 \dots J_1 \\ \left( \begin{array}{c} \\ \mathbf{X}_1 \\ \end{array} \right) \end{array} \dots \begin{array}{c} \text{Grupo } t \\ 1 \dots J_t \\ \left( \begin{array}{c} \\ \mathbf{X}_t \\ \end{array} \right) \end{array} \dots \begin{array}{c} \text{Grupo } T \\ 1 \dots J_T \\ \left( \begin{array}{c} \\ \mathbf{X}_T \\ \end{array} \right) \end{array} \quad (4.2)$$

Siendo:

$\mathbf{X}$ : tabla completa de orden  $(I \times J)$ , de término general  $x_{ij}$ , con  $i = 1, \dots, I$ ;  $j = 1, \dots, J$ .

$\mathbf{X}_t$ :  $t$ -ésima tabla, de término general  $x_{ij}^t$ , con  $i = 1, \dots, I$ ;  $j = 1, \dots, J_t$ . Es decir:  $x_{ij}^t$  valor de la  $j$ -ésima variable para el  $i$ -ésimo individuo de la tabla  $\mathbf{X}_t$ .

$I$ : conjunto de individuos.

$J$ : conjunto de variables o el número total de variables.

$J_t$ : conjunto de variables de la  $t$ -ésima tabla o el  $t$ -ésimo grupo de variables.

$T$ : número total de tablas o el conjunto de grupos de variables. Es decir:

$$\sum_{t=1}^T J_t = J$$

$J, J_t, T, I$  son cardinales.

En este caso no se realiza la estandarización de la matriz  $\mathbf{X}$  debido a la forma en que las tablas están yuxtapuestas, por tanto el análisis consta de cuatro etapas, que se detallan a continuación.

#### 4.4.1. Primera Etapa: Análisis Individuales

Se realiza un PCA de cada una de las  $T$  tablas. En cada uno se selecciona el primer valor propio. Denominaremos con  $\lambda_1^t$  al primer valor propio de la  $t$ -ésima tabla. Los valores de cada una de las  $T$  tablas se ponderan por la inversa de su primer valor propio; es decir, los valores de la  $t$ -ésima tabla están ponderados por  $\frac{1}{\lambda_1^t}$ , para todo  $t = 1, \dots, T$ . Esta ponderación es la misma que la que utiliza el análisis Factorial Múltiple.

Obtenemos así una matriz, que denominaremos  $\tilde{\mathbf{X}}$  de orden  $(IxJ)$ .

$$\tilde{\mathbf{X}} = \left( \begin{array}{ccc} \frac{1}{\lambda_1^1} \mathbf{X}_1 & \dots & \frac{1}{\lambda_1^T} \mathbf{X}_T \end{array} \right)$$

Por tanto:  $\tilde{\mathbf{X}} = \mathbf{X} \mathbf{M}_\lambda$

$\mathbf{M}_\lambda$  es de orden  $(J \times J)$  y su  $t$ -ésimo elemento es una matriz diagonal de orden  $(J_t \times J_t)$ .

$$\tilde{\mathbf{X}} = \mathbf{X} \left[ \begin{array}{ccc} \left( \begin{array}{ccc} \frac{1}{\lambda_1^1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \frac{1}{\lambda_1^1} \end{array} \right) & \dots & \left( \begin{array}{ccc} 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{array} \right) & \dots & \left( \begin{array}{ccc} 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{array} \right) \\ \vdots & \ddots & \vdots & & \vdots \\ \left( \begin{array}{ccc} 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{array} \right) & \dots & \left( \begin{array}{ccc} \frac{1}{\lambda_1^t} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \frac{1}{\lambda_1^t} \end{array} \right) & \dots & \left( \begin{array}{ccc} 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{array} \right) \\ \vdots & & \vdots & \ddots & \vdots \\ \left( \begin{array}{ccc} 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{array} \right) & \dots & \left( \begin{array}{ccc} 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{array} \right) & \dots & \left( \begin{array}{ccc} \frac{1}{\lambda_1^T} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \frac{1}{\lambda_1^T} \end{array} \right) \end{array} \right]$$

#### 4.4.2. Segunda Etapa: Biplot ponderado de la matriz $\tilde{\mathbf{X}}$

En esta etapa, se realiza un Biplot ponderado de  $\tilde{\mathbf{X}}$  y se proponen dos factorizaciones Biplot:

- cuando el objetivo es aproximar los valores de  $\mathbf{X}$

$$\text{JK (Gabriel, 1971): } \mathbf{A}_q \approx (\mathbf{U}\mathbf{\Lambda})_q \text{ y } \mathbf{B}_q \approx (\mathbf{V})_q$$

- cuando el objetivo es obtener una representación conjunta óptima de individuos y variables.

$$\text{HJ (Galindo, 1986): } \mathbf{A}_q \approx (\mathbf{U}\mathbf{\Lambda})_q \text{ y } \mathbf{B}_q \approx (\mathbf{V}\mathbf{\Lambda})_q$$

En ambos Biplot el subíndice  $q$  indica las  $q$  primeras columnas de las matrices, siendo  $q \leq R$ , si  $R$  es el rango de la matriz  $\tilde{\mathbf{X}}$ .

$\mathbf{A}_{(I \times q)}$ : matriz que contiene los  $q$  marcadores para las filas.

$a_i$ : marcador para el  $i$ -ésimo individuo de la matriz  $\tilde{\mathbf{X}}$ .

$\mathbf{B}_{(J \times q)}$ : matriz que contiene los  $q$  marcadores para las columnas.

$b_j^t$ : marcador para la  $j^t$ -ésima variable de la matriz  $\tilde{\mathbf{X}}$ .

Las propiedades y las interpretaciones geométricas en ambos Biplot son las mismas explicadas en la sección 4.3.

#### 4.4.3. Tercera Etapa: Medidas de la calidad de la representación Biplot para matriz $\tilde{\mathbf{X}}$

Las contribuciones se interpretan del mismo modo que en la sección 4.3.4.

- Contribución relativa a la variabilidad total del elemento fila  $i$ :  $CRT_i =$

$$\frac{\sum_{q=1}^R a_{iq}^2}{\sum_{q=1}^R \lambda_q^2}$$

- Contribución relativa a la variabilidad total del elemento columna  $j^t$ :

$$CRT_{j^t} = \frac{\sum_{q=1}^R (b_{iq}^t)^2}{\sum_{q=1}^R \lambda_q^2}$$

- Contribución Relativa del Elemento fila  $i$  al Factor  $q$ :  $CRE_i F_q = \frac{a_{iq}^2}{\lambda_q^2}$
- Contribución Relativa del Elemento columna  $j^t$  al Factor  $q$ :  $CRE_{j^t} F_q = \frac{(b_{jq}^t)^2}{\lambda_q^2}$
- Contribución Relativa del Factor  $q$  al Elemento fila  $i$ :  $CRF_q E_i = \frac{a_{iq}^2}{\sum_{q=1}^R a_{iq}^2}$
- Contribución Relativa del Factor  $q$  al Elemento columna  $j^t$ :  $CRF_q E_{j^t} = \frac{(b_{jq}^t)^2}{\sum_{q=1}^R (b_{jq}^t)^2}$
- Contribución Relativa a la variabilidad Total del grupo  $t$ :  $CRT_t = \frac{\sum_{q=1}^R \sum_{j=1}^{J_t} (b_{jq}^t)^2}{\sum_{q=1}^R \lambda_q^2}$
- Contribución Relativa del Grupo  $t$  al  $q$ -ésimo Factor:  $CRG_t F_q = \frac{\sum_{j=1}^{J_t} (b_{jq}^t)^2}{\lambda_q^2}$   
representa la parte de variabilidad del  $q$ -ésimo factor explicada por el grupo  $t$ .
- Contribución Relativa del Factor  $q$  al  $t$ -ésimo Grupo.  $CRF_q G_t = \frac{\sum_{j=1}^{J_t} (b_{jq}^t)^2}{\sum_{q=1}^R \sum_{j=1}^{J_t} (b_{jq}^t)^2}$   
mide la parte de la variabilidad del grupo  $t$  explicada por el  $q$ -ésimo factor.

Respecto a las medidas de Bondad de Ajuste son iguales a las que se han desarrollado para el caso anterior, con las mismas consideraciones, es decir:

- La Bondad del Ajuste sólo se calcula para el JK Biplot.
- La calidad de aproximación para las filas y las columnas es óptima en el HJ Biplot.

#### 4.4.4. Cuarta Etapa: Bondad del Ajuste para la matriz $\mathbf{X}$

La Bondad del Ajuste para  $\mathbf{X}$ , sólo se calcula si utilizamos el JK Biplot.

En primer lugar, se calcula la aproximación de  $\mathbf{X}$ .

Se ha calculado  $\tilde{\mathbf{X}}$  como:

$$\tilde{\mathbf{X}} = \mathbf{X}\mathbf{M}_\lambda$$

Si se multiplican ambos términos por  $\mathbf{M}_\lambda^{-1}$ , se obtiene:

$$\mathbf{X}\mathbf{M}_\lambda\mathbf{M}_\lambda^{-1} = \mathbf{X} = \tilde{\mathbf{X}}\mathbf{M}_\lambda^{-1}$$

Como la estimación de  $\tilde{\mathbf{X}}$  viene dada por:

$$\hat{\tilde{\mathbf{X}}} = \mathbf{A}_q\mathbf{B}_q^\top$$

Se obtiene que:

$$\hat{\mathbf{X}} = \hat{\tilde{\mathbf{X}}}\mathbf{M}_\lambda^{-1}$$

Mediante la suma de cuadrados explicada por la estimación se puede calcular el porcentaje de variación total explicada por el método.

$$\text{Porcentaje de Variación Explicada} = \frac{SCExplicada}{SCT}$$

Siendo:

$$SCT = \text{traza}(\mathbf{X}^\top\mathbf{X}) \text{ y}$$

$$SCExplicada = \text{traza}(\hat{\mathbf{X}}^\top\hat{\mathbf{X}}).$$

## 4.5. Bootstrap sobre datos de tres vías

Siguiendo con el desarrollo de versiones inferenciales para técnicas multivariantes, se ha realizado una revisión de las distintas aplicaciones de los métodos Bootstrap sobre técnicas que analizan datos de tres vías. Kiers, 2004 propone el cálculo de intervalos de confianza para los métodos CANDECOM/PARAFAC y TUCKER3 (Tucker, 1966). El remuestreo que utiliza para dicho cálculo se basa en la descomposición de los cubos de datos en capas que contienen la información de cada individuo. De esta forma, remuestra matrices completas que contienen la información de cada individuo en las otras dos vías estudiadas. Abdi et al., 2013; Husson et al., 2005; Pagès y Husson, 2005 utilizan un tipo de remuestreo que consiste en considerar como unidades de remuestreo las tablas que conforman los datos que se van a analizar. Otros autores como Abdi et al., 2009 utilizan otro tipo de remuestreo que denominan *remuestreo dividido a la mitad* ya que consideran que las observaciones de los datos que analizan no siempre son independientes y pueden tener una correlación temporal. Dicho remuestreo consiste en dividir cada una de las matrices en dos partes que llaman *de entrenamiento* y *de test*. En cada una de ellas realizan el remuestreo bootstrap como se ha explicado en el caso de datos de dos vías. Si se considera  $B$  el número de submuestras extraídas y  $T$  el número de matrices de las que se dispone, este tipo de remuestreo proporciona  $2 \times T \times B$  matrices. Esta estrategia proporciona la independencia de las matrices necesaria para la utilización de los métodos Bootstrap. Finalmente se remuestran las matrices obtenidas en el paso anterior y se obtienen los parámetros de interés. Dehlholm et al., 2012 presentan una función que forma parte del paquete **FactoMineR**. El planteamiento es remuestrear las filas dentro de cada matriz pero partiendo de las coordenadas resultantes del MFA en lugar de los datos originales. Con las filas remuestreadas calculan los centroides para



cada matriz. Cadoret y Husson, 2013 proponen un método para construir elipses de confianza que denominan *bootstrap total truncado*, implementado en el paquete **SensoMineR** (Le y Husson, 2008). El esquema que siguen es contruir las muestras bootstrap, aplicar el método y luego utilizar rotaciones Procrustes en un número reducido de dimensiones.

Analizando todas las alternativas anteriormente descritas, se propone a continuación una versión inferencial de los métodos Biplot Múltiples.

Dado que se dispone de dos maneras de analizar los datos según el modo que compartan las matrices que se van a analizar, se proponen dos estrategias diferentes para cada uno de los análisis.

#### 4.5.1. **Bootstrap sobre Biplot Múltiple**

##### **Varios conjuntos de individuos sobre los que se ha medido un mismo conjunto de variables**

Para este análisis se propone una nueva estrategia para realizar remuestreo bootstrap combinando las diferentes propuestas explicadas anteriormente. Se plantea un remuestreo *dobles* que consta de dos pasos:

- En primer lugar, se realiza un remuestreo bootstrap sobre los índices de las matrices.
- En segundo lugar, se realiza un remuestreo bootstrap sobre los individuos que componen cada una de las matrices que formen parte de la muestra bootstrap seleccionada en el paso anterior.

De esta forma se consiguen submuestras del mismo tamaño que el de partida y los individuos son independientes entre sí al hacer un remuestreo diferente para cada una de las matrices que formen parte del conjunto de datos.

## Varios conjuntos de variables medidos sobre un mismo conjunto de individuos

En este análisis, se propone como estrategia de remuestreo la descrita en Abdi et al., 2013; Husson et al., 2005 que utilizan como unidades de remuestreo las tablas que contienen información de cada individuo. En nuestro caso, al no tener un cubo de datos, si no que solamente se comparte el modo de los individuos, se considera como unidad de remuestreo la fila correspondiente a cada individuo que contiene las mediciones de todas las variables de los conjuntos considerados sobre dicho individuo. Esta estrategia es semejante a la que se utilizó en el capítulo 2 para proporcionar una versión inferencial de los métodos Biplot.

### 4.6. Programa *MultibiplotGUI*

Debido a que los Biplot Múltiples no disponen de un software para su utilización, se ha desarrollado en el entorno R un nuevo programa en forma de interfaz gráfica (GUI) para permitir su ejecución de una manera interactiva a través de ventanas, menús y botones. Además se ha añadido la opción de proporcionar los resultados de una forma inferencial tal como se ha explicado en la sección anterior. Este paquete se llama `multibiplotGUI` y está disponible para su descarga en <http://cran.r-project.org/web/packages/multibiplotGUI>.

En primer lugar, es necesario bajar R de la web [cran.r-project.org](http://cran.r-project.org) e instalarlo. A continuación se descarga el paquete `multibiplotGUI` y sus dependencias que son los paquetes: `rgl`, `tcltk`, `tcltk2`, `tkrplot`, `shapes`, `cluster` y `dendroextras`; (Adler y Murdoch, 2012; Dryden, 2014; Grosjean, 2012; Jefferis, 2014; Maechler et al., 2015; Tierney, 2012).

Para poder utilizar el paquete `multibiplotGUI` en el software R, se debe introducir el comando `library(multibiplotGUI)` en la ventana principal de R.

A continuación es necesario cargar los datos que se quieren analizar. Si el modo que comparten las distintas matrices son las variables, hay que introducir una matriz en la que se hayan concatenado por filas todas las matrices disponibles; si el modo compartido son los individuos, la matriz que se introduce para analizar es la que resulta de concatenar las matrices por columnas. También es necesario disponer de un vector en el que se informe de las dimensiones del modo que no comparten las matrices.

Si suponemos que la matriz concatenada en la forma adecuada es **data** y el vector que contiene las dimensiones no compartidas se denomina **ni**, la interfaz se inicializa mediante el comando `multibiplot(data,ni)`.

A continuación, se abre la ventana principal. Dicha ventana (figura 4.1) consta de dos radiobuttons que permiten al usuario seleccionar el tipo de datos que quiere analizar, es decir:

- Varios conjuntos de individuos sobre un mismo conjunto de variables.
- Varios conjuntos de variables sobre un único conjunto de individuos.



Figura 4.1: Ventana para seleccionar el tipo de datos.

Una vez que se ha elegido el tipo de datos, aparece la ventana de opciones (figura 4.2). En ella se permite:

- Elegir el tipo de Biplot que se quiere realizar (JK o HJ).

- Cambiar color, tamaño, etiqueta y símbolo de los individuos en los gráficos.
- Cambiar color, tamaño y etiqueta de las variables en los gráficos.
- Mostrar los ejes de coordenadas en los gráficos.
- Cambiar el tamaño de las ventanas que se presentan a posteriori para que el programa sea adaptable a los diferentes tamaños de las pantallas.

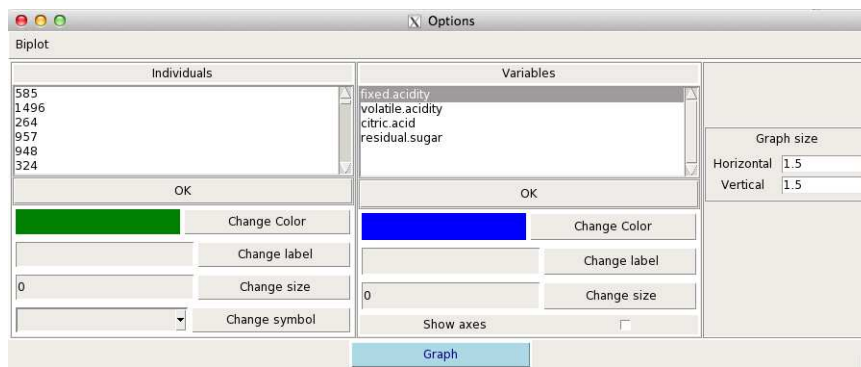


Figura 4.2: Ventana de opciones.

Una vez que el usuario ha configurado las opciones gráficas y pulsa el botón **Graph** aparece una nueva ventana (figura 4.3) en la que se muestra un diagrama de barras con la inercia absorbida por cada eje con el objetivo de que se pueda elegir el número de ejes deseado.

Una vez elegidos el número de ejes y pulsado el botón **Choose**, aparece la ventana (figura 4.4) que contiene el gráfico con los resultados en las dos primeras dimensiones.

En esta ventana se pueden distinguir dos bloques y seis menús. En el bloque de la derecha se tiene la representación gráfica conjunta de las coordenadas biplot para individuos y variables. En este gráfico se pueden mover las etiquetas de los puntos con el botón izquierdo del ratón y se pueden cambiar las características gráficas de los puntos con el botón derecho del ratón. En el bloque de la izquierda

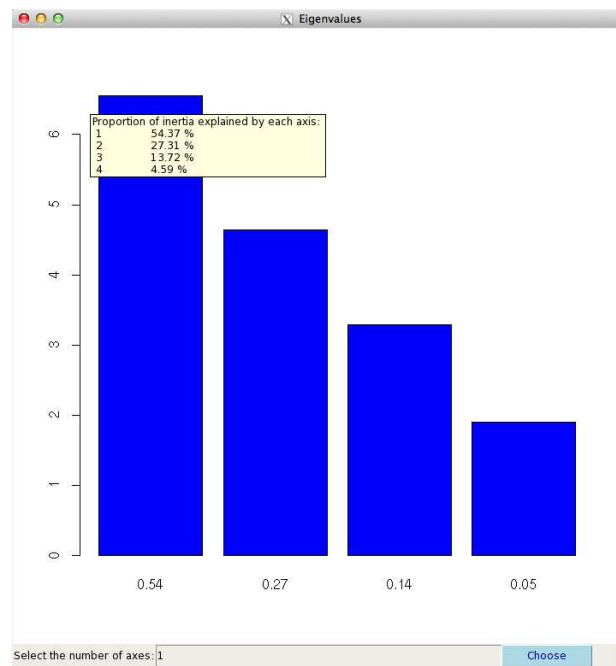


Figura 4.3: Ventana con el gráfico de barras representando la inercia absorbida por cada eje.

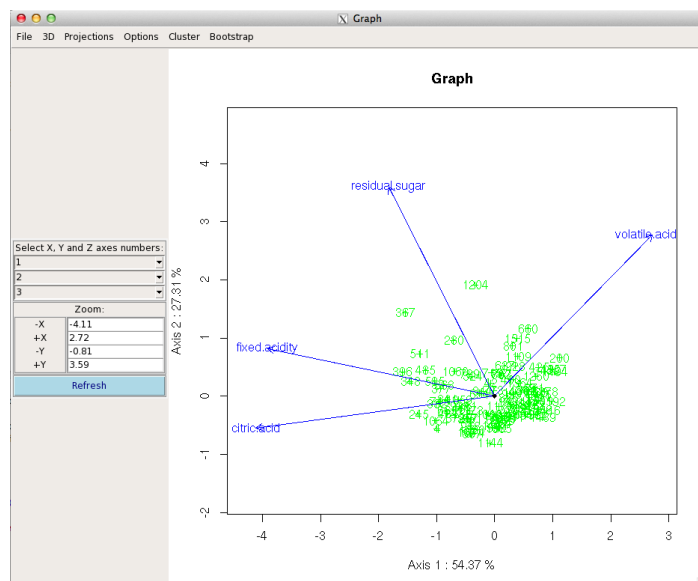


Figura 4.4: Ventana que muestra la representación Biplot en dos dimensiones.

se dispone de tres listbox para elegir la dimensión que se quiere ver en cada gráfico (2 y 3 dimensiones); cuatro textbox que permiten elegir los límites de los ejes  $x$  e  $y$  para poder aumentar o disminuir una zona determinada y un botón **Refresh** que

es necesario pulsar para que se haga efectivo el cambio de los anteriores elementos.

Junto con la ventana aparece un archivo de texto donde se ha guardado toda la información proporcionada por el software: bondad de ajuste, calidad de representación de filas y columnas, contribuciones de los individuos, de las variables y de la matrices a la variabilidad total y a la conformación de los ejes factoriales y viceversa. En la parte superior de la ventana se dispone de seis menús con sus correspondientes submenús:

- File
  - Copy image
  - Save image
    - PDF file
    - Eps file
    - Png file
    - Jpg/Jpeg file
  - Exit
- 3D
  - 3D
- Projections
  - Variables
  - Back to original data
- Options
  - Change title

- Show/Hide axes
- Cluster
  - Hierarchical cluster with biplot coordinates
  - K-means with biplot coordinates
  - K-medoids with biplot coordinates
  - Back to original graph
- Bootstrap
  - Bootstrap

Los cinco primeros son los mismos que se explicaron en el capítulo 2. El primero de ellos sirve para copiar la imagen al portapapeles, guardar en varios tipos de formato el gráfico y salir del programa. El siguiente es el menú de 3 dimensiones. Mediante este menú es posible visualizar el gráfico en tres dimensiones. En este gráfico se puede rotar la imagen con el botón izquierdo del ratón y ampliar o disminuir la imagen con el botón derecho del ratón. El tercer menú que se encuentra el usuario en esta ventana sirve para proyectar los puntos que representan a los individuos sobre la dirección de una variable que se puede seleccionar mediante un listado. Si se desea volver al gráfico original se elige el submenú `Back to original data`. El siguiente menú disponible es el de opciones del gráfico. Contiene dos submenús: cambiar el título del gráfico y mostrar u ocultar los ejes de coordenadas. El quinto menú permite analizar las coordenadas biplot mediante técnicas de Cluster. Los métodos de los que se dispone son:

- Cluster Jerárquico. Se dispone de una ventana para elegir el número de clusters, la distancia que se quiere utilizar (Euclidean, Maximum, Manhat-

tan, Canberra, Binary, Minkowski) y el método de agrupación (Ward.D, Ward.D2, Single, Complete, Average, Mcquitty, Median, Centroid). Se obtiene como resultado un dendrograma de la agrupación de los individuos y se muestra sobre el gráfico en dos dimensiones en distintos colores y encerrados en una línea poligonal cada uno de los clusters obtenidos.

- Cluster de k-medias. Se proporciona una ventana para elegir el número de clusters, el número máximo de iteraciones que se va a permitir realizar hasta llegar a la solución óptima, el número de conjuntos aleatorios de centroides que se van a utilizar como solución inicial en el algoritmo y el algoritmo que se va a aplicar para la búsqueda de la solución (Hartigan-Wong, Lloyd, Forgy, MacQueen). Este método muestra sobre el gráfico en dos dimensiones en distintos colores y encerrados en una línea poligonal tanto los clusters obtenidos como los centroides de cada uno de ellos.
- Cluster de K-medoides. En este método aparece una ventana para elegir el número de clusters deseado por el usuario y la distancia que se quiere utilizar (Euclidean, Maximum, Manhattan, Canberra, Binary, Minkowski).

En este menú también se encuentra disponible la opción de volver al gráfico original.

El último menú que se encuentra disponible es que nos va a permitir tener medidas de precisión de los resultados proporcionados por el Biplot Múltiple. Si se pulsa en este menú aparece una ventana (figura 4.5) donde se pueden elegir los parámetros para los que se quieren medidas de precisión, el número de réplicas bootstrap para generar dichos resultados y el nivel de confianza con los que se van a calcular los intervalos de confianza que se presentan a posteriori. Además de proporcionar los resultados guardados en un archivo de texto, el programa también genera gráficos que contienen los histogramas y los gráficos de normalidad



de cada uno de los parámetros que se hayan elegido. Se guardan en formato .eps y en .pdf y en la ventana es posible elegir si se generan en blanco y negro o en color.

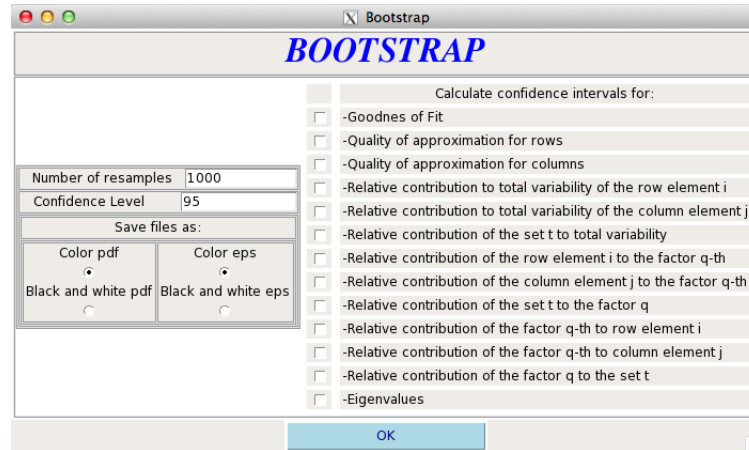


Figura 4.5: Ventana para el análisis Bootstrap.

El programa también genera un gráfico con todas las coordenadas de las variables que se han calculado para todas las muestras bootstrap. Para eliminar el posible efecto espejo que pudiera existir entre las diferentes configuraciones, se han utilizado rotaciones Procrustes como se ha explicado en capítulos anteriores. Cada grupo de coordenadas que representan a la misma variable se muestra en el mismo color y se encuentra encerrado bajo una línea poligonal envolvente (convex hull).

## 4.7. Aplicación a Datos

A continuación se desarrolla un ejemplo con unos datos de prueba para mostrar el funcionamiento del programa y la interpretación de los resultados del Biplot Múltiple tanto en su versión clásica como en la versión inferencial explicada anteriormente.

Los datos que se utilizan son una submuestra de los datos iris que se explicaron

en el ejemplo que se desarrolló para ilustrar la versión inferencial de los Biplot clásicos. Se han elegido 20 individuos de cada una de las variedades de iris (*setosa*, *versicolor* y *virginica*) y se han considerado como tres conjuntos de individuos diferentes sobre los que se han medido las mismas variables.

En primer lugar, se explican los resultados obtenidos para la versión exploratoria del Biplot Múltiple.

Como se ha explicado antes, la situación de los datos corresponde con varios conjuntos de individuos sobre los que se ha medido un mismo conjunto de variables. Por lo tanto, se eligió dicha opción, se seleccionó la factorización HJ Biplot y se retuvieron 3 ejes que suponen una inercia absorbida del 99%. A continuación se muestra la tabla 4.1 con la información relativa a los valores propios. Se observa que el primer eje absorbe la mayor parte de la información y tan sólo una pequeña parte es explicada por los otros dos.

No.	Valor Propio	Variabilidad	Variabilidad Acumulada
1	7.52	78.94	78.94
2	3.45	16.60	95.54
3	1.58	3.46	99.00

Tabla 4.1: Valores propios y variabilidad explicada (%) por cada eje para los datos iris.

Las contribuciones de los individuos a la variabilidad total, las contribuciones relativas de los individuos a los factores y de los factores a los individuos se encuentran en el apéndice A. Según se puede observar, la contribución de cada individuo a la variabilidad total está repartida entre todos ellos sin que ninguno tenga una especial relevancia. En cuanto a la contribución de los individuos a la conformación de los ejes factoriales se puede ver observar que ninguno tiene una contribución alta. Respecto de la contribución de cada factor a la variabilidad de cada individuo se puede apreciar que la mayoría tienen una contribución alta del primer eje excepto los individuos 26, 35, 37, 48, 56 y 59 que tienen mayor contribución del eje 2, el caso 57 que tiene mayor carga factorial para el eje 3

y algunos que tienen repartida su variabilidad entre los tres ejes.

En las tablas 4.2, 4.3 y 4.4 se observa esta información para las variables analizadas.

Variable	Contribución
Sepal.Length	377.73
Sepal.Width	534.48
Petal.Length	38.80
Petal.Width	48.99

Tabla 4.2: Contribuciones relativas de las variables a la variabilidad total.

Las variables que más contribuyen a la variabilidad total son la longitud y la anchura del sépalo mientras que la longitud y anchura del pétalo tienen menos influencia. Si tenemos en cuenta a qué eje están contribuyendo en mayor medida se observa que la anchura del sépalo es la que más influye en la conformación del primer eje, la longitud del sépalo tiene mayor contribución para el segundo eje y la variable que más contribuye al tercer eje es la anchura del pétalo.

Elemento	Eje 1	Eje 2	Eje 3
Sepal.Length	358.07	543.73	30.30
Sepal.Width	590.24	376.84	18.82
Petal.Length	25.51	73.16	177.09
Petal.Width	26.18	6.27	773.79

Tabla 4.3: Contribuciones relativas de las variables a la conformación de los tres primeros ejes.

En cuanto a las contribuciones relativas de los ejes a las variables, se puede ver que la longitud y anchura del sépalo tienen una contribución alta del eje 1, la longitud del pétalo tiene las contribuciones más altas repartidas entre los ejes 1 y 2 y la anchura del pétalo entre los ejes 1 y 3.

Elemento	Eje 1	Eje 2	Eje 3
Sepal.Length	755.88	241.32	2.81
Sepal.Width	880.57	118.20	1.23
Petal.Length	524.25	316.12	159.63
Petal.Width	426.17	21.46	552.36

Tabla 4.4: Contribuciones relativas de los ejes a las variables.

Estas contribuciones son las mismas que se explicaron en la sección 2.5. Pero como se están analizando varios conjuntos (matrices) también se obtiene la información relativa a las matrices analizadas. Por lo tanto, el Biplot Múltiple proporciona la contribución relativa a la variabilidad total de cada matriz, la contribución relativa de cada matriz a la conformación de los ejes factoriales y la contribución relativa de cada eje a la variabilidad de cada matriz. Esta información se recoge en las tablas 4.5, 4.6 y 4.7.

Matriz	Contribución
Setosa	318.94
Versicolor	356.26
Virginica	324.80

Tabla 4.5: Contribuciones relativas de las matrices a la variabilidad total.

Según se observa, las tres matrices contribuyen de una manera parecida a la variabilidad total de los datos. Las que más contribuyen a la conformación del eje 1 son *setosa* y *versicolor* y en menor medida *virginica*. Para los ejes 2 y 3 las que más influyen son las matrices con individuos de las especies *versicolor* y *virginica*.

Matriz	Eje 1	Eje 2	Eje 3
Setosa	361.93	140.25	195.41
Versicolor	353.96	369.80	343.82
Virginica	284.12	489.95	460.77

Tabla 4.6: Contribuciones relativas de las matrices a la conformación de los tres primeros ejes.

Si analizamos las contribuciones de cada eje a la variabilidad total de cada una de las matrices se observa que para las tres especies el primer eje tiene una contribución alta.

Elemento	Eje 1	Eje 2	Eje 3
Setosa	904.86	73.72	21.43
Versicolor	792.23	174.02	33.75
Virginica	697.50	252.89	49.61

Tabla 4.7: Contribuciones relativas de los ejes a las matrices.

Por último se presenta el gráfico en dos dimensiones con la representación conjunta de las tres especies y las variables analizadas (figura 4.6).

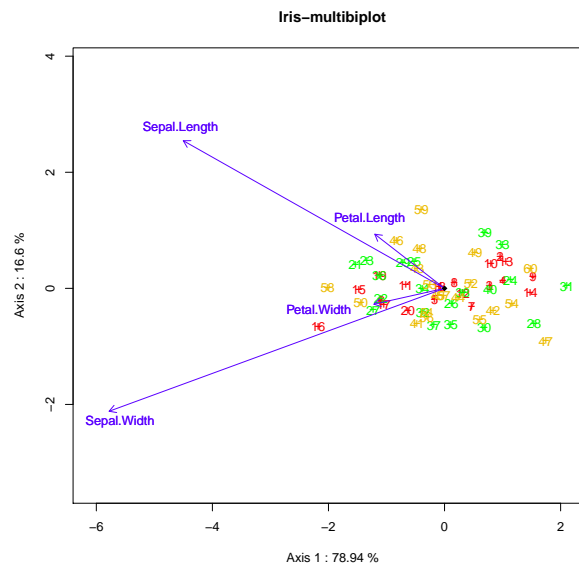


Figura 4.6: Gráfico en las dos primeras dimensiones de las coordenadas resultantes del Biplot Múltiple.

Una vez que se han analizado los datos de una manera exploratoria, el siguiente paso es obtener los resultados inferenciales para las medidas explicadas anteriormente.

En primer lugar, se presentan los resultados para los valores propios. Al igual que en los capítulos anteriores, se presentan en forma de tabla con el valor observado, la media de los valores calculados a partir de las submuestras, el error estándar, el sesgo y los extremos inferior y superior de los intervalos de confianza t-bootstrap, percentiles y BCa. Dicha información está complementada con los gráficos que representan los histogramas de las réplicas bootstrap de cada medida y los gráficos de normalidad. Todos los resultados se han obtenido con 1000 muestras bootstrap a un nivel de confianza del 95 %.

La tabla 4.8 y la figura 4.7 muestran los resultados para los valores propios. Se observa que hay pequeñas diferencias entre los valores observados y las medias

de las réplicas bootstrap sobretodo en el primer valor propio. Respecto a la distribución de las réplicas, se puede ver que el primer histograma tiene más asimetría que el resto y el gráfico de normalidad muestra una cierta desviación de la normalidad. Esto nos indica que el primer valor propio es más inestable que el resto.

	V. Obs	Media	SE	Sesgo	t-b inf	t-b sup	perc inf	perc sup	BCa inf	BCa sup
1	7.52	6.84	1.24	-0.69	4.40	9.27	4.36	8.57	4.59	8.79
2	3.45	3.19	0.82	-0.26	1.58	4.80	1.91	5.09	2.44	6.24
3	1.58	1.61	0.33	0.04	0.96	2.27	1.03	2.37	1.01	2.33
4	0.85	0.84	0.18	-0.01	0.49	1.19	0.55	1.24	0.58	1.36

Tabla 4.8: Resultados bootstrap para valores propios.

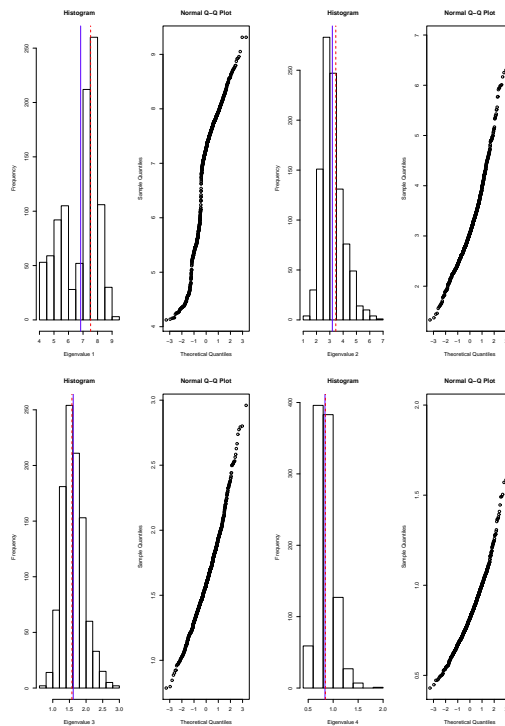


Figura 4.7: Histogramas y gráficos de normalidad para los valores propios.

A continuación se presentan los resultados referentes a la contribución relativa de cada variable a la variabilidad total presente en los datos analizados. En la tabla 4.9 y en la figura 4.8 se observa que la longitud del sépalo presenta diferencias más pequeñas que la anchura del sépalo y la longitud y anchura

del pétalo. Estas dos últimas variables además, presentan distribuciones de las réplicas asimétricas y los gráficos de normalidad indican desviaciones respecto a la distribución normal. Por lo tanto, la contribución relativa de las variables a la variabilidad total no es estable a lo largo del conjunto de muestras bootstrap para estos datos.

	V. Obs	Media	SE	Sesgo	t-b inf	t-b sup	perc inf	perc sup	BCa inf	BCa sup
S.L	377.73	356.99	76.08	-20.74	207.70	506.28	231.97	518.08	252.67	550.59
S.W	534.48	471.46	121.49	-63.02	233.05	709.87	242.72	707.64	365.67	784.75
P.L	38.80	82.22	74.58	43.42	-64.14	228.57	14.73	256.18	13.15	242.86
P.W	48.99	89.33	69.05	40.34	-46.17	224.83	22.21	264.36	14.10	157.11

Tabla 4.9: Resultados bootstrap para las contribuciones relativas de cada variable a la variabilidad total.

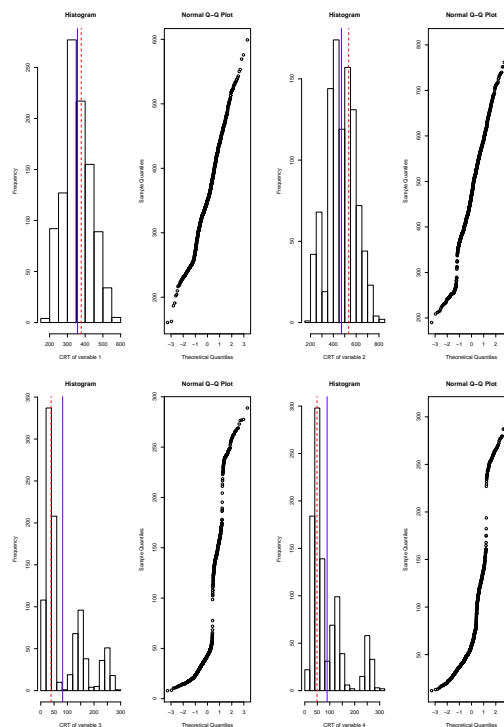


Figura 4.8: Histogramas y gráficos de normalidad para las contribuciones relativas a la variabilidad total de las variables.

Si estudiamos los resultados de la contribución relativa de cada variable a la conformación de los ejes factoriales (tabla 4.10 y figura 4.9), se observan

diferencias entre los valores observados y las medias de las réplicas bootstrap y las amplitudes de los intervalos de confianza no son pequeñas. Respecto a los gráficos, destaca la asimetría y la falta de normalidad de aquellas contribuciones que son pequeñas en la muestra original.

	V. Obs	Media	SE	Sesgo	t-b inf	t-b sup	perc inf	perc sup	BCa inf	BCa sup
S.L										
E.1	358.07	355.91	99.39	-2.16	160.86	550.95	172.37	574.26	197.68	584.44
E.2	543.73	464.27	158.45	-79.46	153.33	775.21	61.14	729.77	294.85	791.91
E.3	30.30	48.11	68.29	17.81	-85.91	182.12	0.14	305.61	0.41	334.81
S.W										
E.1	590.24	508.83	155.26	-81.42	204.16	813.49	209.76	803.08	354.15	873.49
E.2	376.84	365.70	135.24	-11.14	100.31	631.09	76.80	616.23	86.25	623.24
E.3	18.82	86.87	138.38	68.05	-184.68	358.42	0.26	568.22	0.00	222.26
P.L										
E.1	25.51	73.77	86.67	48.26	-96.31	243.85	3.65	295.59	2.76	280.53
E.2	73.16	108.15	143.31	35.00	-173.06	389.37	6.00	782.16	2.94	216.68
E.3	177.09	186.44	134.20	9.35	-76.92	449.79	5.69	482.26	9.26	509.88
P.W										
E.1	26.18	61.50	66.97	35.32	-69.91	192.91	8.65	257.68	5.28	161.69
E.2	6.27	61.88	105.95	55.61	-146.02	269.78	0.07	428.21	0.00	83.82
E.3	773.79	678.59	162.69	-95.21	359.33	997.84	258.30	918.68	533.79	968.99

Tabla 4.10: Resultados bootstrap para las contribuciones relativas de las variables a los ejes factoriales.



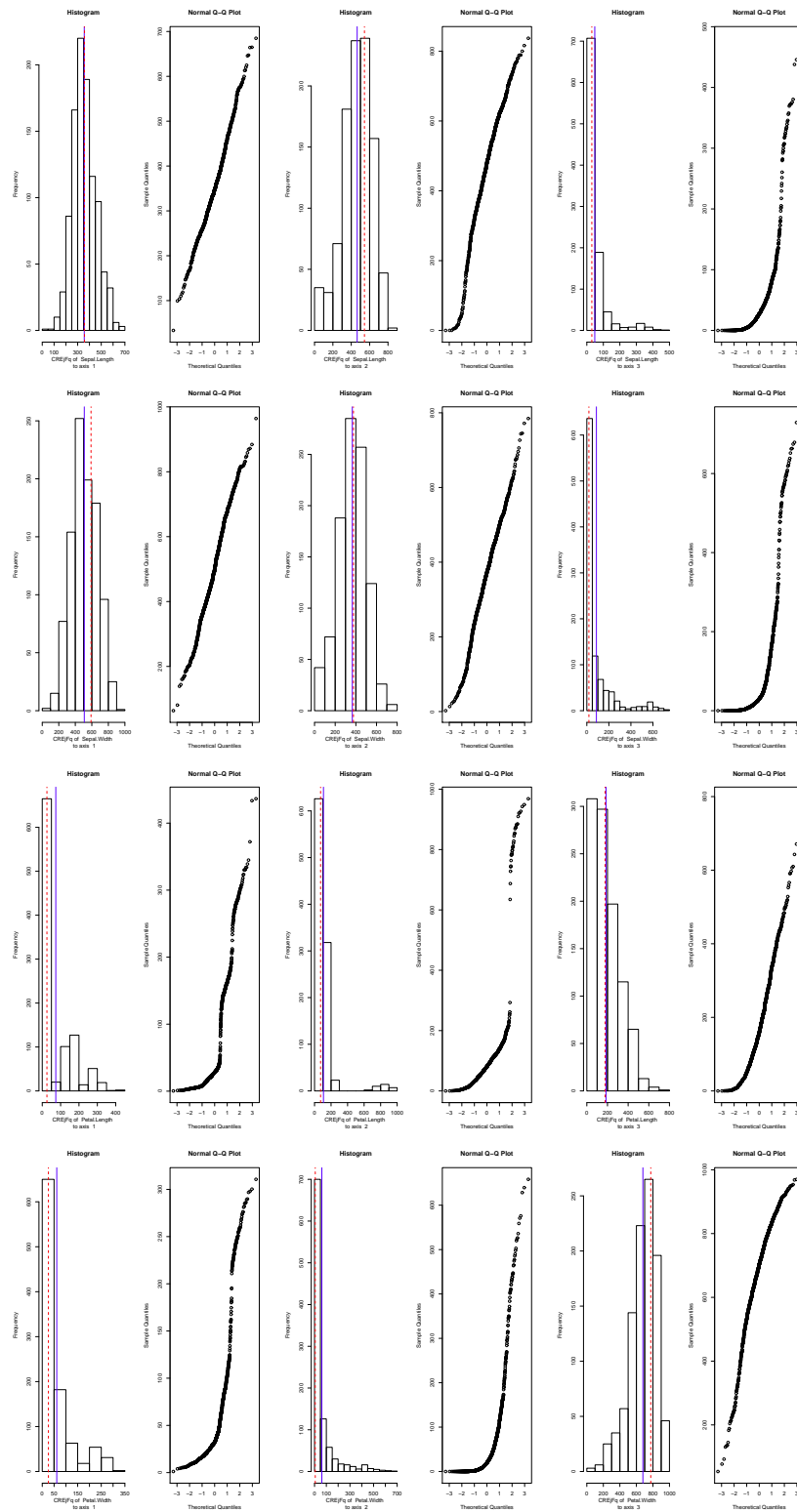


Figura 4.9: Histogramas y gráficos de normalidad para las contribuciones relativas de las variables a la conformación de los ejes factoriales.

A continuación, analizamos los resultados relativos a la contribución relativa de los ejes factoriales a las variables (tabla 4.11 y figura 4.10). Al igual que en las medidas anteriores, se observan algunos sesgos considerables (anchura del pétalo respecto de los ejes 2 y 3). Las contribuciones del eje 1 a las variables longitud y anchura del sépalo presentan gran asimetría. Las relativas al eje 2 son en general asimétricas, siendo más pronunciada la distribución de las réplicas relativas a la variable anchura del pétalo. Las réplicas bootstrap de las contribuciones relativas del eje 3 a las variables longitud y anchura del sépalo poseen una gran asimetría y grandes desviaciones de la normalidad. Lo mismo sucede con la variable longitud del pétalo aunque en menor medida.

	V. Obs	Media	SE	Sesgo	t-b inf	t-b sup	perc inf	perc sup	BCa inf	BCa sup
E.1										
S.L	755.88	764.70	100.61	8.82	567.26	962.13	526.69	915.46	439.88	887.87
S.W	880.57	830.66	116.26	-49.91	602.51	1058.81	537.95	974.59	662.41	985.87
P.L	524.25	533.99	220.44	9.75	101.41	966.58	144.71	901.24	144.68	901.15
P.W	426.17	458.88	137.77	32.70	188.52	729.23	181.30	750.47	120.85	676.56
E.2										
S.L	241.32	224.08	106.32	-17.24	15.45	432.71	45.12	470.81	93.65	584.58
S.W	118.20	149.52	98.92	31.32	-44.60	343.64	24.44	374.65	18.50	335.56
P.L	316.12	277.44	191.87	-38.68	-99.07	653.96	22.47	714.59	70.97	904.73
P.W	21.46	95.71	116.00	74.24	-131.93	323.34	0.18	438.41	0.00	188.98
E.3										
S.L	2.81	11.22	29.11	8.42	-45.90	68.35	0.01	124.70	0.02	132.30
S.W	1.23	19.82	47.69	18.59	-73.77	113.40	0.01	192.38	0.00	35.69
P.L	159.63	188.56	184.26	28.94	-173.03	550.15	2.48	663.13	7.66	732.42
P.W	552.36	445.42	163.32	-106.94	124.93	765.91	85.59	739.89	343.86	904.27

Tabla 4.11: Resultados bootstrap para la contribuciones relativas de los ejes factoriales a las variables.

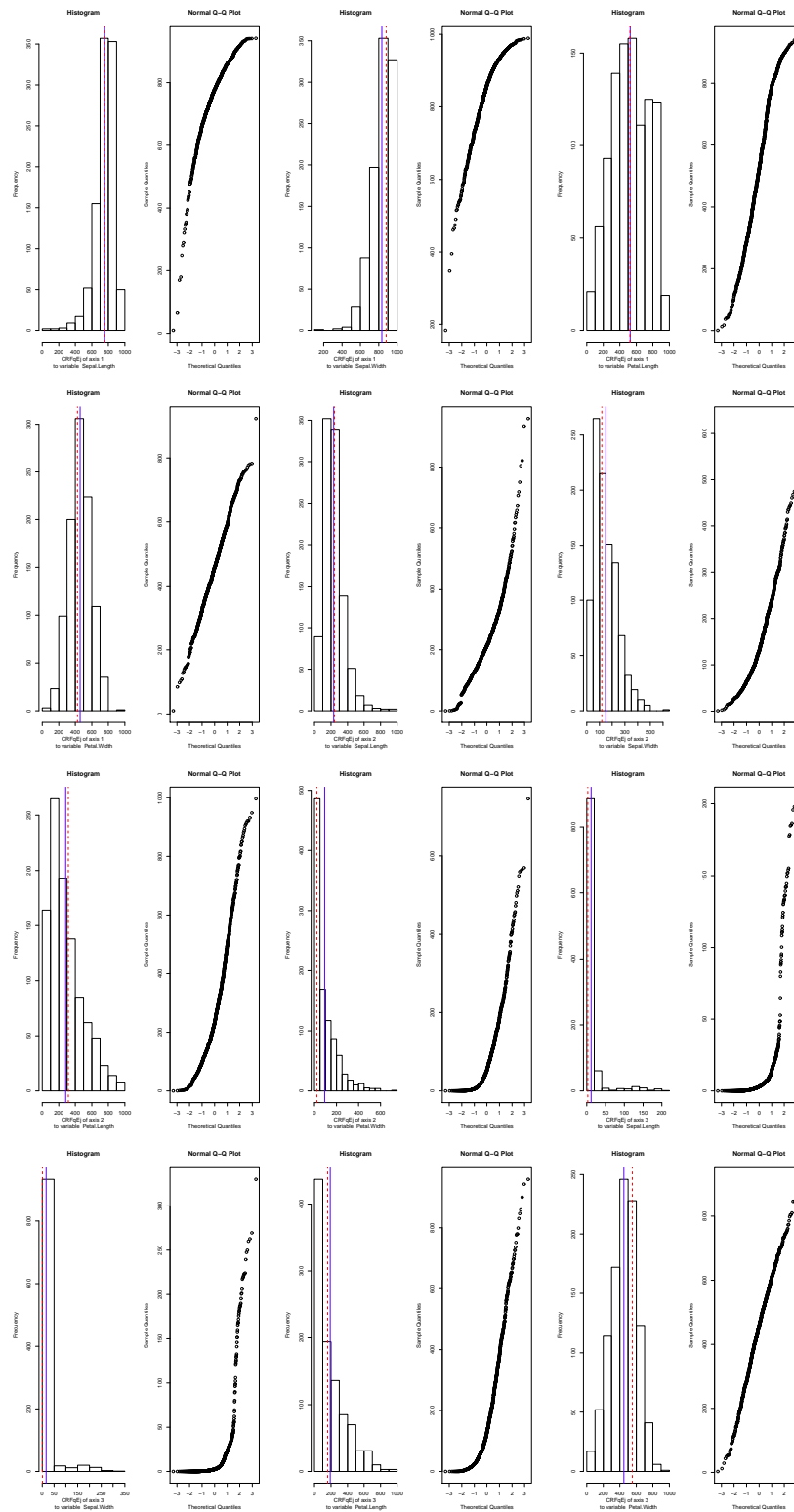


Figura 4.10: Histogramas y gráficos de normalidad para las contribuciones relativas de los ejes factoriales a las variables.

Los siguientes parámetros que se estudian son las contribuciones relativas a la variabilidad total de cada una de las matrices (*setosa*, *versicolor* y *virginica*) que forman parte del estudio. Los resultados se presentan en la tabla 4.12 y la figura 4.11. Estas medidas presentan pequeñas diferencias y los histogramas y gráficos de normalidad de las réplicas bootstrap muestran una simetría y normalidad muy aceptable. Por lo tanto, se considera que estas contribuciones son estables.

	V. Obs	Media	SE	Sesgo	t-b inf	t-b sup	perc inf	perc sup	BCa inf	BCa sup
1	318.94	331.37	81.13	12.43	172.16	490.58	179.06	501.08	169.46	483.76
2	356.26	338.67	92.94	-17.59	156.30	521.04	187.31	534.58	221.51	629.25
3	324.80	329.93	101.27	5.13	131.20	528.65	162.57	552.28	166.08	574.68

Tabla 4.12: Resultados bootstrap para las contribuciones relativas de las matrices a la variabilidad total.

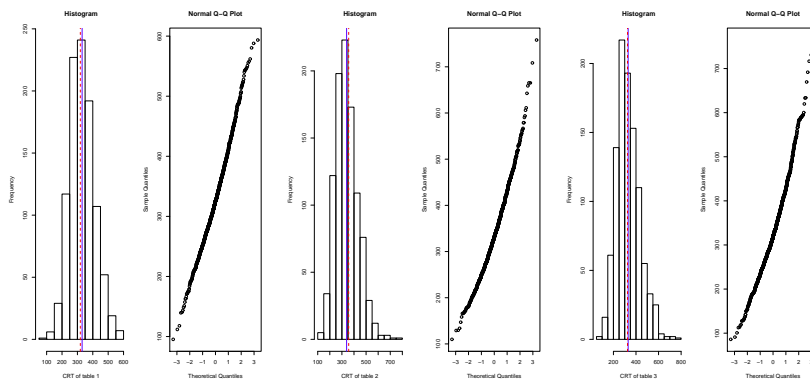


Figura 4.11: Histogramas y gráficos de normalidad para las contribuciones relativas a la variabilidad total de las matrices.

Si analizamos las contribuciones relativas de las matrices a la conformación de los ejes factoriales (tabla 4.13 y figura 4.12), las diferencias no son elevadas aunque la amplitud de los intervalos de confianza si que es grande en general. Respecto a los histogramas y a los gráficos de normalidad, se puede ver que en general presentan simetría y normalidad a excepción de la contribución de la matriz *setosa* al eje 2 y de *virginica* al eje 1.

	V. Obs	Media	SE	Sesgo	t-b inf	t-b sup	perc inf	perc sup	BCa inf	BCa sup
1										
E.1	361.93	363.32	78.13	1.40	210.00	516.65	220.76	531.06	225.73	538.28
E.2	140.25	188.10	111.49	47.85	-30.70	406.89	43.14	462.29	33.01	405.96
E.3	195.41	209.98	106.57	14.57	0.86	419.10	50.29	456.85	52.23	459.18
2										
E.1	353.96	335.22	80.01	-18.74	178.21	492.23	196.53	501.57	238.47	601.33
E.2	369.80	370.94	183.00	1.14	11.83	730.06	81.25	762.65	101.90	785.83
E.3	343.82	375.49	174.88	31.66	32.31	718.67	94.04	734.59	85.06	718.83
3										
E.1	284.12	301.57	79.22	17.46	146.12	457.03	175.73	475.11	166.32	455.50
E.2	489.95	440.23	201.83	-49.72	44.18	836.29	108.43	838.10	161.51	872.64
E.3	460.77	413.87	190.83	-46.90	39.40	788.34	109.46	804.54	156.92	865.65

Tabla 4.13: Resultados bootstrap para las contribuciones relativas de las matrices a los ejes factoriales.

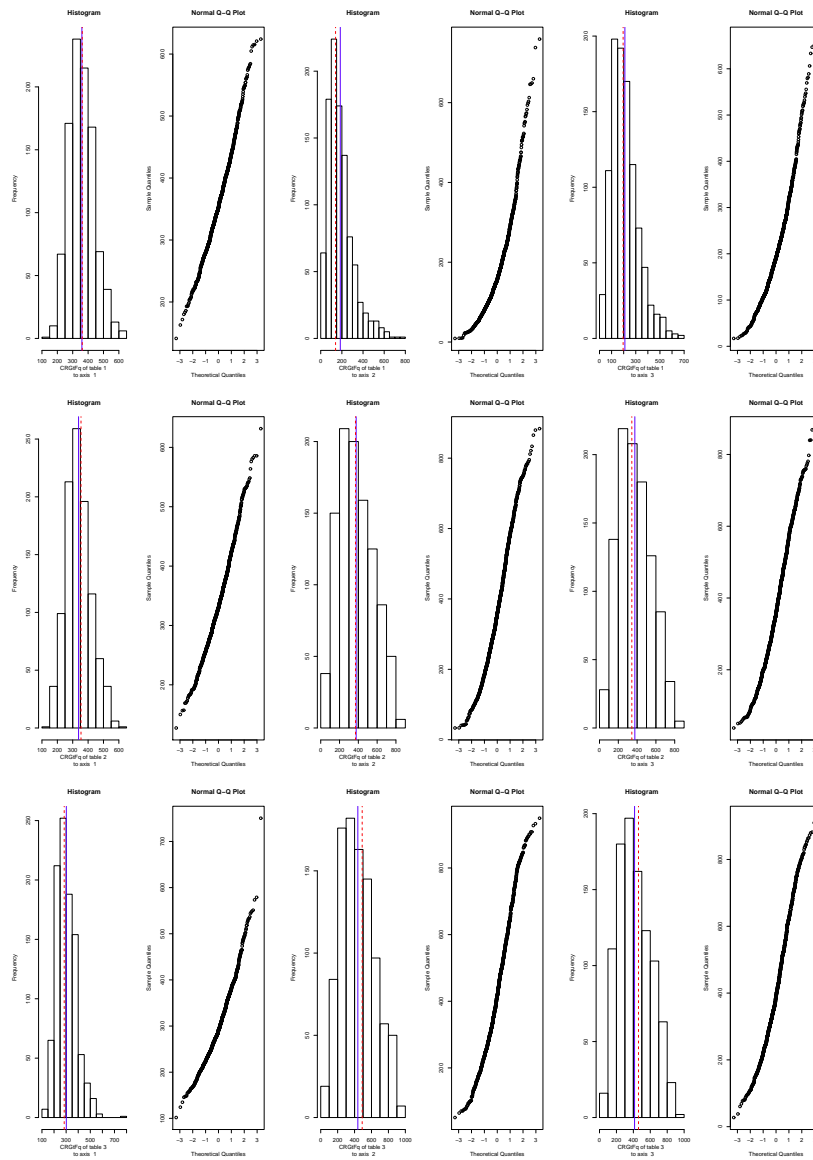


Figura 4.12: Histogramas y gráficos de normalidad para las contribuciones relativas de las matrices a la conformación de los ejes factoriales.

Para finalizar con el estudio inferencial de las medidas relativas a las diferentes matrices que forman parte del estudio, se muestran los resultados bootstrap relativos a las contribuciones relativas de los ejes factoriales a las matrices (tabla 4.14 y figura 4.13). En este caso, los sesgos y las desviaciones estándar son bastante bajos y las distribuciones de las réplicas en general, son simétricas y tienden a la normalidad a excepción de las contribuciones a la tabla *setosa* que presentan grandes asimetrías y desviaciones de la normalidad.

	V. Obs	Media	SE	Sesgo	t-b inf	t-b sup	perc inf	perc sup	BCa inf	BCa sup
E.1										
1	904.86	889.89	78.02	-14.96	736.79	1042.99	631.30	960.70	595.78	954.44
2	792.23	774.47	66.68	-17.77	643.61	905.32	627.33	884.19	666.53	898.63
3	697.51	697.68	88.99	0.18	523.06	872.30	505.77	849.21	494.33	841.03
E.2										
1	73.72	82.18	53.35	8.46	-22.51	186.88	28.51	256.08	33.54	282.42
2	174.02	169.25	57.54	-4.76	56.35	282.16	76.40	299.70	87.83	328.52
3	252.89	245.50	79.32	-7.39	89.84	401.16	115.63	422.66	133.06	457.05
E.3										
1	21.43	27.93	29.43	6.50	-29.82	85.67	4.83	126.20	7.56	164.08
2	33.75	56.28	26.26	22.53	4.75	107.82	21.71	121.89	14.70	58.65
3	49.61	56.82	25.76	7.21	6.27	107.38	22.56	117.27	21.51	112.56

Tabla 4.14: Resultados bootstrap para las contribuciones relativas de los ejes factoriales a las matrices.

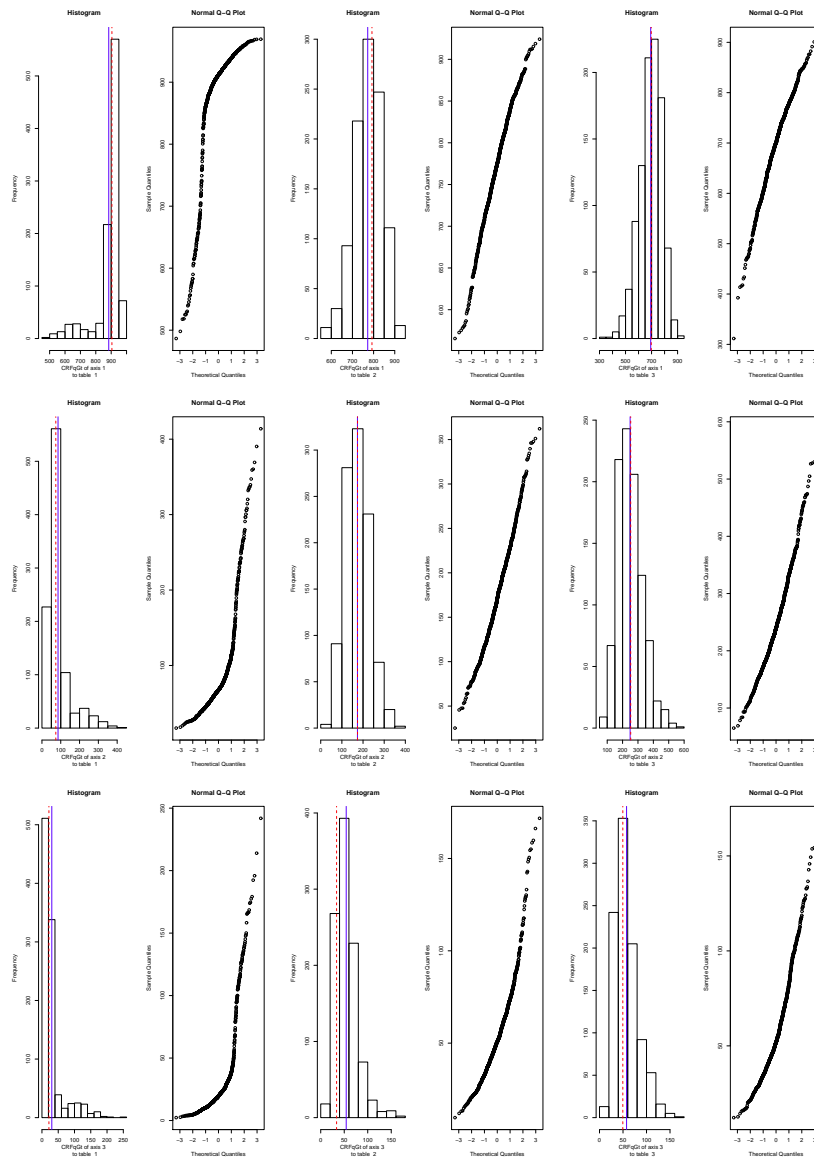


Figura 4.13: Histogramas y gráficos de normalidad para las contribuciones relativas de los ejes factoriales a las matrices.



En el apéndice A se encuentran los resultados relativos a los individuos. En general, las contribuciones relativas de los individuos a la variabilidad total y las contribuciones relativas de los individuos a la conformación de los ejes factoriales son bastante estables, no siendo así para el caso de las contribuciones de los ejes factoriales a los individuos que presentan diferencias significativas en algunos casos e intervalos de confianza con amplitudes demasiado grandes que cubren prácticamente todos los valores posibles.

Por último se presenta el gráfico (figura 4.14) con las coordenadas calculadas para las variables de las 1000 muestras bootstrap. El conjunto de coordenadas que se refiere a una misma variable se ha representado con el mismo color y se han generado polígonos envolventes (convex-hull) para cada uno de ellos.

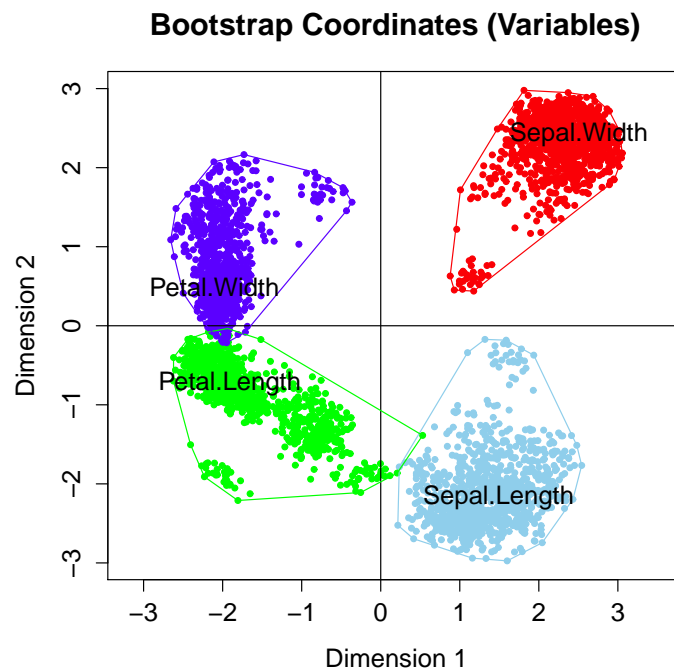


Figura 4.14: Coordenadas de las variables para las 1000 muestras bootstrap.

A modo de resumen, en este capítulo se ha expuesto una revisión de las técnicas disponibles en la literatura para el análisis de datos de tres vías y las

principales versiones inferenciales basadas en remuestreo que existen para ellas. Se ha desarrollado un nuevo software para la utilización del Biplot Múltiple ya que no existía ninguna implementación para su puesta en práctica y además se ha propuesto una versión inferencial que puede ser ejecutada desde el mismo software. Por último, se ha ilustrado el manejo y la interpretación de los resultados de los Biplot Múltiples y su nueva versión inferencial mediante su utilización con un conjunto de datos.

## Capítulo 5

# CLUSTERING DISJOINT BIPLOT



## 5.1. Clustering Biplot

Generalmente se suelen utilizar técnicas multivariantes de reducción de la dimensión para inspeccionar una posible estructura de grupos de las unidades que se pretenden analizar (PCA, FA, Biplot...). Es muy frecuente que una vez extraídas las primeras dimensiones que más información aportan, se aplique una técnica de clasificación sobre las cargas obtenidas para las unidades. Sin embargo, hay autores (De Soete y Carroll, 1994; DeSarbo et al., 1990b) que cargan en contra de este proceso, llamado *análisis tandem* por Arabie y Hubert, 1994. Esto es debido a que, por lo general, estas técnicas no buscan las dimensiones que más separan los posibles grupos que pueda haber si no que buscan direcciones de máxima variabilidad del conjunto total de datos disponibles. Un ejemplo de cómo este procedimiento puede ser no adecuado se puede ver en Vichi y Kiers, 2001.

Para resolver este problema, De Soete y Carroll, 1994 proponen una alternativa al algoritmo de k-means de tal forma que cada cluster se representa mediante un único punto, el centroide, y con dichos puntos se buscan las dimensiones que más variabilidad expliquen para ellos. Una vez extraídas dichas dimensiones, se proyectan los puntos de los datos de partida sobre el espacio generado por estas dimensiones y se obtiene una representación en dimensión reducida para unidades y centroides. Sin embargo, esta técnica tiene el problema de que no representa bien los datos de partida, ya que lo que se ha buscado es la máxima separación entre centroides que no tiene porque ser la que más variabilidad explique de los datos originales.

Para poder optimizar al mismo tiempo tanto la separación entre clusters como la variabilidad explicada por las dimensiones extraídas, se han propuesto varias alternativas que combinan métodos de clusters con técnicas de reducción de la

dimensión. Algunos utilizan escalamiento multidimensional o análisis unfolding (DeSarbo y Heiser, 1993; DeSarbo et al., 1990a; Heiser, 1993). Otros utilizan el ACP o el AF para este propósito (Vichi y Kiers, 2001).

Otra forma de actuar cuando se dispone de una matriz de distancias o similitudes es utilizar el escalamiento multidimensional (MDS). El problema es que los factores y la representación obtenida no son fáciles de interpretar. Para facilitar esta interpretación, se han propuesto procedimientos para establecer relaciones entre dichos factores y variables externas (Cox y Cox, 1994; Green et al., 1989; Hair et al., 1998; Heiser y Meulman, 1983). En ellos, se describen métodos que combinan MDS, técnicas de cluster y regresión múltiple. De esta forma, se utiliza el MDS para obtener coordenadas de los objetos en un espacio de dimensión reducida, con estas coordenadas se realiza una regresión múltiple con la matriz que contiene las variables externas y por último se aplica el análisis de clusters sobre los resultados de la regresión. También es posible intercambiar los pasos de regresión y análisis de clusters. Sin embargo, en ambas alternativas ni el análisis de clusters ni la regresión tienen influencia sobre la solución proporcionada por el MDS. Para que en la extracción de las dimensiones en el MDS se tengan en cuenta las variables externas, se realiza en primer lugar la regresión múltiple. Así, se proyecta la matriz de distancias sobre el espacio de las variables externas y posteriormente se analizan estas distancias proyectadas mediante MDS y análisis de clusters o viceversa. Pero esta aproximación tampoco resulta óptima ya que las funciones objetivo se optimizan secuencialmente y no simultáneamente. Bock, 1987 propone combinar aditivamente dos funciones objetivo relativas al análisis de clusters k-means y al MDS. Heiser, 1993; Heiser y Groenen, 1997 muestran un procedimiento en el que en primer lugar divide los objetos en un número de clases y simultáneamente busca espacios de representación en dimensión reducida para los centroides de los clusters conformados. Kiers et al., 2005 proponen como

alternativa óptima el modelo MDSCLUREG que combina los tres procedimientos en uno de tal forma que los centroides de los clusters se explican mediante un número reducido de dimensiones (MDS) que son consideradas como variables dependientes predecidas por las variables externas. Rocci et al., 2011 proponen un algoritmo en el que convergen las propuestas de De Soete y Carroll, 1994; Vichi y Kiers, 2001, llamado análisis Discriminante Factorial k-means (FDKM) que alterna pasos del análisis Discriminante Lineal con el análisis de clusters k-means.

Si la partición de los objetos en clusters es conocida a priori, Amaro et al., 2004; Gabriel, 1995 proponen el método Biplot Canónico o MANOVA-Biplot que representa tanto los individuos como las variables en el mismo espacio de dimensión reducida y maximiza la distancia entre los clusters utilizando la distancia de Mahalanobis en lugar de la distancia Euclídea.

Siguiendo las ideas de los autores señalados anteriormente, se propone un nuevo algoritmo que aborda el doble problema de clasificación de objetos en grupos y representación en un espacio de dimensión reducida. Dicho algoritmo se denomina *Clustering Biplot* (CBiplot) y combina el análisis de clusters k-means con el método HJ Biplot (Galindo, 1986). El objetivo de este método es, por tanto, encontrar las direcciones que maximizan la separación entre los centroides de los  $P$  clusters que se pretenden encontrar a partir de los datos analizados y obtener una representación HJ Biplot de los individuos y de las variables en el espacio que conforman dichas dimensiones.

### 5.1.1. Notación

Para un mejor seguimiento del algoritmo que se va a proponer en la siguiente sección se expone brevemente la terminología que se va a utilizar:

- $\mathbf{I}, \mathbf{J}, \mathbf{P}, \mathbf{Q}$ . Número de objetos, variables, clusters y componentes respectivamente.
- $\mathbf{X}_{ij}$ . Matriz de datos que contiene la información de  $I$  objetos sobre los que se han medido  $J$  variables. Se supone estandarizada.
- $\mathbf{E}_{ij}$ . Matriz de errores.
- $\mathbf{U}_{ip}$ . Matriz binaria que contiene la localización de los  $I$  individuos en los  $P$  clusters. Es decir,  $u_{ip} = 1$  si el individuo  $i$  pertenece al cluster  $p$  y  $u_{ip} = 0$  en otro caso. Es una matriz estocástica por filas, es decir, todos los elementos son no negativos y  $\sum_{t=1}^P u_{it} = 1 \forall i = 1, \dots, I$ .
- $\bar{\mathbf{X}}_{pj}$ . Matriz de centroides en el espacio original.
- $\mathbf{Z}_{ij}$ . Matriz en la que se ha sustituido el elemento original por el centroide del cluster al que pertenece.
- $\mathbf{A}_{iq}$ . Matriz que contiene las coordenadas HJ Biplot de los objetos en el espacio de las  $Q$  componentes.
- $\bar{\mathbf{A}}_{pq}$ . Matriz que contiene las coordenadas HJ Biplot de los centroides de los  $P$  clusters en el espacio de las  $Q$  componentes.
- $\mathbf{B}_{jq}$ . Matriz que contiene las coordenadas HJ Biplot de las  $J$  variables en el espacio de las  $Q$  componentes.
- $\mathbf{\Lambda}$ . Matriz diagonal que contiene los valores propios de la dvs de  $\mathbf{Z}$ .

### 5.1.2. Modelo Clustering Biplot

El modelo asociado al CBiplot se obtiene de la aplicación del HJ Biplot a la matriz en la que se ha sustituido cada valor original por el valor del centroide del



cluster al que pertenezca. Por tanto, en primer lugar, se aplica el método k-means sobre dicha matriz obteniendo el modelo:

$$\mathbf{X} = \mathbf{U}\bar{\mathbf{X}} + \mathbf{E}_1 \quad (5.1)$$

A continuación, se busca la descomposición de la matriz de centroides mediante el método HJ Biplot. Así, se obtienen las coordenadas para variables  $\mathbf{B}$  y para los centroides  $\bar{\mathbf{A}}$  y la matriz diagonal que contiene los valores propios  $\mathbf{\Lambda}$ .

Si:

$$\bar{\mathbf{X}} = \mathbf{D}\mathbf{A}\mathbf{M}^\top$$

con  $\mathbf{D}^\top\mathbf{D} = \mathbf{M}^\top\mathbf{M} = \mathbf{I}$ , entonces:

$$\bar{\mathbf{A}} = \mathbf{D}\mathbf{\Lambda}$$

y

$$\mathbf{B} = \mathbf{M}\mathbf{\Lambda}.$$

Como el método HJ Biplot no reproduce el dato de partida, se introduce un factor para hacer posible dicha recuperación. De esta forma la matriz de centroides queda expresada como:

$$\bar{\mathbf{X}} = \bar{\mathbf{A}}\mathbf{\Lambda}^{-1}\mathbf{B}^\top + \mathbf{E}_2 \quad (5.2)$$

Por lo tanto, el modelo asociado al CBiplot es:

$$\mathbf{X} = \mathbf{U}\bar{\mathbf{A}}\mathbf{\Lambda}^{-1}\mathbf{B}^\top + \mathbf{E} \quad (5.3)$$

donde  $\mathbf{E} = \mathbf{E}_1 + \mathbf{E}_2$ .

A partir de este modelo es fácil ver que:  $\mathbf{E} = \mathbf{X} - \mathbf{U}\bar{\mathbf{A}}\boldsymbol{\Lambda}^{-1}\mathbf{B}^\top$ . Por lo tanto, se plantea el problema de minimizar la norma de la matriz de errores  $\mathbf{E}$ , lo que resulta en el problema de optimización:

$$\begin{aligned} F(\mathbf{U}, \bar{\mathbf{A}}, \mathbf{B}) &= \|\mathbf{X} - \mathbf{U}\bar{\mathbf{A}}\boldsymbol{\Lambda}^{-1}\mathbf{B}^\top\|^2 \\ &= \sum_{i=1}^I \sum_{j=1}^J \left( x_{ij} - \sum_{p=1}^P \sum_{q=1}^Q u_{ip} \bar{a}_{pq} \lambda_q^{-1} b_{jq} \right)^2 \rightarrow \min_{\mathbf{U}\bar{\mathbf{A}}\mathbf{B}} \end{aligned}$$

Por otro lado, se puede probar que se cumple la siguiente descomposición:

$$\|\mathbf{X}\|^2 = \|\mathbf{X} - \mathbf{U}\bar{\mathbf{A}}\boldsymbol{\Lambda}^{-1}\mathbf{B}^\top\|^2 + \|\mathbf{U}\bar{\mathbf{A}}\boldsymbol{\Lambda}^{-1}\mathbf{B}^\top\|^2$$

*Demostración.* Dado que  $\mathbf{B} = \mathbf{M}\boldsymbol{\Lambda}$ , se cumple que  $\mathbf{B}^\top\mathbf{B} = \boldsymbol{\Lambda}\mathbf{M}^\top\mathbf{M}\boldsymbol{\Lambda}$  y como  $\mathbf{M}$  se ha definido con la condición  $\mathbf{M}^\top\mathbf{M} = \mathbf{I}$ , se tiene que  $\mathbf{B}^\top\mathbf{B} = \boldsymbol{\Lambda}^2$ .

Con esta afirmación, sabiendo que  $\mathbf{X} = \mathbf{U}\bar{\mathbf{X}}$  y  $\bar{\mathbf{A}} = \bar{\mathbf{X}}\mathbf{B}\boldsymbol{\Lambda}^{-1}$ , se tiene que:

$$\begin{aligned} &\|\mathbf{X} - \mathbf{U}\bar{\mathbf{A}}\boldsymbol{\Lambda}^{-1}\mathbf{B}^\top\|^2 + \|\mathbf{U}\bar{\mathbf{A}}\boldsymbol{\Lambda}^{-1}\mathbf{B}^\top\|^2 \\ &= \text{traza} \left( [\mathbf{X} - \mathbf{U}\bar{\mathbf{X}}\mathbf{B}\boldsymbol{\Lambda}^{-2}\mathbf{B}^\top] [\mathbf{X} - \mathbf{U}\bar{\mathbf{X}}\mathbf{B}\boldsymbol{\Lambda}^{-2}\mathbf{B}^\top]^\top \right) \\ &\quad + \text{traza} \left( [\mathbf{U}\bar{\mathbf{X}}\mathbf{B}\boldsymbol{\Lambda}^{-2}\mathbf{B}^\top] [\mathbf{U}\bar{\mathbf{X}}\mathbf{B}\boldsymbol{\Lambda}^{-2}\mathbf{B}^\top]^\top \right) \\ &= \text{traza} (\mathbf{X}\mathbf{X}^\top) - 2\text{traza} (\mathbf{U}\bar{\mathbf{X}}\mathbf{B}\boldsymbol{\Lambda}^{-2}\mathbf{B}^\top\mathbf{X}^\top) \\ &\quad + 2\text{traza} \left( [\mathbf{U}\bar{\mathbf{X}}\mathbf{B}\boldsymbol{\Lambda}^{-2}\mathbf{B}^\top] [\mathbf{U}\bar{\mathbf{X}}\mathbf{B}\boldsymbol{\Lambda}^{-2}\mathbf{B}^\top]^\top \right) \end{aligned}$$

$$\begin{aligned}
&= \text{traza}(\mathbf{X}\mathbf{X}^\top) - 2\text{traza}(\mathbf{U}\bar{\mathbf{X}}\mathbf{B}\mathbf{\Lambda}^{-2}\mathbf{B}^\top\mathbf{X}^\top) \\
&\quad + 2\text{traza}(\mathbf{U}\bar{\mathbf{X}}\mathbf{B}\mathbf{\Lambda}^{-2}\mathbf{B}^\top\mathbf{B}\mathbf{\Lambda}^{-2}\mathbf{B}^\top[\mathbf{U}\bar{\mathbf{X}}]^\top) \\
&= \text{traza}(\mathbf{X}\mathbf{X}^\top) - 2\text{traza}(\mathbf{U}\bar{\mathbf{X}}\mathbf{B}\mathbf{\Lambda}^{-2}\mathbf{B}^\top\mathbf{X}^\top) \\
&\quad + 2\text{traza}(\mathbf{U}\bar{\mathbf{X}}\mathbf{B}\mathbf{\Lambda}^{-2}\mathbf{B}^\top\mathbf{X}^\top) \\
&= \text{traza}(\mathbf{X}\mathbf{X}^\top)
\end{aligned}$$

□

Por lo tanto, minimizar  $F$  es equivalente a maximizar  $\|\mathbf{U}\bar{\mathbf{A}}\mathbf{\Lambda}^{-1}\mathbf{B}^\top\|^2$  que corresponde con la deviance entre clusters de la clasificación de  $\mathbf{X}$  dada por  $\mathbf{U}$ .

Por otro lado se tiene que:

$$\begin{aligned}
\|\mathbf{U}\bar{\mathbf{A}}\mathbf{\Lambda}^{-1}\mathbf{B}^\top\|^2 &= \|\mathbf{U}\bar{\mathbf{X}}\mathbf{B}\mathbf{\Lambda}^{-2}\mathbf{B}^\top\|^2 = \text{traza}([\mathbf{U}\bar{\mathbf{X}}\mathbf{B}\mathbf{\Lambda}^{-2}\mathbf{B}^\top][\mathbf{U}\bar{\mathbf{X}}\mathbf{B}\mathbf{\Lambda}^{-2}\mathbf{B}^\top]^\top) \\
&= \text{traza}([\mathbf{U}\bar{\mathbf{X}}\mathbf{B}\mathbf{\Lambda}^{-1}]\mathbf{\Lambda}^{-1}\mathbf{B}^\top\mathbf{B}\mathbf{\Lambda}^{-1}[\mathbf{U}\bar{\mathbf{X}}\mathbf{B}\mathbf{\Lambda}^{-1}]^\top) \\
&= \text{traza}([\mathbf{U}\bar{\mathbf{X}}\mathbf{B}\mathbf{\Lambda}^{-1}][\mathbf{U}\bar{\mathbf{X}}\mathbf{B}\mathbf{\Lambda}^{-1}]^\top)
\end{aligned}$$

$$= \| \mathbf{U} \bar{\mathbf{X}} \mathbf{B} \boldsymbol{\Lambda}^{-1} \|^2 = \| \mathbf{U} \bar{\mathbf{A}} \|^2 .$$

Por lo tanto, minimizar  $F$  es equivalente a maximizar la deviance entre clusters en el espacio reducido.

Ahora bien, si se parte del modelo de k-means  $\mathbf{X} = \mathbf{U} \bar{\mathbf{X}} + \mathbf{E}_1$ , hay que minimizar la norma  $\| \mathbf{X} - \mathbf{U} \bar{\mathbf{X}} \|^2$ .

Pero,

$$\| \mathbf{X} - \mathbf{U} \bar{\mathbf{X}} \|^2 = \| \mathbf{X} \mathbf{B} \boldsymbol{\Lambda}^{-1} - \mathbf{U} \bar{\mathbf{X}} \mathbf{B} \boldsymbol{\Lambda}^{-1} \|^2 = \| \mathbf{A} - \mathbf{U} \bar{\mathbf{A}} \|^2 .$$

A continuación, se demuestra que:

$$\| \mathbf{A} \|^2 = \| \mathbf{A} - \mathbf{U} \bar{\mathbf{A}} \|^2 + \| \mathbf{U} \bar{\mathbf{A}} \|^2$$

*Demostración.* Sabiendo que  $\mathbf{A} = \mathbf{U} \bar{\mathbf{A}}$ , se tiene que:

$$\begin{aligned} & \| \mathbf{A} - \mathbf{U} \bar{\mathbf{A}} \|^2 + \| \mathbf{U} \bar{\mathbf{A}} \|^2 \\ &= \text{traza} \left( [\mathbf{X} - \mathbf{U} \bar{\mathbf{A}}] [\mathbf{X} - \mathbf{U} \bar{\mathbf{A}}]^\top \right) \\ & \quad + \text{traza} \left( [\mathbf{U} \bar{\mathbf{A}}] [\mathbf{U} \bar{\mathbf{A}}]^\top \right) \\ &= \text{traza} (\mathbf{A} \mathbf{A}^\top) - 2 \text{traza} (\mathbf{U} \bar{\mathbf{A}} \mathbf{A}^\top) \\ & \quad + 2 \text{traza} \left( [\mathbf{U} \bar{\mathbf{A}}] [\mathbf{U} \bar{\mathbf{A}}]^\top \right) \end{aligned}$$

$$= \text{traza}(\mathbf{A}\mathbf{A}^\top)$$

□

Como se ha demostrado antes, minimizar  $F$  es equivalente a maximizar  $\|\mathbf{U}\bar{\mathbf{A}}\|^2$  y según la descomposición  $\|\mathbf{A}\|^2 = \|\mathbf{A} - \mathbf{U}\bar{\mathbf{A}}\|^2 + \|\mathbf{U}\bar{\mathbf{A}}\|^2$ , maximizar  $\|\mathbf{U}\bar{\mathbf{A}}\|^2$  es lo mismo que minimizar  $\|\mathbf{A} - \mathbf{U}\bar{\mathbf{A}}\|^2$ , lo que corresponde con la deviance dentro de los clusters en el espacio reducido.

En el algoritmo que se explica en la siguiente sección se toma como función objetivo la maximización de la deviance entre clusters. Como criterios de parada en el proceso iterativo, se fija un umbral de tolerancia de tal forma que si la diferencia entre la función objetivo en el paso  $k$  y la función objetivo en el paso  $k + 1$  es menor que dicho umbral y la función objetivo en el paso  $k + 1$  es mayor que la del paso  $k$  se interrumpe el proceso.

### 5.1.3. Algoritmo CBiplot

El algoritmo parte de una clasificación aleatoria de los individuos en los  $P$  clusters y mediante un procedimiento iterativo se busca la clasificación óptima que conduzca a la maximización de las distancias entre los centroides de los grupos resultantes.

Sea:

$\mathbf{X}$  la matriz de datos de partida de orden  $I \times J$  que contiene la información de  $I$  individuos sobre los que se han medido  $J$  variables.

El algoritmo consta de un paso inicial que se detalla a continuación:

Se genera la matriz  $\mathbf{U}_0$  aleatoria binaria de orden  $I \times P$  que contiene la localización de los  $I$  individuos en los  $P$  clusters. Es decir,  $u_{ip} = 1$  si el individuo  $i$  pertenece al cluster  $p$  y  $u_{ip} = 0$  en otro caso. Es una matriz estocástica por filas,

es decir, todos los elementos son no negativos y  $\sum_{t=1}^P u_{it} = 1 \forall i = 1, \dots, I$ . Si en este proceso se genera alguna columna vacía para  $\mathbf{U}_0$ , es decir, algún cluster resulta vacío, se elige el de mayor tamaño y se divide aleatoriamente a la mitad. Este procedimiento se repite hasta que todos los clusters tengan algún elemento.

A partir de las matrices  $\mathbf{X}$  y  $\mathbf{U}_0$  se calcula la matriz de centroides  $\bar{\mathbf{X}}_0$  de orden  $P \times J$  mediante la ecuación:

$$\bar{\mathbf{X}}_0 = (\mathbf{U}_0^\top \mathbf{U}_0)^{-1} \mathbf{U}_0^\top \mathbf{X}.$$

Con dicha matriz y la matriz de localización de los individuos en los clusters  $\mathbf{U}_0$ , se calcula  $\mathbf{Z}_0$ . Esta matriz es de orden  $I \times J$  y se construye sustituyendo los valores iniciales de  $\mathbf{X}$  por el valor de los centroides de los clusters a los que pertenece cada individuo.

$$\mathbf{Z}_0 = \mathbf{U}_0 \bar{\mathbf{X}}_0.$$

A partir de la matriz  $\mathbf{Z}_0$  se van a buscar las direcciones de máxima separación entre los centroides calculados previamente. Por tanto, se realiza la descomposición en valores singulares (dvs) de la matriz  $\mathbf{Z}_0$ .

$$\mathbf{Z}_0 = \mathbf{R}\mathbf{A}\mathbf{T}^\top.$$

Se eligen  $\mathbf{A}_0 = \mathbf{R}\mathbf{A}$  como coordenadas para los individuos y  $\mathbf{B}_0 = \mathbf{T}\mathbf{A}$  como coordenadas para las variables. Así mismo, se calculan las coordenadas de los centroides  $\bar{\mathbf{A}}_0$  en el nuevo espacio de representación.

Con estas coordenadas se evalúa la función objetivo  $F_0$ .

Una vez terminado el paso inicial, se comienza con el proceso iterativo.

Se supone que se ha ejecutado hasta el paso  $k - 1$ , por lo tanto, se dispone de

las matrices  $\mathbf{U}_{k-1}$ ,  $\mathbf{Z}_{k-1}$ ,  $\mathbf{A}_{k-1}$ ,  $\mathbf{B}_{k-1}$  y  $\bar{\mathbf{A}}_{k-1}$  y la función objetivo en dicho paso  $F_{k-1}$ . A partir de ellas se va a realizar el paso  $k$ .

A partir de las matrices  $\mathbf{A}_{k-1}$  y  $\bar{\mathbf{A}}_{k-1}$  se actualiza la matriz  $\mathbf{U}_k$  asignando cada individuo al centroide más cercano en el nuevo espacio mediante el algoritmo de k-means. Si algún cluster resulta vacío, se procede de la misma forma que con la matriz inicial  $\mathbf{U}_0$ . Con la nueva matriz de localización de los individuos en los  $P$  clusters  $\mathbf{U}_k$  y la matriz  $\mathbf{X}$  se calcula la nueva matriz de centroides  $\bar{\mathbf{X}}_k = (\mathbf{U}_k^\top \mathbf{U}_k)^{-1} \mathbf{U}_k^\top \mathbf{X}$ . Con ella y  $\mathbf{U}_k$  se actualiza  $\mathbf{Z}_k = \mathbf{U}_k \bar{\mathbf{X}}_k$ .

A continuación se vuelve a calcular la dvs de  $\mathbf{Z}_k$  y se actualizan las matrices de coordenadas para individuos, variables y centroides  $\mathbf{A}_k$ ,  $\mathbf{B}_k$  y  $\bar{\mathbf{A}}_k$ .

Con dichas coordenadas se computa el valor de la función objetivo y se comprueba si se cumplen los criterios de parada explicados al comienzo de esta sección. En este caso, se termina el proceso iterativo y se considera como solución los resultados obtenidos en este paso, si no se repite este paso hasta que se cumplan los criterios de parada.

En la figura 5.1 se resumen los pasos principales del algoritmo Clustering Biplot.

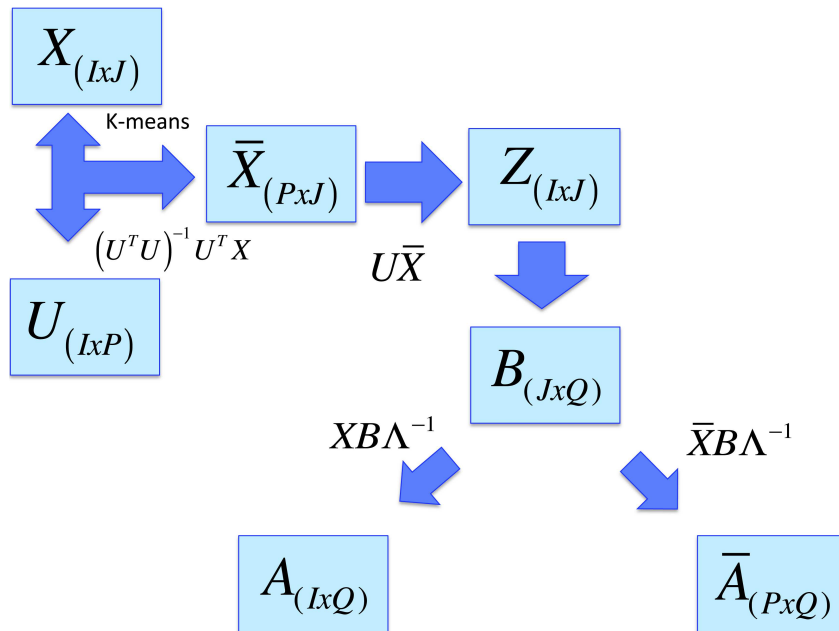


Figura 5.1: Esquema del algoritmo CBiplot.

## 5.2. Disjoint Biplot

Los métodos Biplot (Gabriel, 1971; Galindo, 1986) son herramientas ampliamente utilizadas para obtener una representación conjunta de los objetos y de las variables recogidas en una matriz de datos en un espacio de dimensión reducida. Como se ha detallado en el capítulo 2.1 se han aplicado en numerosos campos. Al igual que ocurre en el PCA, los ejes que se eligen para representar los datos en el espacio de dimensión reducida se obtienen de tal forma que expliquen la mayor parte de la variabilidad presente en los datos. Estos ejes tienen correlaciones cero entre sí y son combinaciones lineales de las variables de partida. Sin embargo, este hecho se convierte en una desventaja a la hora de interpretar los resultados obtenidos. En el contexto del PCA se han desarrollado diferentes alternativas para resolver este problema. Una posible solución es el uso de técnicas de rotación (Jolliffe, 1995). Vines, 2000 propone un nuevo método que denomina Simple PCA, cuya idea básica es forzar a que las cargas factoriales sean  $-1, 0$  o  $1$ . Una



forma artificial que también se ha utilizado frecuentemente es convertir en nulas las cargas factoriales menores que un cierto umbral prefijado (Cadima y Jolliffe, 1995). Jolliffe et al., 2003 presentan el método SCoTLASS que introduce una penalización en la suma de los valores absolutos de las cargas factoriales de tal forma que algunas de ellas se hacen nulas. En el mismo sentido, Zhang et al., 2002, 2004 proponen SLRAs (*Sparse Low Rank Approximations*). Posteriormente, Zou et al., 2006 proponen el Sparse PCA, cuyo objetivo es la búsqueda de cargas *sparse*, es decir, buscan matrices de cargas factoriales que contengan escasos elementos no nulos. Dicho método se basa en la reformulación del PCA como un problema de regresión y así poder integrar directamente técnicas utilizadas en este contexto que persiguen el propósito de mejorar la interpretabilidad de los resultados. Entre dichas técnicas se encuentran *lasso* (Tibshirani, 1996) y *elastic net* (Zou y Hastie, 2003). D'Aspremont et al., 2007 presentan una aproximación del Sparse PCA basada en programación semidefinida. Una revisión exhaustiva desde la formulación clásica del PCA hasta estas aproximaciones se puede consultar en Trendafilov, 2014. McCabe, 1984 presenta una alternativa en la que se selecciona un conjunto reducido de variables que llama *variables principales*. En este sentido, Mahoney y Drineas, 2009 proponen una alternativa a la dvs para aproximar una matriz de datos a la que han denominado descomposición *CUR*. La ventaja que tiene sobre la dvs es que permiten expresar la matriz de datos respecto de un número reducido de columnas y/o filas en lugar de extraer ejes factoriales que son combinaciones lineales de las variables de partida. Esto mejora la interpretación ya que lo que se obtiene son las propias variables y/u objetos. Dicha denominación se debe a que la matriz de partida se descompone en tres matrices **C**, **U** y **R** donde **C** y **R** contienen un número pequeño de columnas y filas respectivamente. Para seleccionar este número se genera una distribución de probabilidad a partir de la influencia de cada columna en la mejor aproximación de dimensión reducida

de la matriz de datos. Existen diferentes descomposiciones  $CUR$  según el criterio definido para elegir las columnas y las cotas de error obtenidas en ellos (Drineas et al., 2006, 2008; Frieze et al., 2004; Goreinov y Tyrtshnikov, 2001; Stewart, 1999). En Vichi y Saporta, 2009, se propone un método en el que, mediante un algoritmo de mínimos cuadrados alternados, se busca un espacio de dimensión reducida en el que cada variable únicamente contribuye a la conformación de uno sólo de los ejes factoriales extraídos a la vez que busca la mejor clasificación de los objetos situados en las filas de la matriz. En esta sección se va a presentar un nuevo algoritmo para obtener ejes factoriales disjuntos en la representación HJ Biplot. Se utiliza para la búsqueda de dichos ejes factoriales el procedimiento utilizado en Vichi y Saporta, 2009 relacionado con el propuesto por Vigneau y Qannari, 2004.

### 5.2.1. Notación

Para una lectura más clara del algoritmo explicado posteriormente se realiza una breve reseña de la terminología que se va a utilizar:

- $I, J, Q$ . Número de objetos, variables y componentes respectivamente.
- $\mathbf{X}_{ij}$ . Matriz de datos que contiene la información de  $I$  objetos sobre los que se han medido  $J$  variables. Se supone estandarizada.
- $\mathbf{E}_{ij}$ . Matriz de errores.
- $\mathbf{V}_{jq}$ . Matriz binaria que contiene la información relativa a la componente sobre la que va a contribuir cada una de las  $J$  variables. Es decir,  $v_{jq} = 1$  si la variable  $j$  contribuye a la componente  $q$  y  $v_{jq} = 0$  en otro caso. Al igual que sucedía con la matriz  $\mathbf{U}$  en el método descrito anteriormente, es una

matriz estocástica por filas, es decir, todos los elementos son no negativos y  $\sum_{t=1}^Q v_{jt} = 1 \forall j = 1, \dots, J$ .

- $\mathbf{A}_{iq}$ . Matriz que contiene las coordenadas HJ Biplot de los objetos en el espacio de las  $Q$  componentes disjuntas.
- $\mathbf{B}_{jq}$ . Matriz que contiene las coordenadas HJ Biplot de las  $J$  variables en el espacio de las  $Q$  componentes disjuntas.
- $\mathbf{\Lambda}$ . Matriz diagonal que contiene los valores propios de la dvs disjunta de  $\mathbf{X}$ .

### 5.2.2. Modelo Disjoint Biplot

El modelo asociado al *Disjoint Biplot* (DBiplot) se obtiene de la aplicación del HJ Biplot a la matriz que contiene los datos de partida. Así, se obtienen las coordenadas para variables  $\mathbf{B}$  y para objetos  $\mathbf{A}$  y la matriz diagonal que contiene los valores propios  $\mathbf{\Lambda}$ .

Si:

$$\mathbf{X} = \mathbf{D}\mathbf{A}\mathbf{M}^T$$

con  $\mathbf{D}^T\mathbf{D} = \mathbf{M}^T\mathbf{M} = \mathbf{I}$ , entonces:

$$\mathbf{A} = \mathbf{D}\mathbf{\Lambda}$$

y

$$\mathbf{B} = \mathbf{M}\mathbf{\Lambda}.$$

Al igual que ocurría en el método anterior, como el método HJ Biplot no

reproduce el dato de partida, se introduce un factor para hacer posible dicha recuperación. De esta forma se obtiene el modelo:

$$\mathbf{X} = \mathbf{A}\mathbf{\Lambda}^{-1}\mathbf{B}^{\top} + \mathbf{E} \quad (5.4)$$

A partir de este modelo es fácil ver que:  $\mathbf{E} = \mathbf{X} - \mathbf{A}\mathbf{\Lambda}^{-1}\mathbf{B}^{\top}$ . Por lo tanto, el problema planteado en este modelo es minimizar la norma de la matriz de errores  $\mathbf{E}$ , lo que resulta en el problema de optimización:

$$\begin{aligned} F(\mathbf{A}, \mathbf{B}) &= \|\mathbf{X} - \mathbf{A}\mathbf{\Lambda}^{-1}\mathbf{B}^{\top}\|^2 \\ &= \sum_{i=1}^I \sum_{j=1}^J \left( x_{ij} - \sum_{q=1}^Q a_{iq} \lambda_q^{-1} b_{jq} \right)^2 \rightarrow \min_{\mathbf{A}, \mathbf{B}} \end{aligned}$$

Por otro lado, se puede probar que se cumple la siguiente descomposición:

$$\|\mathbf{X}\|^2 = \|\mathbf{X} - \mathbf{U}\mathbf{A}\mathbf{\Lambda}^{-1}\mathbf{B}^{\top}\|^2 + \|\mathbf{U}\mathbf{A}\mathbf{\Lambda}^{-1}\mathbf{B}^{\top}\|^2$$

*Demostración.* Como  $\mathbf{B} = \mathbf{M}\mathbf{\Lambda}$ , se cumple que  $\mathbf{B}^{\top}\mathbf{B} = \mathbf{\Lambda}\mathbf{M}^{\top}\mathbf{M}\mathbf{\Lambda}$  y como  $\mathbf{M}$  se ha definido con la condición  $\mathbf{M}^{\top}\mathbf{M} = \mathbf{I}$ , se tiene que  $\mathbf{B}^{\top}\mathbf{B} = \mathbf{\Lambda}^2$ .

Con esta igualdad y  $\mathbf{A} = \mathbf{X}\mathbf{B}\mathbf{\Lambda}^{-1}$ , se tiene que:

$$\begin{aligned} &\|\mathbf{X} - \mathbf{A}\mathbf{\Lambda}^{-1}\mathbf{B}^{\top}\|^2 + \|\mathbf{A}\mathbf{\Lambda}^{-1}\mathbf{B}^{\top}\|^2 \\ &= \text{traza} \left( [\mathbf{X} - \mathbf{X}\mathbf{B}\mathbf{\Lambda}^{-2}\mathbf{B}^{\top}] [\mathbf{X} - \mathbf{X}\mathbf{B}\mathbf{\Lambda}^{-2}\mathbf{B}^{\top}]^{\top} \right) \\ &\quad + \text{traza} \left( [\mathbf{X}\mathbf{B}\mathbf{\Lambda}^{-2}\mathbf{B}^{\top}] [\mathbf{X}\mathbf{B}\mathbf{\Lambda}^{-2}\mathbf{B}^{\top}]^{\top} \right) \end{aligned}$$

$$\begin{aligned}
&= \text{traza}(\mathbf{X}\mathbf{X}^\top) - 2\text{traza}(\mathbf{X}\mathbf{B}\mathbf{\Lambda}^{-2}\mathbf{B}^\top\mathbf{X}^\top) \\
&\quad + 2\text{traza}\left([\mathbf{X}\mathbf{B}\mathbf{\Lambda}^{-2}\mathbf{B}^\top][\mathbf{X}\mathbf{B}\mathbf{\Lambda}^{-2}\mathbf{B}^\top]^\top\right) \\
&= \text{traza}(\mathbf{X}\mathbf{X}^\top) - 2\text{traza}(\mathbf{X}\mathbf{B}\mathbf{\Lambda}^{-2}\mathbf{B}^\top\mathbf{X}^\top) \\
&\quad + 2\text{traza}(\mathbf{X}\mathbf{B}\mathbf{\Lambda}^{-2}\mathbf{B}^\top\mathbf{B}\mathbf{\Lambda}^{-2}\mathbf{B}^\top\mathbf{X}^\top) \\
&= \text{traza}(\mathbf{X}\mathbf{X}^\top) - 2\text{traza}(\mathbf{X}\mathbf{B}\mathbf{\Lambda}^{-2}\mathbf{B}^\top\mathbf{X}^\top) \\
&\quad + 2\text{traza}(\mathbf{X}\mathbf{B}\mathbf{\Lambda}^{-2}\mathbf{B}^\top\mathbf{X}^\top) \\
&= \text{traza}(\mathbf{X}\mathbf{X}^\top)
\end{aligned}$$

□

Por tanto, minimizar  $F$  es lo mismo que maximizar  $\|\mathbf{A}\mathbf{\Lambda}^{-1}\mathbf{B}^\top\|^2$  que corresponde con la variabilidad explicada por la aproximación de  $\mathbf{X}$  dada por la descomposición HJ Biplot.

En el algoritmo que se explica a continuación se fija como función objetivo la maximización de la variabilidad explicada por la aproximación. Los criterios de parada se asumen igual que en la sección anterior.

### 5.2.3. Algoritmo DBiplot

A continuación se explica el algoritmo creado para llevar a cabo el Disjoint Biplot. El objetivo de este método es encontrar  $Q$  componentes que definan un nuevo espacio donde representar los datos de partida pero con la peculiaridad de que cada variable sólomente va a contribuir a una componente. Se parte de  $\mathbf{X}$  matriz de datos de orden  $I \times J$  que contiene la información de  $I$  individuos sobre los que se han medido  $J$  variables.

El algoritmo consta de un paso inicial que se explica a continuación:

Se genera la matriz  $\mathbf{V}_0$  aleatoria binaria de orden  $J \times Q$  que contiene la información relativa a la componente sobre la que va a contribuir cada una de las  $J$  variables. Es decir,  $v_{jq} = 1$  si la variable  $j$  contribuye a la componente  $q$  y  $v_{jq} = 0$  en otro caso. Al igual que sucedía con la matriz  $\mathbf{U}_0$  en el método descrito anteriormente, es una matriz estocástica por filas, es decir, todos los elementos son no negativos y  $\sum_{t=1}^Q v_{jt} = 1 \forall j = 1, \dots, J$ . También se asegura que todas las componentes tengan cargas de alguna variable, es decir, que ninguna componente tenga cargas nulas para todas las variables. El procedimiento es similar al caso de los clusters. Si una componente no tiene ninguna variable asignada para que contribuya a su conformación, aquella que tenga mayor número de variables se divide aleatoriamente en dos.

Partiendo de las matrices  $\mathbf{X}$  y  $\mathbf{V}_0$  se van a calcular las coordenadas para las variables  $\mathbf{B}_0$  de orden  $J \times Q$  de la siguiente manera:

Para cada componente  $q$ :

Se genera la submatriz  $\mathbf{W}_{0q}$  que contiene tantas filas como la matriz de partida  $\mathbf{X}$  y tantas columnas como variables que contribuyan a dicha componente  $q$ , es decir, se seleccionan las columnas de la matriz  $\mathbf{X}$  para las que la variable asociada tenga un uno en la columna  $q$  de la matriz  $\mathbf{V}_0$ .

La matriz  $\mathbf{W}_{0q}$  se descompone en valores singulares.

$$\mathbf{W}_{0q} = \mathbf{R}\tilde{\mathbf{\Lambda}}\mathbf{T}^\top$$

y se calculan las coordenadas para dichas variables como  $\tilde{\mathbf{B}}_{0q} = \mathbf{T}\tilde{\mathbf{\Lambda}}$ . Se elige la primera columna de esta matriz que corresponde con la componente asociada al primer valor propio para este subconjunto de variables. Estas coordenadas van a ser las coordenadas de las variables que se han considerado para la componente  $q$  en el nuevo espacio y el resto de variables que no han formado parte de la submatriz  $\mathbf{W}_{0q}$  tendrán coordenada 0 para esta componente.

En la figura 5.2 se muestra cómo se construye  $\mathbf{W}_q$  a partir de las matrices  $\mathbf{X}$  y  $\mathbf{V}$ .

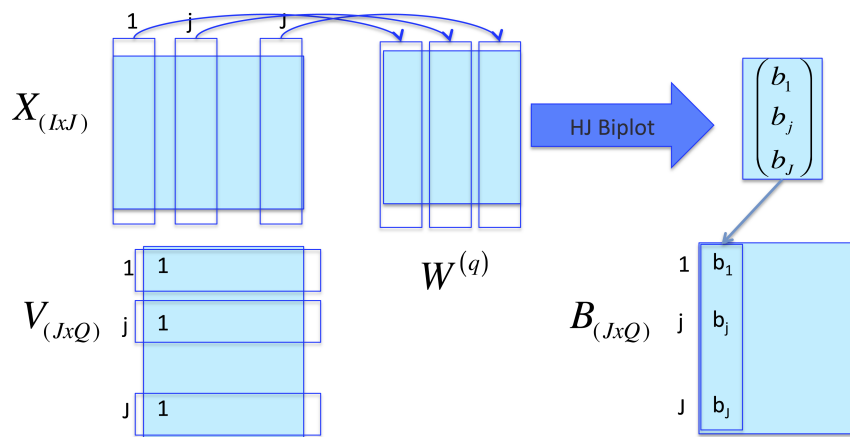


Figura 5.2: Esquema de construcción de la submatriz  $\mathbf{W}_q$ .

Una vez que se tiene la matriz  $\mathbf{B}_0$  de coordenadas para las variables, se calcula la matriz de coordenadas para los individuos:

$$\mathbf{A}_0 = \mathbf{X}\mathbf{B}_0\mathbf{\Lambda}_0^{-1}$$

donde  $\mathbf{\Lambda}_0$  contiene el mayor valor propio resultante de las dvs de cada  $\mathbf{W}_{0q}$ .

Con las coordenadas calculadas en este paso se computa el valor inicial de la función objetivo  $F_0$ .

Una vez terminado el paso inicial, se comienza con el proceso iterativo.

Se supone que se dispone de los resultados obtenidos en el paso  $k - 1$ , por tanto se tienen las matrices  $\mathbf{V}_{k-1}$ ,  $\mathbf{A}_{k-1}$ ,  $\mathbf{\Lambda}_{k-1}$  y  $\mathbf{B}_{k-1}$  y el valor de la función objetivo en dicho paso  $F_{k-1}$ . A partir de ellas se va a realizar el paso  $k$ .

En este paso, es necesario actualizar la matriz  $\mathbf{V}_k$ . Para ello se va a actuar por filas. Para cada fila  $j$  se procede de la siguiente manera:

Se mantienen las filas  $t = 1, \dots, j - 1, j + 1, \dots, J$ . Se ponen a cero toda la fila  $j$  y se crea  $\tilde{\mathbf{V}}_{kq}$  en las que el uno de dicha fila se encuentra en la columna  $q$ . Para cada una de las  $q$   $\tilde{\mathbf{V}}_{kq}$  se calculan las matrices de coordenadas para individuos y variables  $\tilde{\mathbf{A}}_{kq}$  y  $\tilde{\mathbf{B}}_{kq}$  y se evalúa la función objetivo  $\tilde{F}_{kq}$ . Aquella posición para la cual la  $\tilde{F}_{kq}$  sea máxima, será la posición para la cual se fije el uno de dicha fila.

Si de este proceso resulta alguna columna de la matriz  $\mathbf{V}_k$  con todos los elementos nulos, se procede de igual forma que se explicó para  $\mathbf{V}_0$ .

Una vez que se ha actualizado la matriz  $\mathbf{V}_k$  se calculan las coordenadas para las variables  $\mathbf{B}_k$ . El procedimiento es similar al explicado en el paso inicial.

Para cada componente  $q$ :

Se genera la submatriz  $\mathbf{W}_{kq}$  que contiene las columnas de la matriz  $\mathbf{X}$  para las que la variable asociada tenga un uno en la columna  $q$  de la matriz  $\mathbf{V}_k$  actualizada anteriormente.

La matriz  $\mathbf{W}_{kq}$  se descompone en valores singulares.

$$\mathbf{W}_{kq} = \mathbf{R}\tilde{\mathbf{\Lambda}}\mathbf{T}^\top$$

y se calculan las coordenadas para dichas variables como  $\tilde{\mathbf{B}}_{kq} = \mathbf{T}\tilde{\mathbf{\Lambda}}$ . Se elige la primera columna de esta matriz como coordenadas de las variables que se han



considerado para la componente  $q$  en el nuevo espacio y el resto de variables que no han formado parte de la submatriz  $\mathbf{W}_{kq}$  tendrán coordenada 0 para esta componente.

Con la matriz  $\mathbf{B}_k$  se calcula la matriz de coordenadas de individuos  $\mathbf{A}_k$ :

$$\mathbf{A}_k = \mathbf{X}\mathbf{B}_k\mathbf{\Lambda}_k^{-1}$$

donde  $\mathbf{\Lambda}_k$  contiene el mayor valor propio resultante de las dvs de cada  $\mathbf{W}_{kq}$ .

Por último, se calcula el valor correspondiente de la función objetivo  $F_k$ . Se comprueba si se verifican los criterios de parada explicados al comienzo de esta sección. En este caso, se termina el proceso iterativo y se considera como solución las coordenadas obtenidas en este paso, si no, se repite este paso hasta que se cumplan los criterios de parada.

En la figura 5.3 se esquematizan los pasos básicos del algoritmo Disjoint Biplot.

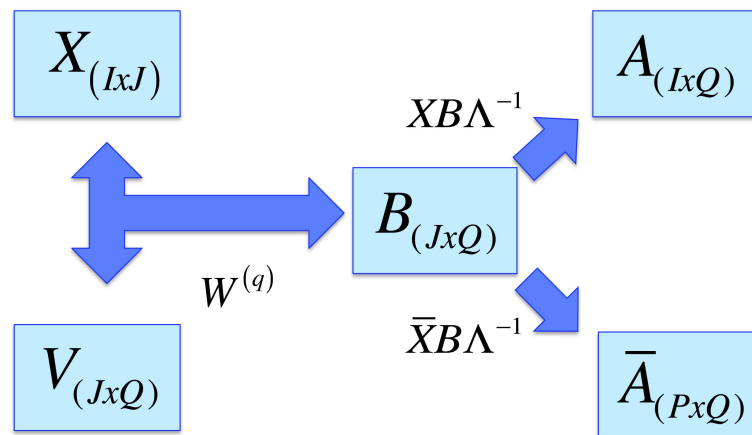


Figura 5.3: Esquema del algoritmo DBiplot.

### 5.3. Clustering Disjoint Biplot

En las secciones anteriores se han explicado diferentes métodos tanto para buscar una estructura subyacente de grupos como para mejorar la interpretabilidad de los ejes factoriales extraídos mediante técnicas de reducción de la dimensión. Sin embargo, hay autores que intentan buscar una solución conjunta a ambos tipos de problemas. El análisis de clusters tradicionalmente se utiliza para la clasificación de objetos. Sin embargo, hay ocasiones en las que se utilizan para buscar una partición de las variables en lugar de las técnicas factoriales. Esto es debido a que en las técnicas factoriales, incluso después de haber aplicado algún tipo de rotación, los factores extraídos pueden compartir variables en su conformación, hecho que dificulta su interpretación. Más relacionado con la idea de buscar una clasificación de objetos y una partición de variables que mejore la interpretación de los resultados, se encuentran las técnicas de biclustering (Hartigan, 1972, 1975). Estas técnicas identifican submatrices de la matriz de partida en las que los objetos y las variables forman clusters. Sin embargo, estos métodos no tienen claramente definido un problema de optimización mediante una función objetivo que mida las discrepancias entre las aproximaciones que se encuentran en el proceso iterativo y los datos originales. En este sentido, Bock, 1979, 2003; DeSarbo, 1982 proponen alternativas al biclustering basadas en diferentes criterios de optimización. Vichi, 2000 presenta el algoritmo Doble k-means. Una revisión completa de los métodos de biclustering se puede ver en Van Mechelen et al., 2004. Witten y Tibshirani, 2010 proponen el *Sparse k-means Clustering*, cuyo doble objetivo es encontrar una clasificación de los objetos mediante el algoritmo k-means y buscar las variables que tengan más influencia en dicha clasificación. Para ello utilizan la penalización *lasso* (Tibshirani, 1996).

En Vichi y Saporta, 2009 se propone un método iterativo que busca la partición de los objetos alrededor de un conjunto de centroides y la de las variables en torno a un conjunto de componentes. Dicho método se denomina CDPCA (*Clustering Disjoint Principal Component Analysis*). Este método tiene la ventaja, respecto del sparse PCA, de obtener particiones de las variables disjuntas, es decir, cada variable del conjunto de datos únicamente va a contribuir a la conformación de una componente, hecho que no ocurre en el sparse PCA ya que aunque se obtienen matriz de cargas con bastantes ceros, una misma variable puede tener cargas para varias componentes. Esta forma de obtener las componentes disjuntas está relacionada con el método propuesto por Vigneau y Qannari, 2004. Recientemente, Macedo y Freitas, 2015 desarrollan un programa para su utilización.

Teniendo en cuenta las aportaciones anteriores, se propone el algoritmo CDBiplot (*Clustering Disjoint Biplot*), que alterna la búsqueda de la mejor clasificación de los objetos mediante la representación HJ Biplot (Galindo, 1986) en un espacio de dimensión reducida donde los ejes factoriales extraídos son disjuntos, es decir, cada variable de la matriz de partida sólo contribuye a la conformación de un eje teniendo contribuciones nulas para el resto. Esto se consigue dividiendo el espacio total en subespacios disjuntos y extrayendo de cada uno de ellos la dirección de máxima variabilidad.

### 5.3.1. Notación

A continuación, se expone brevemente la terminología utilizada en la descripción del algoritmo:

- **I,J,P,Q.** Número de objetos, variables, clusters y componentes respectivamente.

- $\mathbf{X}_{ij}$ . Matriz de datos que contiene la información de  $I$  objetos sobre los que se han medido  $J$  variables. Se supone estandarizada.
- $\mathbf{E}_{ij}$ . Matriz de errores.
- $\mathbf{U}_{ip}$ . Matriz binaria que contiene la localización de los  $I$  individuos en los  $P$  clusters. Es decir,  $u_{ip} = 1$  si el individuo  $i$  pertenece al cluster  $p$  y  $u_{ip} = 0$  en otro caso. Es una matriz estocástica por filas, es decir, todos los elementos son no negativos y  $\sum_{t=1}^P u_{it} = 1 \forall i = 1, \dots, I$ .
- $\bar{\mathbf{X}}_{pj}$ . Matriz de centroides en el espacio original.
- $\mathbf{Z}_{ij}$ . Matriz en la que se ha sustituido el elemento original por el centroide del cluster al que pertenece.
- $\mathbf{V}_{jq}$ . Matriz binaria que contiene la información relativa a la componente sobre la que va a contribuir cada una de las  $J$  variables. Es decir,  $v_{jq} = 1$  si la variable  $j$  contribuye a la componente  $q$  y  $v_{jq} = 0$  en otro caso. Es una matriz estocástica por filas, es decir, todos los elementos son no negativos y  $\sum_{t=1}^Q v_{jt} = 1 \forall j = 1, \dots, J$ .
- $\mathbf{A}_{iq}$ . Matriz que contiene las coordenadas HJ Biplot de los objetos en el espacio de las  $Q$  componentes disjuntas.
- $\bar{\mathbf{A}}_{pq}$ . Matriz que contiene las coordenadas HJ Biplot de los centroides de los  $P$  clusters en el espacio de las  $Q$  componentes disjuntas.
- $\mathbf{B}_{jq}$ . Matriz que contiene las coordenadas HJ Biplot de las  $J$  variables en el espacio de las  $Q$  componentes disjuntas.
- $\mathbf{\Lambda}$ . Matriz diagonal que contiene los valores propios de la dvs disjunta de  $\mathbf{Z}$ .

### 5.3.2. Modelo Clustering Disjoint Biplot

El modelo para el CDBiplot y las consideraciones para la función objetivo son los explicados en la sección 5.1.2 pero teniendo en cuenta que la obtención de los marcadores para las variables siguen un proceso diferente al establecido en el algoritmo CBiplot.

### 5.3.3. Algoritmo CDBiplot

En esta sección se detalla el algoritmo Clustering Disjoint Biplot. Como se ha explicado anteriormente, este método tiene un doble objetivo: por un lado busca las componentes que maximicen la distancia entre los centroides de los  $P$  clusters que se buscan en los  $I$  individuos y por otro lado, se buscan  $Q$  componentes de tal forma que las  $J$  variables sólo contribuyan a una de las componentes encontradas.

Como criterios de parada en el proceso iterativo, se fijan los mismos que en las secciones anteriores.

Se parte de  $\mathbf{X}$ , una matriz de datos de orden  $I \times J$  que contiene la información de  $I$  individuos sobre los que se han medido  $J$  variables.

El algoritmo consta de un paso inicial que a su vez se compone de dos partes:

- Referente al primer objetivo (cluster de individuos):

Se crea la matriz  $\mathbf{U}_0$  aleatoria binaria de orden  $I \times P$  que contiene la localización de los  $I$  individuos en los  $P$  clusters. Es decir,  $u_{ip} = 1$  si el individuo  $i$  pertenece al cluster  $p$  y  $u_{ip} = 0$  en otro caso. Es una matriz estocástica por filas tal que todos los elementos son no negativos y  $\sum_{t=1}^P u_{it} = 1 \forall i = 1, \dots, I$ . Como en el algoritmo CBiplot, se asegura que todos los clusters tengan elementos dividiendo el que tenga mayor número en caso de que fuera necesario.

Mediante las matrices  $\mathbf{X}$  y  $\mathbf{U}_0$  se genera la matriz de centroides  $\bar{\mathbf{X}}_0$  de orden  $P \times J$ :

$$\bar{\mathbf{X}}_0 = (\mathbf{U}_0^\top \mathbf{U}_0)^{-1} \mathbf{U}_0^\top \mathbf{X}.$$

Con esta matriz y  $\mathbf{U}_0$ , se calcula  $\mathbf{Z}_0$ . Esta matriz contiene los valores de los centroides de los clusters a los que pertenece cada individuo en lugar de los valores iniciales de  $\mathbf{X}$ .

$$\mathbf{Z}_0 = \mathbf{U}_0 \bar{\mathbf{X}}_0.$$

- Referente al segundo objetivo (variables que sólo contribuyen a una componente):

Se genera la matriz  $\mathbf{V}_0$  aleatoria binaria de orden  $J \times Q$  que informa de la componente a la que contribuye cada variable. Es decir,  $v_{jq} = 1$  si la variable  $j$  contribuye a la componente  $q$  y  $v_{jq} = 0$  en otro caso. Cumple que todos los elementos son no negativos y  $\sum_{t=1}^Q v_{jt} = 1 \forall j = 1, \dots, J$ . Como se explicó en el algoritmo DBiplot, se asegura que todas las componentes tengan alguna variable que contribuya a ella reasignando la mitad de las que contribuyan a la componente con mayor número de cargas no nulas.

Partiendo de las matrices  $\mathbf{Z}_0$  y  $\mathbf{V}_0$  se calculan las coordenadas para las variables  $\mathbf{B}_0$  mediante el siguiente procedimiento:

Para cada componente  $q$ :

Se extrae la submatriz  $\mathbf{W}_{0q}$  de  $\mathbf{Z}_0$  formada por las columnas de la matriz  $\mathbf{Z}_0$  para las que la variable asociada tenga un uno en la columna  $q$  de la matriz  $\mathbf{V}_0$ .

Se realiza la dvs de la matriz  $\mathbf{W}_{0q} = \mathbf{R}\tilde{\mathbf{\Lambda}}\mathbf{T}^\top$  y se calculan las coordenadas para dichas variables como  $\tilde{\mathbf{B}}_{0q} = \mathbf{T}\tilde{\mathbf{\Lambda}}$ . Se selecciona la primera columna de esta matriz como coordenadas de las variables que se han considerado para la componente  $q$  en el nuevo espacio y al resto de variables que no han formado parte de la submatriz  $\mathbf{W}_{0q}$  se les asigna 0 para dicha componente. Una vez obtenida la matriz  $\mathbf{B}_0$ , se obtiene la matriz de coordenadas para los individuos:

$$\mathbf{A}_0 = \mathbf{X}\mathbf{B}_0\mathbf{\Lambda}_0^{-1}$$

donde  $\mathbf{\Lambda}_0$  contiene el mayor valor propio resultante de las dvs de cada  $\mathbf{W}_{0q}$ , al igual que se explicó en el método Disjoint Biplot.

También se obtiene la matriz de coordenadas de los centroides en el nuevo espacio de representación.

$$\bar{\mathbf{A}}_0 = \bar{\mathbf{X}}\mathbf{B}_0\mathbf{\Lambda}_0^{-1}.$$

Con las matrices  $\mathbf{A}_0$ ,  $\bar{\mathbf{A}}_0$ ,  $\mathbf{U}_0$ ,  $\mathbf{B}_0$  y  $\mathbf{\Lambda}_0^{-1}$  se obtiene el valor inicial de la función objetivo  $F_0$ .

A partir de este punto se comienza con el proceso iterativo.

Tal como se ha explicado en los algoritmos anteriores, se suponen conocidos los resultados obtenidos en el paso  $k-1$ , por tanto se tienen las matrices  $\mathbf{U}_{k-1}$ ,  $\bar{\mathbf{X}}_{k-1}$ ,  $\mathbf{Z}_{k-1}$ ,  $\mathbf{A}_{k-1}$ ,  $\bar{\mathbf{A}}_{k-1}$ ,  $\mathbf{V}_{k-1}$ ,  $\mathbf{V}_{k-1}$ ,  $\mathbf{\Lambda}_{k-1}$  y  $\mathbf{B}_{k-1}$  y el valor de la función objetivo en dicho paso  $F_{k-1}$ . Con todas ellas se van a calcular las correspondientes al paso  $k$ .

- Referente al primer objetivo (cluster de individuos):

En esta parte se actualiza la matriz  $\mathbf{U}_k$  a partir de las coordenadas de los

individuos y de los centroides  $\mathbf{A}_{k-1}$  y  $\bar{\mathbf{A}}_{k-1}$  mediante el algoritmo de k-means. Se asegura que todas las columnas de  $\mathbf{U}_k$  tengan algún elemento no nulo como se explicó para  $\mathbf{U}_0$ .

Con dicha matriz y los datos de partida  $\mathbf{X}$  se calculan los centroides  $\bar{\mathbf{X}}_k = (\mathbf{U}_k^\top \mathbf{U}_k)^{-1} \mathbf{U}_k^\top \mathbf{X}$ . Con ella y  $\mathbf{U}_k$  se actualiza  $\mathbf{Z}_k = \mathbf{U}_k \bar{\mathbf{X}}_k$ .

- Referente al segundo objetivo (variables que sólo contribuyen a una componente):

En esta parte se actualiza la matriz  $\mathbf{V}_k$ . Para ello se va a actuar por filas. Para cada fila  $j$  se sigue el siguiente proceso:

Se fijan las filas  $t = 1, \dots, j-1, j+1, \dots, J$ . Se ponen ceros en todas las posiciones de la fila  $j$  y se genera la matriz  $\tilde{\mathbf{V}}_{kq}$  para cada una de las componentes  $Q$  en las que se posiciona el uno en el lugar  $q$ -ésimo. A partir de cada una de las  $q$   $\tilde{\mathbf{V}}_{kq}$  se calculan las matrices de coordenadas para individuos, centroides y variables  $\tilde{\mathbf{A}}_{kq}$ ,  $\tilde{\bar{\mathbf{A}}}_{kq}$  y  $\tilde{\mathbf{B}}_{kq}$  y se calcula el valor de la función objetivo  $\tilde{F}_{kq}$ . Aquella posición que haga máxima  $\tilde{F}_{kq}$ , será la posición para la cual se fije el uno de la fila  $j$ .

Una vez que se ha actualizado la matriz  $\mathbf{V}_k$  con la seguridad de que todas las columnas tienen algún elemento no nulo, se obtienen las coordenadas para las variables  $\mathbf{B}_k$ . El procedimiento es el explicado en el paso inicial.

Para cada componente  $q$ :

Se genera la submatriz  $\mathbf{W}_{kq}$  que contiene las columnas de la matriz  $\mathbf{Z}_k$  para las que la variable asociada tenga un uno en la columna  $q$  de la matriz  $\mathbf{V}_k$  actualizada anteriormente.



La matriz  $\mathbf{W}_{kq}$  se descompone en valores singulares.

$$\mathbf{W}_{kq} = \mathbf{R}\tilde{\mathbf{\Lambda}}\mathbf{T}^\top$$

y se calculan las coordenadas para dichas variables como  $\tilde{\mathbf{B}}_{kq} = \mathbf{T}\tilde{\mathbf{\Lambda}}$ . Se escoge la primera columna de esta matriz como coordenadas de las variables que se han considerado para la componente  $q$  en el nuevo espacio y el resto de variables que no han formado parte de la submatriz  $\mathbf{W}_{kq}$  tendrán coordenada 0 para esta componente.

Con la matriz  $\mathbf{B}_k$  se calcula la matriz de coordenadas de individuos  $\mathbf{A}_k$ :

$$\mathbf{A}_k = \mathbf{X}\mathbf{B}_k\mathbf{\Lambda}_k^{-1}$$

donde  $\mathbf{\Lambda}_k$  contiene el mayor valor propio resultante de las dvs de cada  $\mathbf{W}_{kq}$ .

También es necesario la matriz de coordenadas de los centroides:

$$\bar{\mathbf{A}}_k = \bar{\mathbf{X}}\mathbf{B}_k\mathbf{\Lambda}_k^{-1}$$

Por último, se calcula el valor de la función objetivo  $F_k$ . Se comprueba si se verifican los criterios de parada anteriormente explicados. Si se cumplen, se termina el proceso iterativo y se considera como solución las coordenadas obtenidas en este paso, si no, se repite este paso hasta que se cumplan los criterios de parada.

Con el objetivo de no caer en un mínimo local, es necesario ejecutar el algoritmo varias veces. Se recomienda un mínimo de 1000 para encontrar una solución estable.

En la figura 5.4 se muestran el esquema del algoritmo Clustering Disjoint

Biplot.

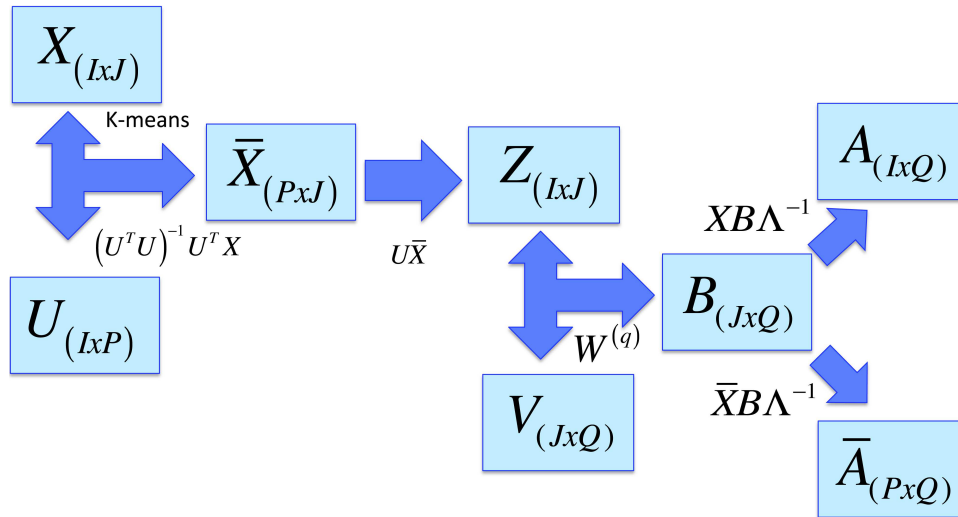


Figura 5.4: Esquema del algoritmo CDBiplot.

## 5.4. Programa *CDBiplot*

Con el objetivo de proporcionar un software que permita la utilización de los tres algoritmos descritos anteriormente, se ha desarrollado una nueva GUI en el entorno R. Este programa permite al usuario elegir el algoritmo que desea utilizar con sus datos y obtener un gráfico con la representación resultante.

El programa se encuentra en forma de función dentro del paquete `biplotbootGUI`. La función se denomina `CDBiplot` y consta de dos argumentos:

- **data.** Matriz que contiene los datos que van a ser analizados.
- **clase.** Vector que contiene la clase a la que pertenece cada objeto en caso de que se conozca y se quiera analizar la clasificación de los objetos.

Una vez que se invoca la función, aparece una ventana (figura 5.5) con tres radiobuttons para poder elegir el algoritmo a ejecutar.

- Clustering Biplot (CBiplot).
- Disjoint Biplot (DBiplot).
- Clustering Disjoint Biplot (CDBiplot).

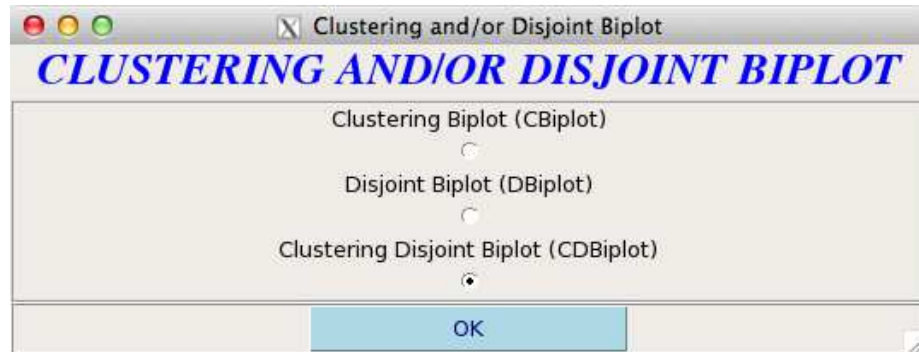


Figura 5.5: Ventana principal de la función CDBiplot.

Una vez elegido el algoritmo, emerge una ventana (figura 5.6) donde se debe introducir el número de componentes, el umbral de tolerancia permitida, el número máximo de iteraciones que va a ejecutar el algoritmo y el número de veces que se quiere repetir el algoritmo completo con el objetivo de encontrar la solución óptima y no caer en un mínimo local. Si los algoritmos que se van a utilizar requieren la clasificación de los objetos, también se pide el número de clusters que se quiere considerar.

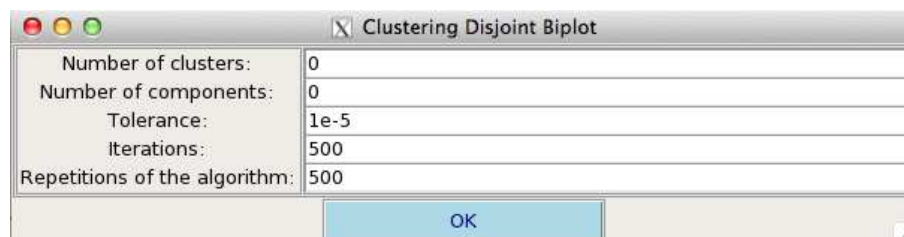


Figura 5.6: Ventana para elegir los parámetros para ejecutar el algoritmo.

A continuación, el programa ejecuta el algoritmo y cuando termina, aparece una ventana (figura 5.7) con la representación de los objetos y las variables analizadas mediante el algoritmo elegido.

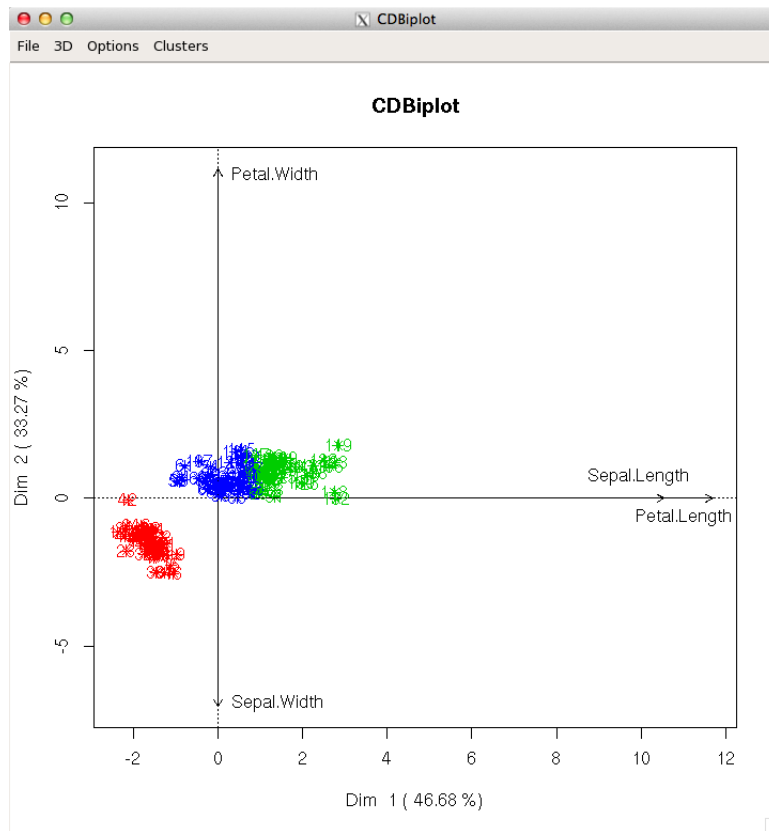


Figura 5.7: Ventana con la representación conjunta de objetos y variables.

En esta ventana se pueden distinguir dos bloques y varios menús. En el bloque de la derecha se tiene la representación gráfica de los objetos y las variables. En este gráfico es posible mover las etiquetas de los puntos con el botón izquierdo del ratón. El bloque de la izquierda sólo aparece si el número de componentes es mayor que dos. En ese caso, se pueden observar tres listbox que sirven para elegir la dimensión que se quiere ver en cada gráfico (2 y 3 dimensiones) y un botón **Refresh** que es necesario pulsar para que los cambios de los anteriores cuadros de texto se actualicen. En la parte superior de la ventana se tienen 4 menús:

- File
  - Copy image
  - Save image

- PDF file
- Eps file
- Png file
- Jpg/Jpeg file
- Exit
- 3D
  - 3D
- Options
  - Change title
  - Show/Hide axes
  - Show/Hide variables
  - Show/Hide row labels
- Cluster
  - Convex-hull

El primero de ellos permite copiar la imagen al portapapeles, guardarla en varios tipos de formato y salir del programa. El siguiente es el menú de 3 dimensiones. Permite ver las coordenadas de objetos y variables en tres dimensiones. En este gráfico existe la posibilidad de rotar la imagen con el botón izquierdo del ratón y ampliar o disminuir la imagen con el botón derecho del ratón. El tercer menú disponible es el de opciones. Contiene cuatro submenús: cambiar el título del gráfico, mostrar o quitar los ejes de coordenadas, mostrar o quitar la representación de las variables y mostrar o quitar las etiquetas de los



Figura 5.8: Ventana para elegir las opciones respecto del polígono envolvente.

objetos. El último menú permite al usuario dibujar convex-hull alrededor de cada cluster. Cuando se elige este submenú aparece una ventana (figura 5.8).

En ella se puede elegir si se dibujan los polígonos envolventes (convex-hull) y en caso de querer representarlos, es posible elegir dibujar sólo la línea exterior o rellenar toda la zona delimitada por el polígono.

Junto a esta ventana emerge un texto con los resultados del análisis:

- Matriz  $U$  de clasificación de los objetos.
- Matriz  $V$  de información sobre a qué componente contribuye cada variable.
- Valores propios.
- Variabilidad explicada por cada componente.
- Coordenadas de objetos y variables.
- Matriz de pseudoconfusión en el caso de haber pasado el argumento `clase` al invocar la función y haber elegido un algoritmo que implique clasificación de los objetos. Esta matriz es una tabla que cruza el número de objetos clasificados mediante el algoritmo con la clasificación introducida.
- Matriz de correlaciones entre componentes.
- Número máximo de iteraciones necesarias para la convergencia del algoritmo.

- Valor máximo de la función objetivo  $F$ .

Dicho archivo se guarda automáticamente en el directorio en el que se encuentre el usuario.

## 5.5. Aplicación a Datos

Para mostrar el funcionamiento del software y comprobar las ventajas de los algoritmos descritos anteriormente se ejecuta el programa sobre dos conjuntos de datos.

En primer lugar, se va a ejecutar el método CDBiplot sobre los datos iris (Anderson, 1935) que ya se explicaron en la sección 2.5. Como se veía en dicha sección, a pesar de que los datos están clasificados en tres grupos, el HJ Biplot en las dos primeras dimensiones (figura 5.9) no diferencia los tres grupos. Únicamente se aprecia gran diferencia entre el grupo *setosa*, representado en rojo, de los grupos *versicolor* y *virginica* (verde y naranja). Pero estos dos grupos se solapan en una región y por tanto, no se ve una separación clara.

Si se aplica el algoritmo CDBiplot a estos datos se obtiene la figura 5.10.

Como se puede observar, se obtiene una separación entre el primer cluster y el resto. También se observa una separación más clara en los dos clusters que no se diferenciaban en el HJ Biplot anterior.

A continuación se presenta el gráfico con la representación conjunta de individuos y variables (figura 5.11). En él se puede ver que las variables resultantes de la separación entre clusters son las longitudes del sépalo y del pétalo mientras que la anchura del sépalo y del pétalo separan el primero del resto.

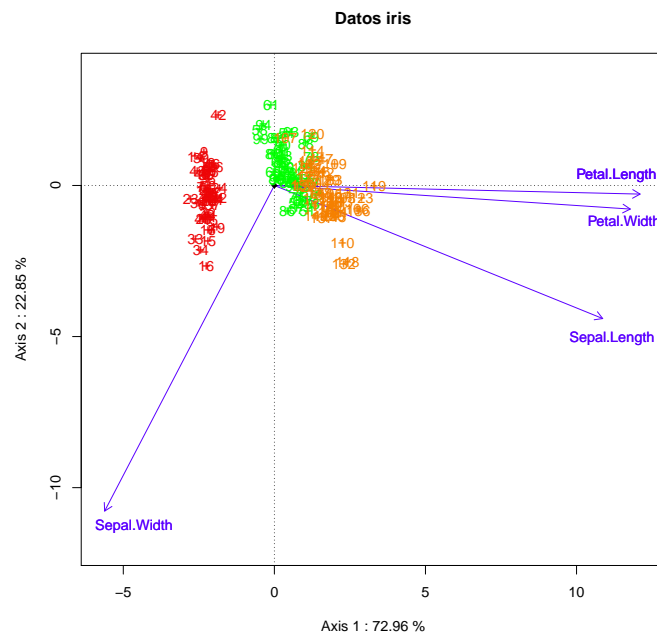


Figura 5.9: Gráfico que muestra la representación HJ Biplot para los datos iris en las dos primeras dimensiones.

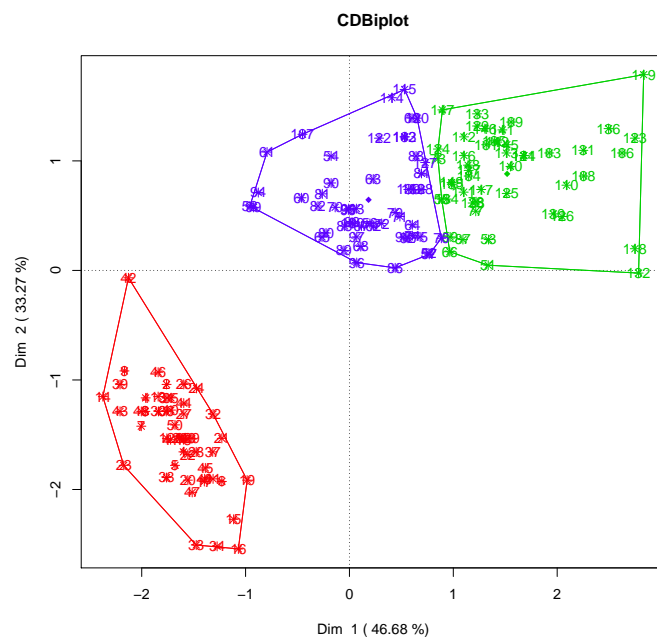


Figura 5.10: Gráfico que muestra los clusters obtenidos mediante el algoritmo CDBiplot en las dos primeras dimensiones.



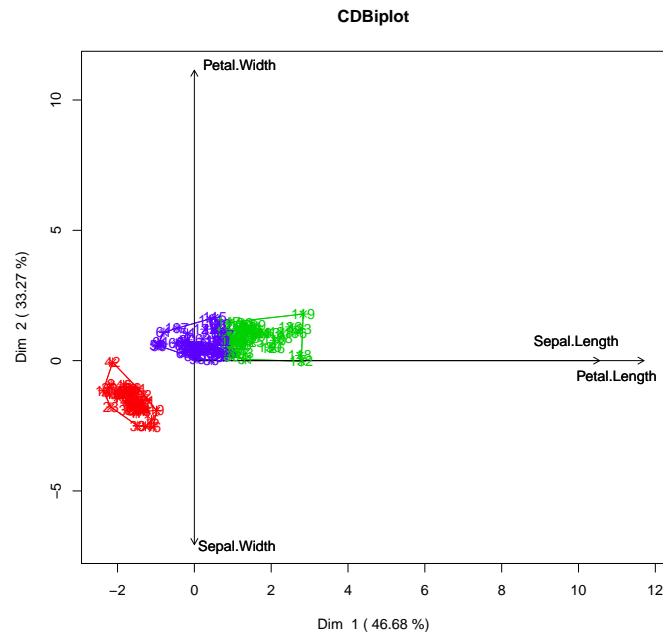


Figura 5.11: Gráfico que muestra los clusters y las variables obtenidas mediante el algoritmo CDBiplot en las dos primeras dimensiones.

Hay que tener en cuenta que la variabilidad explicada por cada eje en este algoritmo es menor que la que explica el HJ Biplot ya que el objetivo del CDBiplot no es sólo maximizar la variabilidad explicada por la aproximación sino buscar la mejor clasificación de los individuos. Además, los ejes extraídos mediante el CDBiplot son *disjoint*, es decir, cada variable sólo contribuye a la conformación de un único eje y este hecho también influye en la disminución de la variabilidad explicada por cada eje. El programa proporciona, además del gráfico, una tabla cruzada donde se puede ver el número de casos bien y mal clasificados en el caso de que se haya incluido como parámetro de llamada a la función un vector con la clasificación real de los individuos. En este caso, se observa que el grupo *setosa* se clasifica correctamente mientras que en el grupo *versicolor* y en el *virginica* existen individuos mal clasificados. Dichos individuos suponen un 13 % del total de la muestra.

Otra cuestión que hay que tener en cuenta es que, por la manera de calcularlos,

los ejes extraídos no tienen correlación 0. Por ello, el programa proporciona la matriz de correlaciones para los ejes factoriales. Para estos datos se obtiene una correlación entre ambos ejes del 81%. Cabe destacar también que, para estos datos, el número máximo de iteraciones que ha necesitado el algoritmo para alcanzar la convergencia ha sido de 14.

El segundo conjunto de datos que se utiliza con este algoritmo son datos disponibles en la página <http://archive.ics.uci.edu/ml/datasets/Wine+Quality>. La muestra original consta de 2 conjuntos de datos relativos a dos variedades (tinto y blanco) de *Vinho Verde* portugués (Cortez et al., 2009). Sobre ellos se han medido las siguientes variables fisico-químicas: acidez fija, acidez volátil, ácido cítrico, azúcar residual, cloruros, dióxido de azufre libre, dióxido de azufre total, densidad, pH, sulfatos y alcohol. Los datos contienen 1599 mediciones de la variedad tinta y 4898 de la variedad blanca. Como ejemplo, se ha extraído una muestra que contiene 500 mediciones de cada una de ellas. Se utiliza el algoritmo CDBiplot para encontrar la representación de la clasificación en ambas variedades y obtener dos ejes factoriales disjuntos.

Al igual que en el ejemplo anterior, se realiza primero un HJ Biplot sobre los datos. En la figura 5.12 se puede ver dicha representación. Según se observa, hay dos grupos pero no están bien diferenciados. La cantidad de variabilidad absorbida por los dos ejes es del 51,10%. En esta representación se aprecia que las variables que más contribuyen a la separación de los dos grupos son la acidez volátil, azúcar residual, dióxido de azufre total y dióxido de azufre libre.

A continuación, se ejecuta el algoritmo CDBiplot sobre los mismos datos. En la figuras 5.13 y 5.14 se presentan los resultados reteniendo dos ejes factoriales.

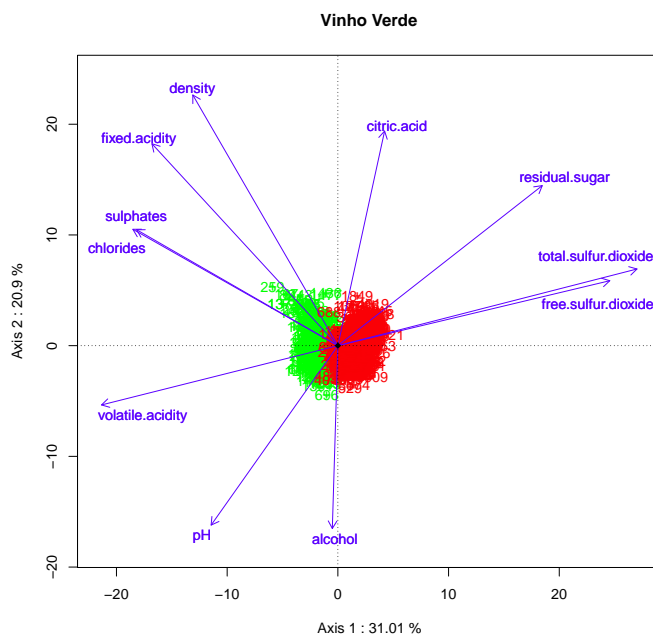


Figura 5.12: Gráfico que muestra la representación HJ Biplot para los datos Vinho Verde en las dos primeras dimensiones.

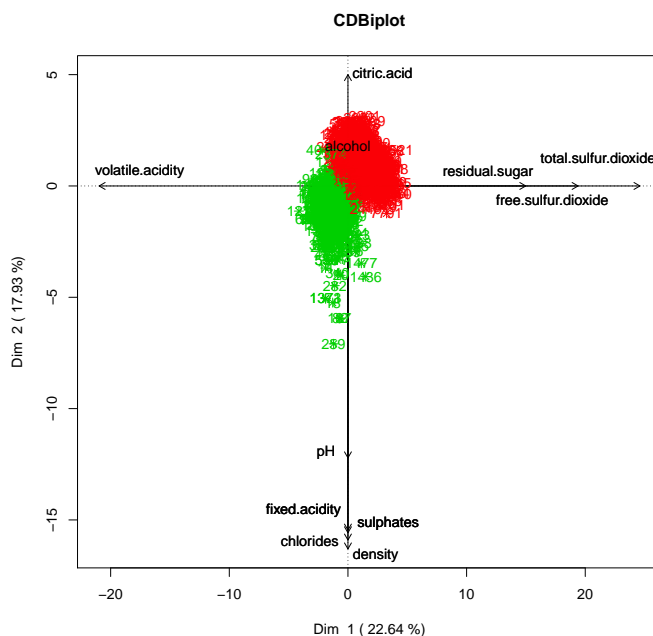


Figura 5.13: Gráfico que muestra los clusters y las variables obtenidas mediante el algoritmo CDBiplot en las dos primeras dimensiones. Datos Vinho Verde.



---

considerados. A partir de esta revisión se proponen tres algoritmos: CBiplot, que busca el mejor espacio de representación común reducida teniendo en cuenta la estructura de grupos de los individuos; DBiplot, que busca un espacio de dimensión reducida para representar individuos y variables con la ventaja de obtener dimensiones disjuntas en las que cada variable sólo contribuye a la conformación de un eje; CDBiplot, que combina ambos algoritmos para conseguir simultáneamente las dos ventajas de los métodos anteriores. Para que los tres algoritmos sean fácilmente utilizables en la práctica, se ha desarrollado un programa denominado `CDBiplot` en el entorno `R` que permite obtener los resultados de una forma interactiva mediante el uso de ventanas y botones. Por último, a través de dos conjuntos de datos reales se ha ilustrado su utilización y la interpretación de los resultados obtenidos.



# CONCLUSIONES





1. De la exhaustiva revisión bibliográfica se deduce que los métodos Bootstrap han tenido un gran desarrollo desde 1979 en que aparecieron. Se han aplicado en análisis Factorial, análisis de Componentes Principales, análisis de Correspondencias, análisis no Simétrico de Correspondencias, rotaciones Procrustes, modelos Multinivel, modelos de TUCKER y análisis Factorial Múltiple, pero no se ha encontrado ninguna referencia en revistas científicas que utilice estos métodos para proporcionar versiones inferenciales de los métodos Biplot.
2. Se han desarrollado versiones inferenciales basadas en los métodos Bootstrap de los métodos Biplot tanto para los clásicos de Gabriel, 1971 como para el HJ Biplot de Galindo, 1986 con su correspondiente programa llamado `biplotbootGUI`, implementado en el lenguaje R para su aplicación.
3. Así mismo, se han desarrollado versiones inferenciales basadas en los métodos Bootstrap tanto para el análisis Canónico de Correspondencias simétrico y no simétrico como para los Biplot Múltiples.
4. Para estas versiones también se han implementado softwares específicos que permiten su puesta en práctica. Estos programas se denominan `cncaGUI` y `multibiplotGUI`.
5. Siguiendo las ideas de Vichi y Saporta, 2009, se ha desarrollado un nuevo método con su algoritmo iterativo denominado Clustering Biplot cuyo objetivo es la búsqueda de la mejor clasificación de los individuos mediante el algoritmo k-means en un espacio de dimensión reducida en el que se representen los individuos y las variables mediante el HJ Biplot.
6. Así mismo, se ha presentado un nuevo método, versión del HJ Biplot, que pretende la búsqueda de ejes factoriales disjuntos, de tal forma que cada

variable sólo contribuye a la conformación de un eje. Esta versión facilita la interpretación de los factores resultantes y se ha denominado Disjoint Biplot.

7. Como combinación de ambas técnicas, se propone el Clustering Disjoint Biplot. Este método tiene el doble objetivo de buscar la mejor clasificación de los individuos mediante el algoritmo de k-means y generar ejes factoriales disjuntos que permitan una interpretación más fácil a la hora de representar los individuos y los grupos generados en el mismo espacio de representación mediante el HJ Biplot.
8. Al igual que desarrollaron Macedo y Freitas, 2015 para el CDPCA, se ha implementado una interfaz gráfica de usuario llamada CDBiplot para facilitar la ejecución de estos algoritmos.

# ARTÍCULOS PUBLICADOS



## A Methodology for Biplots Based on Bootstrapping with R

Una metodología para biplots basada en bootstrapping con R

ANA B. NIETO<sup>1,a</sup>, M. PURIFICACIÓN GALINDO<sup>1,b</sup>, VÍCTOR LEIVA<sup>2,3,c</sup>,  
PURIFICACIÓN VICENTE-GALINDO<sup>1,d</sup>

<sup>1</sup>DEPARTAMENTO DE ESTADÍSTICA, UNIVERSIDAD DE SALAMANCA, ESPAÑA

<sup>2</sup>FACULTAD DE INGENIERÍA Y CIENCIAS, UNIVERSIDAD ADOLFO IBÁÑEZ, CHILE

<sup>3</sup>INSTITUTO DE ESTADÍSTICA, UNIVERSIDAD DE VALPARAÍSO, CHILE

---

### Abstract

A biplot is a graphical representation of two-mode multivariate data based on markers for rows and columns often provided in a two-dimensional space. These markers define parameters that help to interpret goodness of fit, quality of the representation and variability and relationships between variables. However, such parameters are estimated as point values by the biplot, thus no information on the accuracy of the corresponding estimators is obtained. We propose a graphical methodology, that may be considered as an inferential version of a biplot, based on bootstrap confidence intervals for the mentioned parameters. We implement our methodology in an R package and validate it with simulated and real-world data.

**Key words:** Bootstrap Confidence Interval, Graphical Methods, Multivariate Data, Quantiles, Software.

### Resumen

Un biplot es una representación gráfica de datos multivariantes de dos vías basada en marcadores para filas y columnas proporcionada usualmente en un espacio bidimensional. Estos marcadores definen parámetros que ayudan a interpretar bondad de ajuste, calidad de representación y variabilidad y relaciones entre variables. Sin embargo, tales parámetros son estimados puntualmente en el biplot, sin proporcionar información acerca de la precisión de los estimadores. Se propone una metodología gráfica, que puede

---

<sup>a</sup>Associate Professor. E-mail: [ananiето@usal.es](mailto:ananiето@usal.es)

<sup>b</sup>Professor. E-mail: [pgalindo@usal.es](mailto:pgalindo@usal.es)

<sup>c</sup>Professor. E-mail: [victor.leiva@yahoo.com](mailto:victor.leiva@yahoo.com)

<sup>d</sup>Professor. E-mail: [purivg@usal.es](mailto:purivg@usal.es)

ser considerada como una versión inferencial de un biplot, basada en intervalos de confianza bootstrap para los parámetros mencionados. La metodología es implementada en un paquete R y validada con datos simulados y reales.

**Palabras clave:** cuantiles, datos multivariantes, intervalos de confianza bootstrap, métodos gráficos, software.

## 1. Introduction and Bibliographical Review

Analyses of high dimension data matrices for individuals and variables can be performed using multivariate techniques, which reduce this dimensionality projecting the data onto an optimal subspace, conserving the patterns of similarity between individuals and of covariation between variables. Differences among these techniques depend on the type of variables and metrics used into the respective subspaces. The biplot methods (biplots in short) proposed by Gabriel (1971) are part of such techniques, but biplots did not diffuse at the same speed as other multivariate techniques, due to the absence of software. Biplots are a graphical display in the context of principal component analysis (PCA in short, and PC for principal components), jointly representing a multivariate data matrix by markers for its rows (individuals) and columns (variables), permitting interrelations between them to be captured visually in a low-dimensional space. Biplots allow us to make description and also modeling and diagnostics (Bradu & Gabriel 1978) and are a powerful data visualization tool that can be considered as a multivariate version of the scatterplot, because biplots are usually performed in the two-dimensional space. This is the classical biplot of Gabriel, which has two parts. First, it approximates the data matrix by a singular value decomposition (SVD). Then, this matrix is factorized to obtain a low dimension Euclidean map through row and column markers represented by points/vectors, similarly to the case of the factorial correspondence analysis (CA). However, in biplots, the interpretation is based on the geometric properties of the scalar product between the rows and columns, allowing an approximation of the data matrix elements to be obtained.

Gower & Harding (1988), Gower (1992) and Gower & Hand (1996) provided a different focus of the classical biplot, ordering the individuals by scaling and then superimposing the variables so that a joint graphical interpretation is possible, as usual in biplots. The two most used biplots are known as GH and JK. Galindo (1986) proved that, with a suitable choice of the markers, it is possible to represent the rows and columns simultaneously on the same Euclidean space with an optimal quality, which is called the HJ biplot. Its coordinates for columns are the column markers in the GH biplot and the coordinates for rows are the row markers in the JK biplot. HJ biplot of Galindo has been applied in several fields. Orfao, González, San-Miguel, Ríos, Caballero, Sanz, Calmuntia, Galindo & López-Borrasca (1988) applied this biplot to histopathology; Rivas-Gonzalo, Gutiérrez, Polanco, Hebrero, Vicente-Villardón, Galindo & Santos-Buelga (1993) to enology; Mendes, Fernández-Gómez, Galindo, Morgado, Maranhão, Azeiteiro & Bacelar-Nicolau (2009) to limnology; Viloria, Gil, Durango & García (2012) to biotechnology; Díaz-Faes, González-Albo, Galindo & Bordons (2013) to bibliome-

try; García-Sánchez, Frías-Aceituno & Rodríguez-Domínguez (2013) to sociology; and Gallego-Álvarez, Galindo & Rodríguez-Rosa (2014) to sustainability. Another biplot is known as GGE which displays the genotype main effect (G) and the genotype by environment interaction (GE) in two-way (two-mode) data. The GGE biplot originates from data graphical analysis of multi-environment variety trials. Technical details of the GH, HJ and JK biplots are provided in Section 2 and of the GGE biplot in Frutos, Galindo & Leiva (2014).

Ter-Braak (1986) used biplots fitted with linear models in the context of direct gradient analysis, which allows a set of species to be ordered according to its relationship with a set of environmental variables. Gauch (1988) employed the biplots for validating and selecting models when the interaction between genotype and environment is studied. Ter-Braak (1990) and Ter-Braak & Looman (1994) took advantage of the relationship between biplot and regression methods to introduce the biplot of the regression coefficient matrix and to propose a biplot based on reduced rank regression. Cárdenas & Galindo (2003) investigated the inferential aspects of biplots using the generalized bilinear models, extending their fitting with external information for variables related to the exponential family.

Vairinhos (2003) showed that the biplots are an ideal basis for the development of a data mining system, because most of the data analysis techniques can be expressed as particular cases of biplots. Amaro, Vicente-Villardón & Galindo (2004) studied the properties of MANOVA biplots within the context of the multivariate general linear model, developing methods for their interpretation. Hernández (2005) studied the performance of biplots in the presence of outliers and Ramírez, Vásquez, Camardiel, Pérez & Galindo (2005) proposed biplots to detect multicollinearity. As an alternative to the multiple CA for the case of presence/absence variables associated with the binomial distribution, Vicente-Villardón, Galindo & Blázquez (2006) considered the prediction biplots and applied it to biplots fitted by generalized linear regression, proposing logistic biplots, later extended by Demey, Vicente-Villardón, Galindo & Zambrano (2008).

Bradú & Gabriel (1974) and Bradú & Gabriel (1978) studied the fitting of bilinear models in two-way tables, analyzing collinearity between rows and columns on the biplots. Gabriel & Zamir (1979) also worked on the fitting of these bilinear models, but they proposed iterative techniques to obtain approximations at low rank using weighted least squares. Denis (1991), Falguerolles (1995), Choulakian (1996) and Gabriel (1998) used biplots to study interactions in two-and-three-way tables. Gabriel (1998) developed diagnostics in models based on contingency tables. Sepúlveda, Vicente-Villardón & Galindo (2008) used biplots as a diagnostic tool of local dependence in latent class models.

Methods for three-way data analysis have shown to be variants of the PCA of the two-way supermatrix, being the two most common ones: (i) TUCKER3 (Tucker 1966) and (ii) STATIS (L'Hermier des Plantes 1976). In (i), the data are summarized by three-mode components, and for their entities (individuals, sampling sites, etc.), component loadings are yielded. In (ii), data are compared on several occasions (time instants) by a PCA linked into column vectors (variables), belonging to different occasions. Based on TUCKER3, Carlier & Kroonenberg

(1996) generalized the SVD to a three-mode table proposing interactive and joint biplots to capture the information from the data. The difference between these two biplots is how the initial data matrix is treated, because in the interactive biplot two modes are combined, whereas the joint biplot is conditioned to one of the modes. Martín-Rodríguez, Galindo & Vicente-Villardón (2002) proposed meta-biplots following the meta-PC and procrustes methods, allowing biplots to be compared for studying individuals with variables, alternatively to the interactive and joint biplots.

Vallejo-Arboleda, Vicente-Villardón & Galindo (2006) and Vallejo-Arboleda, Vicente-Villardón, Galindo, Fernández, Fernández & Bécáres (2008) proposed the canonical STATIS, a biplot for multi-table data. Frequently, multivariate data taken over multiple occasions are found to produce a multi-table experiment. Neither the separate analysis of each occasion, using MANOVA or canonical analysis, nor the joint analysis using STATIS for multiple tables, are adequate to capture the real structure of the data matrices. This is because the former one accounts for group structure, but for not time evolution, whereas the last one confuses between and within group variabilities. Canonical STATIS permits a data group structure to be accounted, as well as time evolution on various occasions, by obtaining common or stable canonical variables across multiple occasions or data sets.

We focus on the classical biplot of Gabriel (1971); see Cárdenas, Galindo & Vicente-Villardón (2007) for a review and the books by Gower & Hand (1996), Greenacre (2010) and Gower, Gardner-Lubbe & Le-Roux (2011) for more details.

The bootstrap method was proposed by Efron (1979, 1987, 1993) and is used for facilitating calculations from statistical inference, which need the modern computer power since they are intensive. Bootstrapping corresponds to a resampling method useful for estimating the standard error (SE) of an estimator and then bootstrap confidence intervals (CIs) can be constructed. Because it is difficult to obtain closed expressions for sampling distributions of statistics associated with biplots (or with the SVD components), bootstrapping seems to be sound and adequate for approximating these distributions. Marcenko & Pastur (1967), Wachter (1978), Stewart (1980), McKay (1981), Edelman (1988), Lambert, Wildt & Durand (1990), Milan & Whittaker (1995), Díaz-García, Leiva & Galea (2002), Díaz-García, Galea & Leiva (2003), Díaz-García & Leiva (2003), Caro-Lopera, Leiva & Balakrishnan (2012) and Sánchez, Leiva, Caro-Lopera & Cysneiros (2015) discussed sampling distributions of SVDs and other decompositions and random matrices. Chatterjee (1984), Daudin, Duby & Trécourt (1988), Holmes (1989) and Linting, Meulman, Groenen & Van der Kooij (2007) combined bootstrapping with several multivariate techniques to provide more accurate results. Meulman (1982), Greenacre (1984), Gifi (1990), Timmerman, Kiers, Smilde & Stouten (2009), Kiers (2004) and Van Ginkel (2011) used bootstrapping in the context of multi-mode data.

The main objective of our work is to introduce a new methodology for biplots based on bootstrapping. We implement it in a graphical user interface (GUI) package developed in the statistical software R ([www.r-project.org](http://www.r-project.org)), named



`biplotbootGUI`. R is an integrated suite of software facilities for data manipulation, calculation and graphical display; see R-Team (2013).

The paper is organized as follows. In Section 2, we provide the technical background of this work. In Section 3, we propose a biplot methodology with bootstrapping and the state-of-art of the software developed for biplots. In addition, in this section, we detail the features of the `biplotbootGUI` package. In Section 4, we perform the numerical application of the proposed computational implementation by using simulated and real-world data sets. Finally, in Section 5, we sketch some discussions, conclusions and future works.

## 2. Background and Technical Preliminaries

In this section, we provide some technical preliminaries useful for facilitating the understanding of the results proposed in this paper.

### 2.1. Biplot Representations

Any  $I \times J$  two-way data matrix  $\mathbf{X}$  can be expressed as the product of two matrices:  $\mathbf{A}$  with  $I$  rows and  $S$  columns and  $\mathbf{B}$  with  $S$  rows and  $J$  columns. If  $S$  is equal to two, then each row in  $\mathbf{A}$  and each column in  $\mathbf{B}$  have two values defining a point in a two-dimensional plot. When both of  $I$  rows of  $\mathbf{A}$  and  $J$  columns of  $\mathbf{B}$  are displayed in a single graphical representation, this is called a biplot. Thus, a biplot is a graph of a matrix  $\mathbf{X}_{I \times J}$  with row and column markers  $\mathbf{a}_1, \dots, \mathbf{a}_I$  and  $\mathbf{b}_1, \dots, \mathbf{b}_J$ , respectively, chosen in such a way that the inner product  $\mathbf{a}_i^\top \mathbf{b}_j$  is the element  $x_{ij}$  of  $\mathbf{X}$ . The rows and columns of this marker matrix are the coordinate points in an Euclidean space related to the same orthogonal axes. A property of a biplot is that each of the  $I \times J$  values can be recovered by viewing its  $I + J$  points, which is a display of a matrix of rank equal to two (rank-two). Decomposition of a matrix  $\mathbf{X}$  into its component  $\mathbf{A}$  and  $\mathbf{B}$  is called a SVD, obtaining as result  $S$  PCs. A two-way data matrix rarely has rank-two, so that approximating  $\mathbf{X}$  by a rank-two matrix means that only the first two PCs are used for representing it. If these explain an important proportion of the total variability of  $\mathbf{X}$ , then it is sufficiently approximated by a rank-two matrix and can be displayed in a biplot.

Let  $\mathbf{X}$  be a data matrix composed by  $I$  individuals measured on  $J$  variables. The SVD of  $\mathbf{X}$  is defined as  $\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top$ , where  $\mathbf{U}$  is a matrix whose column vectors are orthonormal and correspond to the eigenvectors of  $\mathbf{X}\mathbf{X}^\top$ ,  $\mathbf{V}$  is a matrix whose column vectors are also orthonormal and correspond to the eigenvectors of  $\mathbf{X}^\top\mathbf{X}$ , and  $\mathbf{\Lambda}$  is a diagonal matrix containing the singular values arranged in decreasing order. An element of  $\mathbf{X}$  may be written generically as  $x_{ij} = \sum_{s=1}^{\min(I,J)} \lambda_s u_{is} v_{js}$ . The first  $S$  elements of  $\mathbf{u}_s$  and of  $\mathbf{v}_s$  combined with the singular values  $\lambda_s$  in different ways are used as the coordinates for a graphical display of the data. The most common types of biplots are shown in Figure 1.

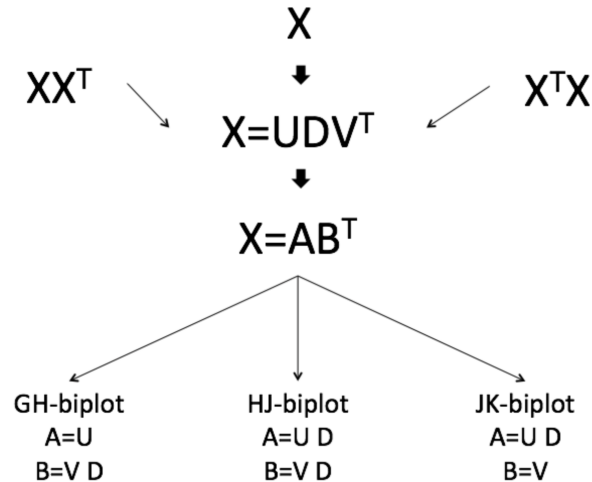


FIGURE 1: Types of biplots.

In a biplot, the column markers  $\mathbf{b}_j$  are shown as arrows and the row markers  $\mathbf{a}_i^\top$  as points; see Figure 1. The biplot representation makes the projection of the row markers onto the column markers easier. The relationships between individuals and variables are studied through the projection of the points representing individuals onto the vectors representing variables, that is,  $x_{ij} \approx \mathbf{a}_i^\top \mathbf{b}_j$  implies  $x_{ij} \approx \|\text{proj } \mathbf{a}_i / \mathbf{b}_j\| \text{sign } \mathbf{b}_j \|\mathbf{b}_j\|$ , where  $\|\text{proj } \mathbf{a}_i / \mathbf{b}_j\|$  is the length of the segment from the origin to the point  $\mathbf{a}_i$  (length of the projection from  $\mathbf{a}_i$  to  $\mathbf{b}_j$ ),  $\text{sign } \mathbf{b}_j$  is the sign of  $\mathbf{b}_j$  and  $\|\mathbf{b}_j\|$  is the module of  $\mathbf{b}_j$  (length of the segment from the origin to  $\mathbf{b}_j$ ). This means that  $x_{ij}$  is approximately the module of the projection of  $\mathbf{a}_i$  onto  $\mathbf{b}_j$  multiplied by the length of  $\mathbf{b}_j$ , with its corresponding sign. The direction of the vector  $\mathbf{b}_j$  shows the increasing direction of the corresponding variable values. The projections of the points  $\mathbf{a}_i$  onto a column vector approximate the  $j$ th column of  $\mathbf{X}$  and provide an ordination of the individuals respect to the corresponding variable. Once a way of representation is defined, it can be interpreted. Thus:

- The distance between points are dissimilarities between the corresponding individuals, specially if they are well represented. Individuals that are far away from each other have a larger Euclidean distance (ED) and vice versa. In Figure 2, the largest ED is observed between individuals  $\mathbf{a}_1$  and  $\mathbf{a}_8$  and the smallest ED is obtained between  $\mathbf{a}_5$  and  $\mathbf{a}_6$ .
- In the JK biplot, the line length approximates the variance of the variable. Hence, the longest line is the largest variance. From Figure 2, the variable  $\mathbf{b}_3$  has the largest variance among the variables, while the variable  $\mathbf{b}_2$  has the smallest variance. The cosine of the angle between the vectors approximates the correlation between the variables they represent. Thus, as this angle goes to 90 (or 270) degrees, the corresponding correlation decreases. An angle of 0 or 180 degrees reflects a correlation of 1 or  $-1$ , respectively. The

biplot in Figure 2 shows a strong relationship between the variables  $b_4$  and  $b_5$ , and a weak relationship between the variables  $b_2$  and  $b_3$ , and between  $b_1$  and  $b_3$ . The correlation between the variables  $b_3$  and  $b_6$  is negative. The variables with the same direction involve multicollinearity, such as observed in Figure 2 for variables  $b_1$  and  $b_2$ . Also, biplots show multivariate outliers that can be used to detect clusters, such as the group formed by individuals  $a_1$ ,  $a_2$ ,  $a_7$  and  $a_9$ .

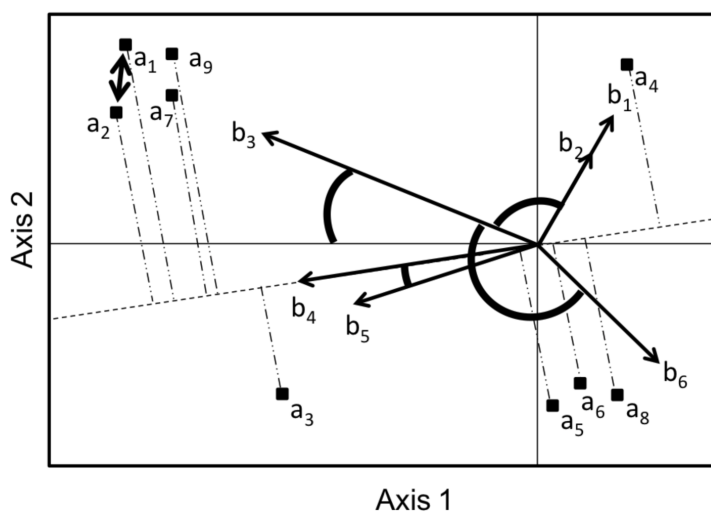


FIGURE 2: Biplot representation of a matrix with 6 variables and 9 individuals.

- The relationships between individuals and variables can be interpreted in terms of scalar products, that is, through the projections of the points onto the arrows. It permits us to know what variables differentiate among groups of individuals. If the projection falls on the origin, the value of the observation is approximately the average of the respective variable. Thus, as the projection of an individual goes increasing onto the direction of a vector representing a variable, this individual goes moving away from the average of that variable, whereas the opposite occurs when the projection goes increasing onto reverse direction. Therefore, in Figure 2, individual  $a_2$  stands out with the largest value of the variable  $b_4$ , followed by  $a_1$ ,  $a_7$  and  $a_9$ .
- In the HJ biplot, the search for the variables that differentiate individuals is made by the interpretation of the factorial axes, that is, the new variables that are a linear combination of the original variables and the relationships of new variables with the observed variables.
- The measure of the relationship between the axes of biplots and each of the observed variables is called relative contribution (RC) of the factor to the element, which represents the variability proportion of each variable explained by the factor. This measure is interpreted such as the coefficient of determination in regression. In fact, if the data are centered, this is

the coefficient of determination in the regression of each variable on the corresponding axis. The RC permits us to know what variables are more related to each axis (Axis 1 and Axis 2) and, therefore, allow us to know the variables involved in the order of the individuals on the projections in each axis. Because the axes are built to be independent, the RCs of each axis to each of the variables are independent and then it is possible to calculate the RC of a plane adding the RCs of the axes that form the plane.

**Properties of the markers in the JK biplot.** In this biplot, we use the metric  $B^T B = I$ , such that:

- The scalar products of the individuals of  $\mathbf{X}$  with the identity metric are the scalar products of row markers included in  $\mathbf{A}$  for the full space  $\mathbf{X}\mathbf{X}^T = \mathbf{A}\mathbf{A}^T$ .
- The ED between two individuals of  $\mathbf{X}$  and the ED between row markers in the full space are the same, that is,  $(\mathbf{x}_i - \mathbf{x}_j)^T (\mathbf{x}_i - \mathbf{x}_j) = (\mathbf{a}_i - \mathbf{a}_j)^T (\mathbf{a}_i - \mathbf{a}_j)$ .
- The row markers and the individual coordinates are equal in the PC space, that is, if  $\Psi$  is a matrix containing the individual coordinates in the PC space, then  $\Psi = (UDV^T)V = UD = \mathbf{A}$ .
- The column coordinates of  $\mathbf{X}$  are the projection of the original axes onto the PC space, that is, the projection of each row marker onto column markers is an approximation of individual values on corresponding variables.
- The quality of representation of the rows is better than the columns.

**Properties of the markers in the GH biplot.** In this biplot, we use the metric  $\mathbf{A}^T \mathbf{A} = \mathbf{I}$ , such that:

- The scalar products of the columns of  $\mathbf{X}$  are the scalar products of the column markers  $\mathbf{X}^T \mathbf{X} = \mathbf{B}\mathbf{B}^T$ .
- If  $\mathbf{X}$  has been centered by columns, the squared length of the vectors representing column markers approximate the covariance of the corresponding variables and as consequence the three following properties arise:
  - The squared length of the column vector approximates the variance of the corresponding variable, whereas the length of the vector approximates the standard deviation (SD) of these variables, that is,  $\|\mathbf{b}_j\| = \|\mathbf{x}_j\| = \sqrt{\text{Var}(\mathbf{x}_j)}$ .
  - The cosine of the angle formed by two column markers approximates the correlation between the corresponding variables, that is,  $\cos(\mathbf{b}_i, \mathbf{b}_j) = \text{Cor}(\mathbf{x}_i, \mathbf{x}_j)$ .
  - The ED between two variables is the ED between the corresponding column markers, that is,  $d^2(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^T (\mathbf{x}_i - \mathbf{x}_j) = \|\mathbf{x}_i\|^2 + \|\mathbf{x}_j\|^2 - 2(\mathbf{x}_i^T \mathbf{x}_j) = \|\mathbf{b}_i\|^2 + \|\mathbf{b}_j\|^2 - 2(\mathbf{b}_i^T \mathbf{b}_j) = d^2(\mathbf{b}_i, \mathbf{b}_j)$ .

- The coordinates in  $\mathbf{B}$  are the importance of the variables on the principal axes.
- The Mahalanobis distance between two individuals can be approximated by the ED between row markers, that is, by  $(\mathbf{x}_i - \mathbf{x}_j)^\top \widehat{\boldsymbol{\Sigma}}^{-1} (\mathbf{x}_i - \mathbf{x}_j) = (\mathbf{a}_i - \mathbf{a}_j)^\top (\mathbf{a}_i - \mathbf{a}_j)$ , where  $\widehat{\boldsymbol{\Sigma}}$  is an estimate of the corresponding variance-covariance matrix.
- If  $\mathbf{X}$  is centered by columns, the row marker coordinates are the individual coordinates in the PC space and then  $\mathbf{A}$  contains the scores on the standardized PCs.
- The scalar products of the row markers are the scalar products of the rows of  $\mathbf{X}$  with the metric  $(\mathbf{X}^\top \mathbf{X})^{-1}$  in the column space, that is,  $\mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top = \mathbf{A}\mathbf{A}^\top$ .
- The quality of representation of columns is better than that for the rows.

**Properties of the markers in the HJ biplot.** In this biplot, the properties of row markers are the same as in the JK biplot, whereas the column markers are the same as in the GH biplot. The rules for interpreting the HJ biplot are a combination of the rules used in classical biplots, CA, factor analysis (FA) and multidimensional scaling. Specifically, we have that:

- The distances between row markers are interpreted as inverse similarities, in such a way that closer individuals are more similar, which allows the clusters of individuals with similar profiles to be identified.
- The lengths of the column vectors approximate the SD of the variables.
- The cosines of the angles between the column vectors approximate the correlations between variables. Hence, small acute angles are associated with highly positive correlated variables; obtuse angles near to the straight angle are associated with highly negative correlated variables; and right angles are associated with non-correlated variables; analogously the cosines of the angles between the variable markers and the PCs approximate the correlations between them, whereas for standardized data they approximate the factor loadings in FA.
- The location in the plot of the orthogonal projections of the row markers onto a column marker allows us to approximate the ranking of the row elements in that column. Thus, as the projection of a point (individual) away from the center of gravity (average coordinate point), the value that this individual takes on the variable is farther from its mean.
- Row and column markers can be shown in the same Cartesian system with optimal quality of representation. In the CA context, Greenacre (1984) and Lebart, Morineau & Piron (1995) proved that the clouds of row and column points have the same eigenvalues and barycentric relationships between them

exist. The relationships proposed by Galindo (1986) are similar, that is, the relations between the eigenvectors  $\mathbf{U}$  and  $\mathbf{V}$  are  $\mathbf{U} = \mathbf{XVD}^{-1}$  and  $\mathbf{V} = \mathbf{X}^{\top}\mathbf{UD}^{-1}$ . Hence, the markers can be written as

$$\mathbf{A} = \mathbf{VD} = \mathbf{X}^{\top}\mathbf{UD}^{-1}\mathbf{D} = \mathbf{X}^{\top}\mathbf{U} = \mathbf{X}^{\top}\mathbf{XVD}^{-1} = \mathbf{X}^{\top}\mathbf{BD}^{-1} \quad \text{and}$$

$$\mathbf{B} = \mathbf{UD} = \mathbf{XVD}^{-1}\mathbf{D} = \mathbf{XV} = \mathbf{XX}^{\top}\mathbf{UD}^{-1} = \mathbf{XAD}^{-1}$$

Therefore, the row coordinates are weighted means of the columns where the weights are the values of  $\mathbf{X}$  and the same applies for columns.

## 2.2. Goodness of Fit

To assess goodness of fit in  $S$  dimensions, we need to know the variability proportion of  $\mathbf{X}$  explained by the approximated matrix  $\tilde{\mathbf{X}}$ , that is, the proportion of total variability  $= \|\mathbf{X}\|^2 = \sum_{i=1}^I \sum_{j=1}^J x_{ij}^2$ . Because of the least-square properties of the SVD and the orthogonormality of  $\mathbf{U}$  and  $\mathbf{V}$ , this total variability can be split into an explained variability and a residual variability expressed in terms of the squared singular values as  $\sum_{s=1}^{\tilde{S}} \lambda_s^2 = \sum_{s=1}^S \lambda_s^2 + \sum_{s=S+1}^{\tilde{S}} \lambda_s^2$ , where  $\tilde{S}$  is the rank of  $\mathbf{X}$ . This expression shows that the sum of the first  $S$  squared singular values divided by the total sum of squared singular values is a way to assess the amount of total variability explained by the first  $S$  vectors. If the explained total variability is large, it means that the graph represented by the first  $S$  singular vectors has a good representation of the initial matrix. If only a small proportion of such a variability is explained by the first singular vectors, the rest of variability can be explained by vectors of higher dimensions. If the data are centered by columns, individuals located near the origin may have measures close to the variable means, or their variability is explained by higher dimensions. In the same way, variables located near the origin may have small variability or may be not well represented in these dimensions. The estimates of row and column markers for each biplot and their quality of representation are shown in Table 1.

TABLE 1: Markers and their quality of representation.

	Rows		Columns	
	Coordinate	Quality	Coordinate	Quality
GH biplot	$\mathbf{U}$	$S/\tilde{S}$	$\mathbf{VD}$	$\sum_{s=1}^S \lambda_s^4 / \sum_{s=1}^{\tilde{S}} \lambda_s^4$
JK biplot	$\mathbf{UD}$	$\sum_{s=1}^S \lambda_s^4 / \sum_{s=1}^{\tilde{S}} \lambda_s^4$	$\mathbf{V}$	$S/\tilde{S}$
HJ biplot	$\mathbf{UD}$	$\sum_{s=1}^S \lambda_s^4 / \sum_{s=1}^{\tilde{S}} \lambda_s^4$	$\mathbf{VD}$	$\sum_{s=1}^S \lambda_s^4 / \sum_{s=1}^{\tilde{S}} \lambda_s^4$

## 2.3. Contributions

The quality of representation detailed in Subsection 2.2 is a way to globally measure the fit of an approximation. However, it is also possible to individually measure its fit related to units and variables, which is important to interpret the results from the biplot. These measures are based on the concepts of RC or

absolute contribution (AC) proposed in Galindo (1986) and Jambu (1991). The total inertia is the sum of the eigenvalues of a matrix, that is, the trace of the matrix, used as a measure of the total variability in a data matrix. It is directly related to the physical concept of inertia, which is the tendency of an object in motion to stay in motion, and the tendency of an object at rest to stay at rest. Note that the total variability of the individual cloud is equal to the total variability of the variable cloud, given by  $\text{trace}(\mathbf{X}\mathbf{X}^\top) = \text{trace}(\mathbf{X}^\top\mathbf{X}) = \sum_{s=1}^S \lambda_s^2$ , where  $\sum_{s=1}^S \lambda_s^2 = \sum_{j=1}^J d^2(\mathbf{b}_j, \mathbf{0}) = \sum_{s=1}^S \sum_{j=1}^J b_{js}^2 = \sum_{i=1}^I d^2(\mathbf{a}_i, \mathbf{0}) = \sum_{s=1}^S \sum_{i=1}^I a_{is}^2$ . The ACs of the individual  $i$  and of the variable  $j$  to the variability of the axis  $s$  are  $\text{AC}_{is} = a_{is}^2$  and  $\text{AC}_{js} = b_{js}^2$ , respectively. The total inertia of the factor  $s$  taking into account the ACs of the individual  $i$  and of the variable  $j$  are  $\sum_{i=1}^I a_{is}^2 = \lambda_s^2$  and  $\sum_{j=1}^J b_{js}^2 = \lambda_s^2$ , respectively. The RCs of the elements  $i$  and  $j$  to the factor  $s$  are  $\text{RC}_{is} = \text{AC}_{is}/\lambda_s$  and  $\text{RC}_{js} = \text{AC}_{js}/\lambda_s$ , respectively, whereas the RCs of the factor  $s$  to the elements  $i$  and  $j$  are  $\text{RC}_{si} = a_{is}^2/d^2(\mathbf{a}_i, \mathbf{0}) = \cos^2(\mathbf{a}_i)$  and  $\text{RC}_{sj} = a_{js}^2/d^2(\mathbf{b}_j, \mathbf{0}) = \cos^2(\mathbf{b}_j)$ , respectively. The RC of the element to the factor measures how this factor can be explained by such an individual or variable.

### 3. A Biplot Methodology with Bootstrapping

In this section, we provide some aspects related to bootstrapping, propose a biplot methodology based on bootstrapping, discuss the state-of-art of the software developed for biplots and detail the features of the `biplotbootGUI` package.

#### 3.1. Bootstrapping

Statistical theory attempts to answer three basic questions. (i) How should the data be collected? (ii) How should the collected data be analyzed and summarized? (iii) How accurate is this data summary? The third question constitutes part of the process known as statistical inference. Bootstrapping can help to answer this question when a sampling distribution is not available. Suppose a random sample  $\mathbf{X} = (X_1, \dots, X_n)^\top$  is obtained from a population with unknown distribution. Let  $\mathbf{x} = (x_1, \dots, x_n)^\top$  be an observation of  $\mathbf{X}$ , from which we can obtain the estimate  $\hat{\theta} = s(\mathbf{x})$  of a parameter of interest  $\theta$ , corresponding to an observed value of the estimator  $\hat{\theta} = s(\mathbf{X})$  for which we want to know its accuracy. A bootstrap sample  $\mathbf{x}^* = (x_1^*, \dots, x_n^*)^\top$  is defined to be a sample of size  $n$  with replacement from the observed sample  $\mathbf{x}$ . A bootstrap replication of  $\hat{\theta}$  results from applying the same function  $s(\cdot)$  to  $B$  bootstrap samples. To calculate the accuracy of the estimator  $\hat{\theta}$ , the bootstrap estimate of the corresponding SE,  $\text{SE}[s(\mathbf{X})]$  say, can be used. Its bias can be empirically calculated from  $\text{B}[s(\mathbf{X})] = s(\mathbf{x}^*) - \theta$ . Algorithm 1 summarizes the bootstrap method to calculate the mentioned SE, which is often used for constructing a CI for a parameter.

**Algorithm 1** Bootstrapping

- 
- 1: Select  $B$  bootstrap samples  $\mathbf{x}_1^*, \dots, \mathbf{x}_B^*$  each consisting of  $n$  data drawn with replacement from  $\mathbf{x}$ .
  - 2: Calculate the estimate  $\hat{\theta}_b^* = s_b(\mathbf{x}^*)$  from the  $b$ th sample corresponding to a bootstrap replication of  $\hat{\theta}$  for  $b = 1, \dots, B$ .
  - 3: Estimate the SE of  $\hat{\theta} = s(\mathbf{X})$  with the sample SD of the  $B$  bootstrap replications, that is, by  $\hat{\text{SE}}[s(\mathbf{X})] = ((1/B) \sum_{b=1}^B (s_b(\mathbf{x}^*) - s(\mathbf{x}^*))^2)^{1/2}$ , where  $s(\mathbf{x}^*) = \sum_{b=1}^B s_b(\mathbf{x}^*)/B$ .
- 

**Normal and  $t$  distributions-based CIs.** Assume the estimator  $\hat{\theta}$  is normally distributed (at least approximately) with unknown expectation  $\theta$  and SE known given by  $(\text{Var}[\hat{\theta}])^{1/2} = \text{SE}[\hat{\theta}]$ , that is,  $\hat{\theta} \sim N(\theta, \text{Var}[\hat{\theta}])$ . Then,  $Z = (\hat{\theta} - \theta)/\text{SE}[\hat{\theta}] \sim N(0, 1)$ . Note that  $P(|Z| \leq z_{1-\alpha/2}) = 1 - \alpha$  is equivalent to

$$P(\theta \in [\hat{\theta} - z_{1-\alpha/2}\text{SE}[\hat{\theta}], \hat{\theta} + z_{1-\alpha/2}\text{SE}[\hat{\theta}]]) = 1 - \alpha$$

Denote  $\hat{\theta}_L = \hat{\theta} - z_{1-\alpha/2}\text{SE}[\hat{\theta}]$  and  $\hat{\theta}_U = \hat{\theta} + z_{1-\alpha/2}\text{SE}[\hat{\theta}]$ . Hence, the random interval  $[\hat{\theta}_L, \hat{\theta}_U]$  has probability  $1 - \alpha$  of containing the true value of  $\theta$ . Thus, a  $100 \times (1 - \alpha)\%$  CI for  $\theta$  is  $[\hat{\theta} \pm z_{1-\alpha/2}\hat{\text{SE}}[s(\mathbf{X})]]$ . These results are meaningful for large enough sample sizes, for example, for  $n \geq 25$ . However, if we have small samples ( $n < 25$ ), these results still can be correct (Bickel & Krieger 1989), but inappropriate for  $n \leq 5$  (Chernick 1999). In addition, if  $\text{SE}[\hat{\theta}]$  is unknown, we can estimate it with the expression given in Step 3 of Algorithm 1,  $\hat{\text{SE}}[s(\mathbf{X})]$  say, but in this case  $Z = (\hat{\theta} - \theta)/\hat{\text{SE}}[s(\mathbf{X})]$  still follows, in an approximate way, for large enough sample sizes, a standard normal distribution. Otherwise (smaller size samples), we have  $Z = (\hat{\theta} - \theta)/\hat{\text{SE}}[s(\mathbf{X})] \sim t(n - 1)$ , that is, now  $Z$  is Student- $t$  with  $n - 1$  degrees of freedom distributed, but we need additionally the normality assumption for the population  $X$ . Thus, in this case, an  $100 \times (1 - \alpha)\%$  CI for  $\theta$  with small sample sizes is  $[\hat{\theta} \pm t_{1-\alpha/2}(n - 1)\hat{\text{SE}}[s(\mathbf{X})]]$ , where  $t_{1-\alpha/2}(n - 1)$  denotes the  $(1 - \alpha/2) \times 100$ th quantile of the  $t(n - 1)$  distribution.

**Bootstrap normal and  $t$  distributions-based CIs.** The normal and  $t$  distributions do not adjust the CI for  $\theta$  to account for skewness and/or other aspects that can result when  $\hat{\theta}$  is not the sample mean. The bootstrap normal and  $t$  CIs are procedures that adjust these aspects. Thus, by using the bootstrap method, we can obtain accurate CIs without having to make the normality assumption. This procedure approaches the population distribution directly from the data and builds CIs in the same way that we have explained in the cases of normal and  $t$  distributions. Algorithm 2 summarizes this procedure.

**Bootstrap quantile-based CI.** An alternative way to the bootstrap  $t$  distribution-based method (boot- $t$ ) for constructing bootstrap CIs is the quantile method (boot- $q$ ). The boot- $t$  and boot- $q$  methods are based on a simple structure. However, several data analyses involve more complex structures such as analysis of variance, regression models or time series. Boot- $t$  and boot- $q$  methods used for a more complex parameter than the mean were recently proposed by Leiva,



**Algorithm 2** Bootstrap normal and  $t$  CIs

- 
- 1: Follow Steps 1-3 of Algorithm 1 and obtain  $\hat{SE}[s(\mathbf{X})]$ .
  - 2: Calculate the value  $z_b^* = (\hat{\theta}_b^* - \hat{\theta})/\hat{SE}_b^*$  from the  $b$ th sample corresponding to a bootstrap replication of  $z = (\hat{\theta} - \theta)/\hat{SE}[\hat{\theta}]$ , where  $\hat{\theta}_b^*$  and  $\hat{SE}_b^*$  are the estimates of  $\theta$  and of  $SE[\hat{\theta}]$  for the  $b$ th bootstrap sample,  $\mathbf{x}_b^*$  say, with  $b = 1, \dots, B$ .
  - 3: Determine the  $(1 - \alpha/2) \times 100$ th quantile of  $z_b^*$  as follows:
    - 3.1 If  $n \geq 25$ , use the value  $\hat{z}_{1-\alpha/2}$  such that  $\#\{z_b^* \leq \hat{z}_{1-\alpha/2}\}/B = \alpha/2$ ;
    - 3.2 If  $n < 25$ , use  $\hat{t}_{1-\alpha/2}(n-1)$  such that  $\#\{z_b^* \leq \hat{t}_{1-\alpha/2}(n-1)\}/B = \alpha/2$ .
  - 4: Compute the bootstrap CI for  $\theta$  as follows:
    - 4.1 If  $n \geq 25$ ,  $[\hat{\theta} \pm \hat{z}_{1-\alpha/2} \hat{SE}[s(\mathbf{X})]]$ ;
    - 4.2 If  $n < 25$ ,  $[\hat{\theta} \pm \hat{t}_{1-\alpha/2}(n-1) \hat{SE}[s(\mathbf{X})]]$ . If  $B\alpha/2$  is not an integer, assume  $\alpha/2 \leq 0.5$  and compute  $k$  as the largest integer less or equal than  $(B+1)\alpha/2$  and define the  $(1 - \alpha/2) \times 100$ th quantile by the  $(B+1-k)$ th largest value of  $z_b^*$ .
- 

Marchant, Saulo, Aslam & Rojas (2014) and can be adapted to data structures still more complex, as occurs with biplots; see Subsection 3.2. Algorithm 3 summarizes the boot-q method.

**Algorithm 3** Bootstrap quantile CIs

- 
- 1: Follow Steps 1 and 2 of Algorithm 1 obtaining the bootstrap replications  $\hat{\theta}_1^*, \dots, \hat{\theta}_B^*$ .
  - 2: Order  $\hat{\theta}_1^*, \dots, \hat{\theta}_B^*$  obtained in Step 1 of Algorithm 3 as  $\hat{\theta}_{(1)}^* < \dots < \hat{\theta}_{(B)}^*$ .
  - 3: Determine the  $(B\alpha/2) \times 100$ th and  $B(1 - \alpha/2) \times 100$ th quantiles of the distribution of  $\hat{\theta}^*$ , denoted by  $\hat{\theta}_{B\alpha/2}$  and  $\hat{\theta}_{B(1-\alpha/2)}$ , respectively.
  - 4: Construct the boot-q CI as  $[\hat{\theta}_{B\alpha/2}, \hat{\theta}_{B(1-\alpha/2)}]$ .
- 

## 3.2. Biplots Based on Bootstrapping

We adapt Algorithms 2 and 3 to measure the accuracy of the following biplot parameters: (B1) goodness of fit; (B2) quality of the approximation for columns; (B3) eigenvalues; (B4) angles between variables; (B5) angles between variables and axes; (B6) RC to the total variability of the  $j$ th column element; (B7) RC of the column element  $j$  to the  $q$ th factor; and (B8) RC of the  $q$ th factor to the  $j$ th column element. Adaptation of Algorithms 2 and 3 is given in Algorithm 4.

**Algorithm 4** Adaptation of Algorithms 2 and 3

- 
- 1: Follow Steps 1 and 2 of Algorithm 1 obtaining the bootstrap replications  $\hat{\theta}_1^*, \dots, \hat{\theta}_B^*$ .
  - 2: Calculate the empirical mean, SE and bias of the estimator  $\hat{\theta}$  with the bootstrap samples by using the expressions  $\hat{E}[s(\mathbf{X})] = \sum_{b=1}^B s_b(\mathbf{x}_b^*)/B$ ,  $\hat{SE}[s(\mathbf{X})] = ((1/B) \sum_{b=1}^B (s_b(\mathbf{x}_b^*) - \hat{E}[s(\mathbf{X})])^2)^{1/2}$  and  $\hat{B}[s(\mathbf{X})] = s(\mathbf{x}^*) - s(\mathbf{x})$ , respectively.
  - 3: Establish boot-t CIs for the parameters (B1)-(B8) with Step 4 of Algorithm 2.
  - 4: Determine boot-q CIs for the parameters (B1)-(B8) with Step 4 of Algorithm 3.
-

### 3.3. Software for Biplots

Macros for biplots have been implemented in main commercial and non-commercial statistical software packages. Currently, most commercial statistical software packages include a procedure or macro for generating biplots; see details in Frutos et al. (2014). Specifically, the `GGEbiplot` software, dedicated to the GGE biplot ([www.ggebplot.com](http://www.ggebplot.com)), can also generate the classical biplot. The `GGEbiplot` program is a commercial software and is widely used by agronomists, crop scientists and geneticists; see Yan & Kang (2003) and Frutos et al. (2014) and references therein. Vicente-Villardón (2010) implemented in the commercial software `MatLab` ([www.mathworks.com/products/matlab](http://www.mathworks.com/products/matlab)) a program to perform biplots called `multbiplot`. It contains classical, HJ and logistic biplots, among other biplots, as well as simple and multiple CA for contingency tables.

Most of the software available for biplots is developed for specific applications, or as part inside general purpose packages. Consequently, they are not very flexible and produce static pictures that limit the interpretation of their results. Tables 2 and 3 contain the main packages in R, which have implemented biplot decompositions and/or representations. In these tables, the name of the package, the approach on which it is based, that is, Gabriel (1971), Galindo (1986) or Gower (1992), the main references, the creation date and last update of the corresponding package are presented and its main contents and functionalities are discussed.

In Table 4, we provide a review of the R packages mentioning the word “biplot”, although it refers to the joint representation of coordinates calculated with other methods instead of using the biplot decomposition.

### 3.4. The `biplotbootGUI` Package

Because all of the packages (commercial and non-commercial) discussed in Subsection 3.3 are not suitable for constructing bootstrap CIs for biplot parameters (B1) through (B8), we developed a new package in the R language that combines the biplots described by Gabriel (1971) and Galindo (1986) and the bootstrap method to display results of these biplots and their statistical accuracy measures.

As mentioned, a GUI is a type of user interface which allows practitioners to interact with electronic devices such as computers. It is characterized by the use of icons and visual indicators, as opposed to text-based interfaces, typed command labels or text navigation, to fully represent the information and actions available to the user. The actions are often linked through direct manipulation of the graphical elements. Below, we discuss the features of a GUI in R language of the methodology for biplots based on bootstrapping proposed in the article and implemented in the `biplotbootGUI` package.

TABLE 2: Biplots in R

Package Method	References	Content	Date Update
<b>calibrate</b> Gower	(Graffelman 2013)	It draws calibrated scales with tick marks on non-orthogonal variable vectors in biplots.	21-01-06 20-03-12
<b>BiplotGUI</b> Gower	(La Grange, Le-Roux & Gardner-Lubbe 2009, La Grange, Le-Roux, Rousseeuw, Ruts & Tukey 2013)	It provides a GUI to construct and interact with biplots and displays the variables as calibrated axes. Then, it is not possible to interpret the variable lengths. It allows us to change the title, show labels and points or hide them, change the type, color and size of lines and font, the color and orientation of labels and tick marks, draw convex-hulls and alpha bags. It also performs non-linear and MDS biplots and allows us to choose the distance and way to calculate the coordinates. It shows the variable correlations and provides interactive 3D graphs.	13-08-08 19-03-13
<b>bpca</b> Gabriel, Galindo	(Faria & Demetrio 2012)	It shows biplots in 2D-and-3D and provides variable lengths, angles between variables, correlations, coordinates to individuals and variables, eigenvalues, eigenvectors and quality of representation. It displays a graph with the correlations and their approximations and the 3D graph is interactive.	17-08-08 21-02-12
<b>GGEbiplotGUI</b> Gabriel Galindo Yang	(Yan, Hunt, Sheng & Szlavnic 2000, Yan & Kang 2003, Frutos & Galindo 2013)	It is a GUI to construct and interact with GGE biplots. It provides eigenvalues, % of variability explained by each of them, coordinates of individuals and variables, contributions of factors to elements. Also, this GUI allows us to change the background color, genotype labels, environments labels and title, font, graph title, in addition to showing genotypes and environments, as well as to hide title, axes and symbols. Furthermore, with this GUI it is possible to move the labels by the mouse button and change the color and text of labels.	29-08-11 22-06-13
<b>multibiplotGUI</b> Gabriel, Galindo	(Nieto, Baccalá, Vicente-Galindo & Galindo 2012)	It provides a GUI to construct and interact with multibiplots. It allows us to obtain the quality of representation, contributions, goodness of fit, eigenvalues and possibility of selecting the number of axes. It shows 2D-and-3D graphs (2D graph moves or removes labels, changes color, size and symbol of the points and selects the axes shown in the graph; 3D graph rotates and makes zoom).	29-10-12

TABLE 3: (continued) biplots in R.

Package Method	References	Content	Date Update
<code>nominallogisticbiplot</code> Gabriel, Galindo	(Hernández & Vicente-Villardón 2013a)	It produces a matrix analysis of polytomous items using nominal logistic biplots, extending the binary logistic biplot to polytomous nominal data.	17-09-13
<code>biplot{stats}</code> Gabriel	(R-Team 2013)	It is part of the basics of R and produces a biplot from the output of <code>princomp</code> or <code>prcomp</code> .	25-09-13
<code>ordinallogisticbiplot</code> Gabriel, Galindo	(Hernández & Vicente-Villardón 2013b)	It produces a matrix analysis of polytomous items using ordinal logistic biplots, extending the binary logistic biplot to polytomous ordinal data.	30-10-13 26-11-13
<code>dynbiplotGUI</code> Gabriel, Galindo	(Egido 2014)	It is a GUI to solve dynamic, classic and HJ biplots and tries with 2-and-3 way data matrices.	04-11-13 08-01-14

TABLE 4: R packages which mention biplots.

Package	References	Content	Dates
<code>vegan</code>	(Oksanen, Blanchet, Kindt, Legendre, Minchin, O'Hara, Simpson, Solyomos, Stevens & Wagner 2013)	It provides tools for describing community ecology. This package has the basic functions of diversity, community ordination and dissimilarity analyses. In addition, it shows biplots from results of redundancy, canonical correlation and canonical correspondence analyses, which can be used for other types of data as well.	06-09-01 19-03-13
<code>ade4</code>	(Chessel, Dufour & Thioulouse 2004, Dray & Dufour 2007, Dray, Dufour & Chessel 2007, Chessel, Dufour, Dray, Jombart, Lobry, Ollier & Thioulouse 2013)	It is characterized by the implementation of graphical and statistical functions, availability of numerical data and writing of technical and thematic documentation. It includes bibliographic references and has functions to show biplots from results of the implemented analysis.	10-12-02 11-04-13
<code>ade4TkGUI</code>	(Thioulouse & Dray 2007, Thioulouse & Dray 2012)	It is a Tcl/Tk GUI for some basic functions of the <code>ade4</code> package.	29-09-06 13-11-12
<code>ca</code>	(Nenadic & Greenacre 2007, Greenacre & Nenadic 2012)	It computes and visualizes simple, multiple and joint CA and shows biplots from the results of the previous analysis.	28-07-07 12-06-12
<code>caGUI</code>	(Markos 2012)	It is a Tcl/Tk GUI for functions of the <code>ca</code> package	04-10-09 29-10-12
<code>ThreeWay</code>	(Del Ferraro, Kiers & Giordani 2013)	It allows us to do component analysis for 3-way data arrays by means of Candecomp/Parafac, Tucker1, Tucker2 and Tucker3 models and shows joint biplots from Tucker3 models.	29-10-12 11-06-13

After the R software has been downloaded from `cran.r-project.org` and installed, the user must download and install the `biplotbootGUI` package and its dependencies, which are the `rgl`, `tcltk`, `tcltk2`, `tkrplot` and `vegan` packages; see Adler & Murdoch (2012), Grosjean (2012), Tierney (2012) and Oksanen et al. (2013). Then, to load the `biplotbootGUI` package into the R software, the command `library(biplotbootGUI)` must be entered at the R prompt. Once all these instructions have been followed, the data must be loaded. Hence, one starts the GUI by entering the command `biplotboot(data)` in the R console, where data to be analyzed must be in a data frame; see details and examples in Section 4 of applications. Once the GUI has been initialized, a window entitled “bootstrap on classical biplots” emerges; see Figure 3. This window allows us to enter the number of replications and the confidence level to calculate the CIs. Also, it is possible to choose the parameters to be considered by the user.

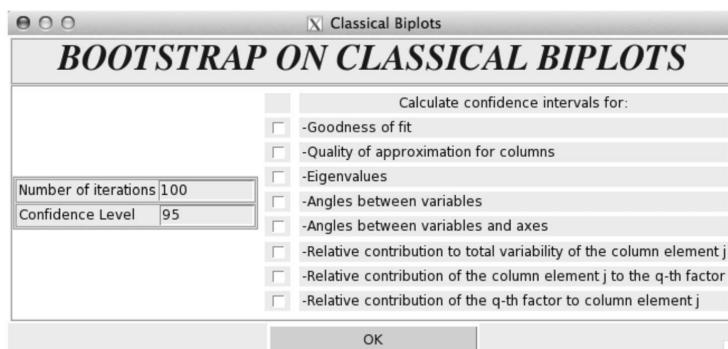


FIGURE 3: Main window.

After entering and selecting the parameters, one must click on the OK button and a window titled “Options” appears; see Figure 4. In this window the following options are available:

- Select the type of biplot to be executed (HJ, GH or JK).
- Select the transformation to be performed on the data considering:
  - Subtract the global mean.
  - Center by columns.
  - Standardize by columns.
  - Center by rows.
  - Standardize by rows.
  - Raw data.
- Change the color, size, label and symbol representing individuals in the graph.
- Change the color, size and label representing variables in the graph.
- Show the axes in the graph.

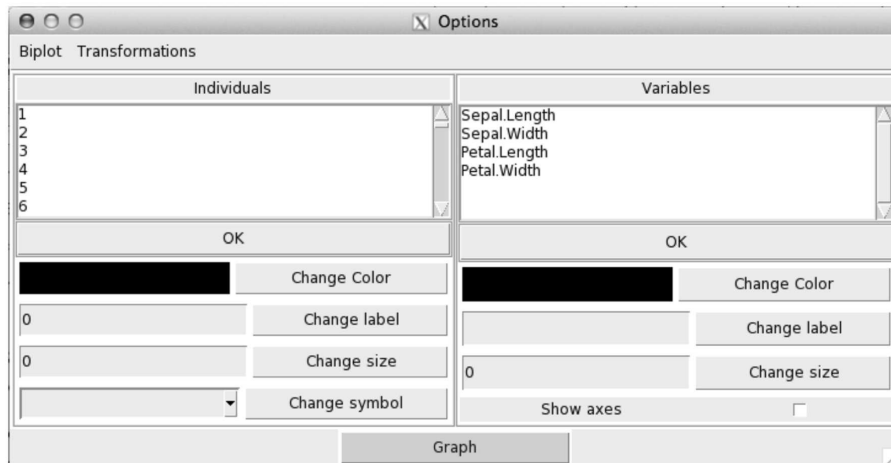


FIGURE 4: Window of options.

Given that not all the data are well represented by the first two axes, a window after clicking the button “graph” emerges with the option to choose the number of axes to be retained, according to the variability explained by each axis. After choosing the number of axes to be retained and clicking the button “choose”, a window showing the resulting graph in 2D appears; see Figure 5. This window displays the labels for the two axes indicating the percentage of variability explained by each of them (72.96 by axis 1 and 22.85% by axis 2). The user can select the axes to be displayed in the graph. At the top of the window, two menus with options to save the graph and show the biplot in 3D are displayed, whereas three text boxes where the user can change the axes displayed in the graph are at the bottom. Also, the user can move or remove the label of a specific element by clicking the left-mouse button and change the graphical displays of such an element by clicking the right-mouse button. This window contains two dropdown menus. In the first one, options to copy, save the graph in different file formats (PDF, postscript, BMP, PNG, JPG/JPEG) or exit are available, whereas the second one provides a 3D-graph made by the `rgl` package; see Adler & Murdoch (2012).

The user can rotate or make zoom in this graph by clicking the left-mouse or right-mouse button. Together with this window, a graph showing the coordinates for variables computed for all of the replications is shown. The GUI provides two text files. In the first one, the parameters of the biplot analysis (see B1-B8) are saved, whereas in the second one, tables with the values for the mean, SE, bias and lower and upper limits of the bootstrap CIs are provided. These two text files are automatically saved together with all the graphs containing the histogram and quantile versus quantile (QQ) plot of the estimates calculated by bootstrapping of the selected parameters in the first window. In the histogram, the solid line represents the estimate of the biplot parameter obtained from bootstrapping, whereas the dashed line is its value obtained from the biplot. In the x-axis of the QQ plot are the theoretical quantiles and in the y-axis the empirical quantiles.

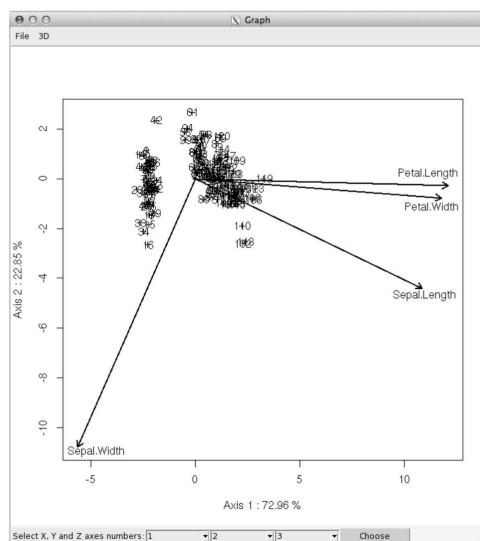


FIGURE 5: Window with a biplot representation in two dimensions.

## 4. Numerical Applications

In this section, we evaluate the performance and potentiality of our methodology by means of the `biplotbootGUI` package using both simulated and real-world data.

### 4.1. Simulated Data

To evaluate the performance of the `biplotbootGUI` package, an HJ biplot with the transformation “centering by columns” has been performed. We simulate data of 100 individuals on 5 variables ( $V_1, \dots, V_5$ ) normally distributed, generated to have correlations  $\text{Cor}(V_1, V_2) = 0.50$ ,  $\text{Cor}(V_2, V_3) = 0.80$  and  $\text{Cor}(V_4, V_5) = 0.90$ . The number of bootstrap replications is 1,000 and the confidence level 95%. The time involved in a bootstrap replication is usually small. For example, the time spent in the calculations of a  $1,000 \times 5$  matrix is less than four minutes for 1,000 replications. First, we explain the main results of the classical biplot. In Table 5, we observe the variability explained by each axis (Axis 1, Axis 2 and Axis 3). Note that the first eigenvalue explains more than 50% and the first three axes explain more than 94.27% of the total variability. Table 6 shows the RCs of the factor to the column elements in the first three axes. Note that all the variables are well represented by the first two axes, except the variable  $V_1$ , which is in the third axis. The biplot representation using the first two axes (Axis 1 and Axis 2) is shown in Figure 7(left). The covariation structure shows a very high correlation between the variables  $V_4$  and  $V_5$ , and  $V_2$  and  $V_3$ , represented by acute angles. Variables  $V_2$  and  $V_3$  have a high correlation with  $V_1$ , however they present no correlation with  $V_4$  and  $V_5$ , since they are almost orthogonal; see Table 7. Second, we explain

the results of applying the bootstrap method. Goodness of fit and eigenvalues are explained next. Figure 8 shows the histogram and QQ plot representing the values of the quality of approximation of 1,000 bootstrap replications.

TABLE 5: Eigenvalues and variability % explained by each of them with simulated data.

No.	Eigenvalue	Variability	Accumulated variability
1	16.06	53.47	53.47
2	12.15	30.59	84.06
3	7.01	10.21	94.27

TABLE 6: RCs of the factors to the column elements for simulated data.

Variable	Axis 1	Axis 2	Axis 3
$V_1$	226.48	16.48	737.49
$V_2$	282.43	118.64	201.18
$V_3$	281.58	89.64	46.09
$V_4$	112.91	421.49	9.12
$V_5$	96.60	353.75	6.12

TABLE 7: Angles between variables for simulated data.

Variable	$V_1$	$V_2$	$V_3$	$V_4$	$V_5$
$V_1$	0.00	14.59	11.58	67.15	66.89
$V_2$	14.59	0.00	3.00	81.73	81.48
$V_3$	11.58	3.00	0.00	78.73	78.47
$V_4$	67.15	81.73	78.73	0.00	0.26
$V_5$	66.89	81.48	78.47	0.26	0.00

We denote by “lower-t” and “upper-t” the lower and upper limits of the CIs based on the boot-t method, respectively, whereas these limits are denoted by “lower-q” and “upper-q” for the boot-q method. Table 11 provides the observed values for the mean, SE, bias and these limits. Notice that the observed value and its approximation are very close. These same results for eigenvalues are provided in Table 8. Figure 6 shows the histogram and QQ plot for the first eigenvalue (a similar behavior is observed for the other four eigenvalues, whose plots are omitted here, but are available under request for interested users). Note that the observed and estimated values practically do not differ and a similar conclusion is reached for the CIs. Each of the five eigenvalues resulting from the SVD of the simulated data shows the values calculated by 1,000 bootstrap replications.

TABLE 8: Results for the eigenvalues with simulated data.

No.	Eigenvalue	Mean	SE	Bias	lower-t	upper-t	lower-q	upper-q
1	16.06	16.09	1.13	0.03	13.85	18.33	13.87	18.24
2	12.15	11.92	0.78	-0.22	10.37	13.47	10.27	13.33
3	7.01	6.9	0.63	-0.12	5.64	8.15	5.74	8.13
4	4.51	4.39	0.31	-0.12	3.78	5.00	3.82	5.05
5	2.69	2.61	0.17	-0.08	2.28	2.94	2.28	2.93



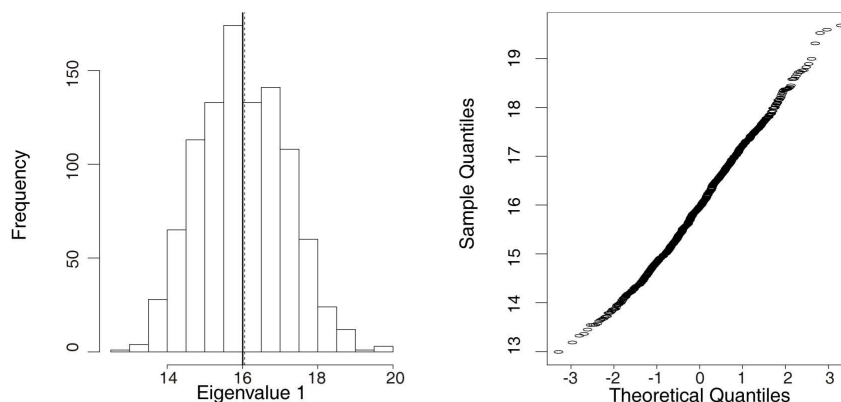


FIGURE 6: Histogram (left) and QQ plot (right) for the first eigenvalue of the simulated data SVD.

## 4.2. Real-World Data

To illustrate the potentiality of the `biplotbootGUI` package, we use real-world data collected by Anderson (1935) and contained in the R software, which can be loaded once the user installs it. The data set corresponds to the measurements in cm of the variables: sepal length ( $Y_1$ ) and width ( $Y_2$ ) and petal length ( $Y_3$ ) and width ( $Y_4$ ), for 50 flowers from each of three species of iris. The species are *iris setosa*, *versicolor* and *virginica*. An HJ biplot with the transformation “standardize by columns” is performed. Once again the number of replications entered is 1,000 and the confidence level 95%. First, we show the main results of the HJ biplot. Table 9 presents the percentage (%) of variability explained by each axis, from where the first eigenvalue explains more than 70% and the first three axes explain almost the 100% of the total variability.

TABLE 9: Eigenvalues and variability % explained by each of them for iris data.

No.	Eigenvalue	Variability	Accumulated variability
1	20.85	72.96	72.96
2	11.67	22.85	95.81
3	4.68	3.67	99.48

Table 10 provides the RCs of the factor to the column elements in the first three axes. Notice that all the variables are well represented by the first axis, except the variable  $Y_4$ , which is well represented by the second axis. The biplot representation using the first two axes is shown in Figure 7(right). The covariation structure shows a very high correlation between  $Y_3$  and  $Y_4$  represented by an acute angle. Both variables have a high correlation with the variable  $Y_1$ . However, there is no relation with  $Y_2$  due to a right angle is obtained. Table 10 also explains the angles between variables in the plane representing the first two axes. Figure 8 shows the histogram and QQ plot representing the values of the quality of approximation of the 1,000 bootstrap replications.

TABLE 10: RCs of the factors to the columns and angles between variables for iris data.

Variable	Axis 1	Axis 2	Axis 3	$Y_1$	$Y_2$	$Y_3$	$Y_4$
$Y_1$	793.52	130.38	76.09	0.00	95.47	20.71	18.27
$Y_2$	211.80	779.43	8.77	95.47	0.00	116.18	113.74
$Y_3$	996.44	0.56	3.00	20.71	116.18	0.00	2.44
$Y_4$	936.50	4.12	59.38	18.27	113.74	2.44	0.00

Table 11 provides the observed values for the mean, SE, bias, lower-t, upper-t, lower-q and upper-q for simulated and real-world (iris) data. Note that there is no difference between the observed value and its approximation, whereas the endpoints of both intervals are similar. Table 12 provides the RCs to the total variability of the variables based on 1,000 bootstrap replications. Note that there are no differences between observed values and their estimates, whereas the width of the CIs is small suggesting a high accuracy of our methodology. Figure 8 shows the histogram and QQ plot for the RCs to total variability of the variable  $Y_1$ . A similar behavior is observed for the other three variables, whose plots are omitted here, but are available under request for interested users.

TABLE 11: Results of the approximation quality for the indicated data set.

Data set	Value	Mean	SE	Bias	lower-t	upper-t	lower-q	upper-q
Simulated	94.27	94.46	0.75	0.19	92.98	95.95	92.9	95.80
Iris	99.48	99.49	0.08	0.01	99.34	99.64	99.33	99.62

TABLE 12: Results of the contributions to the total variability for iris data.

Variable	Value	Mean	SE	Bias	lower-t	upper-t	lower-q	upper-q
$Y_1$	250.95	250.93	0.16	-0.01	250.62	251.24	250.65	251.27
$Y_2$	251.22	251.20	0.17	-0.02	250.86	251.55	250.90	251.57
$Y_3$	247.96	247.99	0.31	0.03	247.39	248.59	247.36	248.53
$Y_4$	249.87	249.87	0.13	0.00	249.63	250.12	249.63	250.10

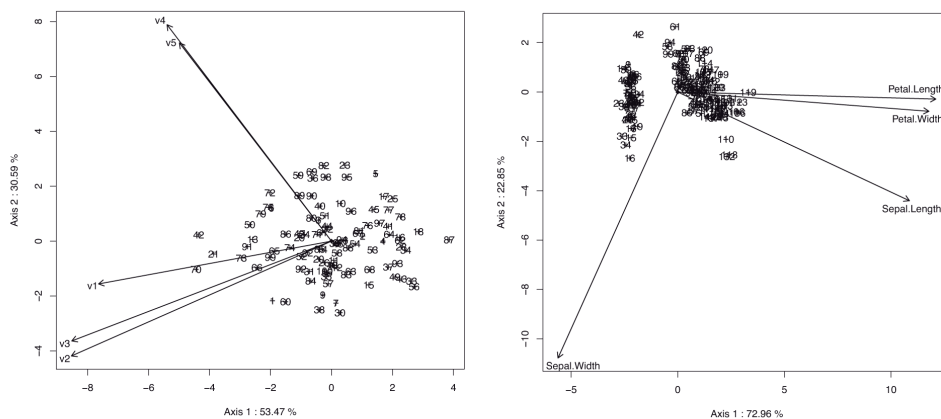


FIGURE 7: Biplots of simulated (left) and iris (right) data sets.

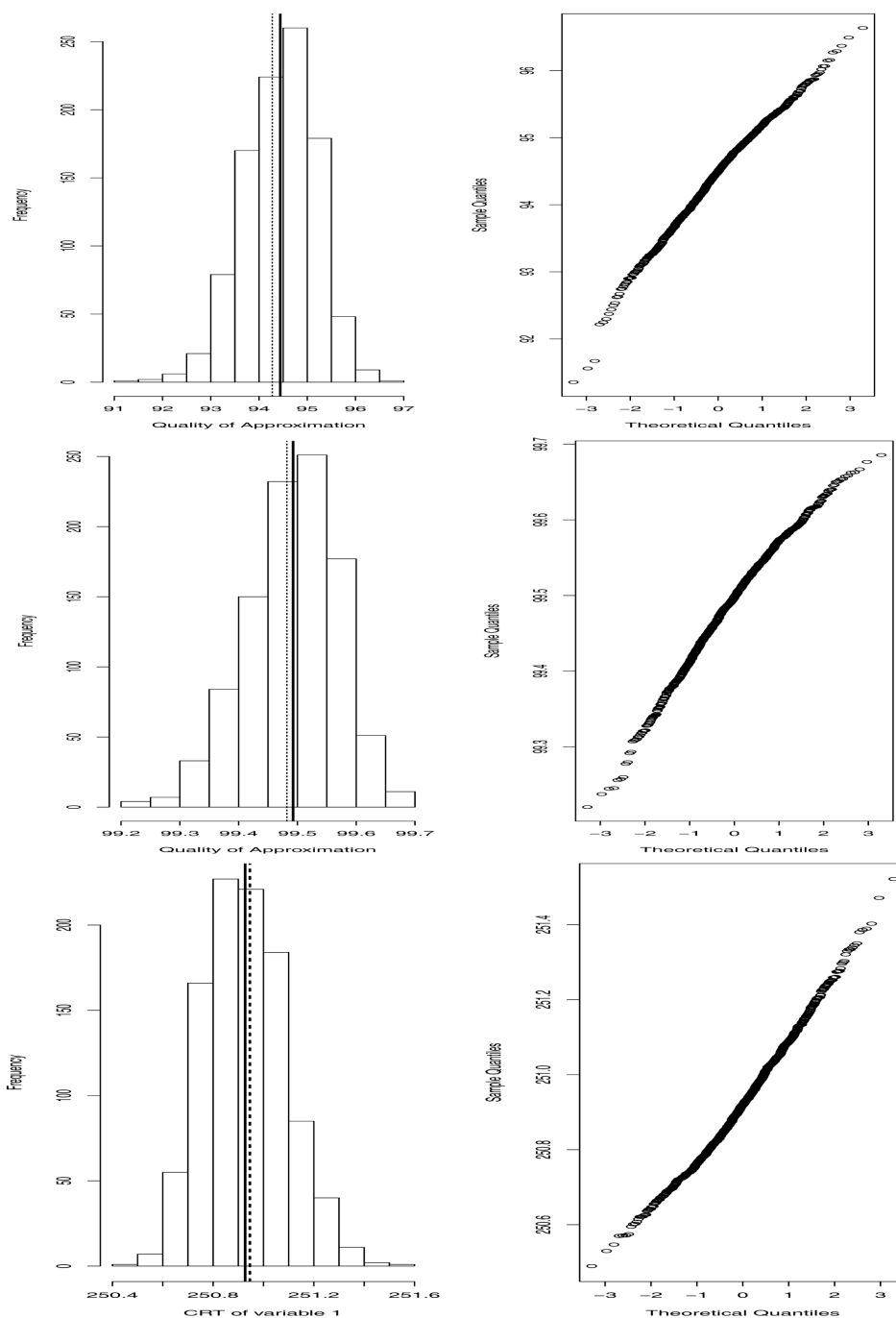


FIGURE 8: Histograms (left) and QQ plots (right) for quality of approximation with simulated (1st panel) and iris (2nd panel) data sets and RCs to total variability of the first variable presented with iris data (3rd panel).

## 5. Discussion and Conclusions

Factorial analysis techniques only provide to researchers point estimates for their results. In this work, we have proposed a methodology that combines bootstrap and biplots methods to calculate confidence intervals for the results from biplots in order to provide measures of their accuracy. This idea has been applied in several multivariate techniques that incorporate a singular value decomposition. Despite there are some packages in the R software to perform biplots, such as detailed in this paper, these packages only provide estimated results as point values and no information about their accuracy is available. For such a reason, we have developed a new package in this software to implement our methodology.

Specifically, in this paper, we have proposed a graphical methodology based on confidence intervals for the main parameters of biplots based in bootstrapping. These parameters help to interpret the contribution from the elements and axes of the biplot and correspond to goodness of fit, quality of the representation, and variability and relationships among variables. The proposed methodology may be considered as an inferential version of classical biplots and has been implemented in the new `biplotbootGUI` R package. We have detailed the features of this package and validated our methodology with numerical applications based on simulated and real-world data. The numerical results have shown the good performance and potentiality of our methodology, as well as the simple and easy manner to work with the `biplotbootGUI` package. As a supplement to our work, we have also provided a review on the key theoretical contributions and the computational implementations for biplot methods, covering the period from 1971 to the present.

Other ways to calculate measures of accuracy, such as jackknife, Markov chain Monte Carlo and permutation methods, are proposed in the literature, as well as ways to calculate confidence intervals other than the intervals proposed in this paper. In future works, some of these methods may be considered by us to provide different measures of accuracy for the results obtained by biplots methods.

## Acknowledgement

The authors wish to thank the Editors of the Special Issue on “Current Topics in Statistical Graphics”, headed by Dr. Fernando Marmolejo-Ramos, the Editor-in-Chief of the journal, Dr. Leonardo Trujillo, and two anonymous referees for their constructive comments on an earlier version of this manuscript which resulted in this improved version. Our research was partially supported by the Chilean Council for Scientific and Technological Research under the project grant FONDECYT 1120879.

[Recibido: mayo de 2014 — Aceptado: octubre de 2014]

## References

- Adler, D. & Murdoch, D. (2012), *The rgl R package version 0.92.894: 3D visualization device system (open GL)*, R project.  
\*cran.r-project.org/package=rgl
- Amaro, I., Vicente-Villardón, J. & Galindo, M. (2004), 'MANOVA biplot for treatment arrays with two factors based on multivariate general linear models', *Interciencia* **29**, 26–32.
- Anderson, E. (1935), 'The irises of the Gaspé peninsula', *Bulletin of the American Iris Society* **59**, 2–5.
- Bickel, P. & Krieger, A. (1989), 'Confidence bands for a distribution function using the bootstrap', *Journal of the American Statistical Association* **84**, 95–100.
- Bradu, D. & Gabriel, K. (1974), 'Simultaneous statistical inference on interactions in two-way analysis of variance', *Journal of the American Statistical Association* **29**, 428–436.
- Bradu, D. & Gabriel, K. (1978), 'The biplot as a diagnostic tool for models of two-way tables', *Technometrics* **20**, 47–68.
- Cárdenas, O. & Galindo, M. P. (2003), *Biplot with External Information based on Generalized Bilinear Models*, Council of Scientific and Humanistic Development of the Central University of Venezuela, Caracas.
- Cárdenas, O., Galindo, M. & Vicente-Villardón, J. (2007), 'Biplot methods: Evolution and applications', *Revista Venezolana de Análisis de Coyuntura* **13**, 279–303.
- Carlier, A. & Kroonenberg, P. (1996), 'Decompositions and biplots in three-way correspondence analysis', *Psychometrika* **61**, 355–373.
- Caro-Lopera, F., Leiva, V. & Balakrishnan, N. (2012), 'Connection between the Hadamard and matrix products with an application to a matrix-variate Birnbaum-Saunders distribution', *Journal of Multivariate Analysis* **104**, 126–139.
- Chatterjee, S. (1984), 'Variance estimation in factor analysis: An application of the bootstrap', *British Journal of Mathematical and Statistical Psychology* **37**, 252–262.
- Chernick, M. (1999), *Bootstrap Methods: A Practitioner's Guide*, Wiley & Sons, New York, US.
- Chessel, D., Dufour, A., Dray, S., Jombart, T., Lobry, J., Ollier, S. & Thioulouse, J. (2013), *The ADE4 R package version 1.5-2: Analysis of ecological data: Exploratory and Euclidean methods in environmental sciences*, R project.  
\*cran.r-project.org/package=ade4

- Chessel, D., Dufour, A. & Thioulouse, J. (2004), 'The ADE4 R package-I: One-table methods', *R Journal* **4**, 5–10.
- Choulakian, V. (1996), 'Generalized bilinear models', *Psychometrika* **61**, 271–283.
- Daudin, J., Duby, C. & Trécourt, P. (1988), 'Stability of principal components studied by the bootstrap method', *Statistics* **19**, 241–258.
- Del Ferraro, M., Kiers, H. & Giordani, P. (2013), *The ThreeWay R package version 1.1.1: Three-way component analysis*, R project.  
\*cran.r-project.org/package=ThreeWay
- Demey, J., Vicente-Villardón, J., Galindo, M. & Zambrano, A. (2008), 'Identifying molecular markers associated with classifications of genotypes by external logistic biplot', *Bioinformatics* **24**, 28–32.
- Denis, J. (1991), 'Ajustements de modèles lineaires et bilineaires sous contraintes lineaires avec données manquantes', *Statistique Appliquée* **39**, 5–24.
- Díaz-Faes, A., González-Albo, B., Galindo, M. & Bordons, M. (2013), 'HJ-biplot as tool of matrix inspection for bibliometrical data', *Revista Española de Documentación Científica* **36**, 1–16.
- Díaz-García, J., Galea, M. & Leiva, V. (2003), 'Influence diagnostics for multivariate elliptic regression linear models', *Communications in Statistics: Theory and Methods* **32**, 625–641.
- Díaz-García, J. & Leiva, V. (2003), 'Doubly non-central t and F distribution obtained under singular and non-singular elliptic distributions', *Communications in Statistics: Theory and Methods* **32**, 11–32.
- Díaz-García, J., Leiva, V. & Galea, M. (2002), 'Singular elliptic distribution: Density and applications', *Communications in Statistics: Theory and Methods* **31**, 665–681.
- Dray, S. & Dufour, A. (2007), 'The ADE4 package: Implementing the duality diagram for ecologists', *Journal of Statistical Software* **22**, 1–20.
- Dray, S., Dufour, A. & Chessel, D. (2007), 'The ADE4 package-II: Two-table and K-table methods', *R Journal* **7**, 47–52.
- Edelman, A. (1988), 'Eigenvalues and condition numbers of random matrices', *SIAM Journal on Matrix Analysis and Applications* **9**, 543–560.
- Efron, B. (1979), 'Bootstrap methods: Another look at the jackknife', *The Annals of Statistics* **7**, 1–26.
- Efron, B. (1987), 'Better bootstrap confidence intervals', *Journal of the American Statistical Association* **82**, 171–185.
- Efron, B. (1993), *An Introduction into the Bootstrap*, Chapman and Hall, New York, US.

- Egido, J. (2014), *The dynBiplotGUI R package version 1.0.1: full interactive GUI for dynamic biplot*, R project.  
\*cran.r-project.org/web/packages/dynBiplotGUI
- Falguerolles, A. (1995), *Generalized Bilinear Models and Generalized Biplots: Some Examples*, Publications du Laboratoire de Statistique et Probabilités. Université Paul Sabatier, Toulouse.
- Faria, J. & Demetrio, C. (2012), *The bpca R package version 1.0-10: Biplot of multivariate data based on principal component analysis*, R project.  
\*cran.r-project.org/package=bpca
- Frutos, E. & Galindo, M. (2013), *The GGEBiplotGUI R package version 1.0-6: interactive GGE biplots in R*, R project.  
\*cran.r-project.org/package=GGEBiplotGUI
- Frutos, E., Galindo, M. & Leiva, V. (2014), 'An interactive biplot implementation in R for modeling genotype-by-environment interaction', *Stochastic Environmental Research and Risk Assessment* **28**, 1629–1641.
- Gabriel, K. (1971), 'The biplot graphic display of matrices with application to principal component analysis', *Biometrika* **58**, 453–467.
- Gabriel, K., G. M. . V.-V. J. (1998), Use of biplots to diagnose independence models in three-way contingency tables, in J. Blasius & M. Grenacre, eds, 'Visualization of Categorical Data', Academic Press, London, UK, pp. 391–404.
- Gabriel, K. & Zamir, S. (1979), 'Lower rank approximation of matrices by least squares with any choice of weights', *Technometrics* **21**, 489–498.
- Galindo, M. (1986), 'An alternative for simultaneous representation: HJ-biplot', *Questúio* **10**, 12–23.
- Gallego-Álvarez, I., Galindo, M. & Rodríguez-Rosa, M. (2014), 'Analysis of the sustainable society index worldwide: A study from the biplot perspective', *Social Indicators Research* **120**, 29–65.
- García-Sánchez, I., Frías-Aceituno, J. & Rodríguez-Domínguez, L. (2013), 'Determinants of corporate social disclosure in Spanish local governments', *Journal of Cleaner Production* **39**, 60–72.
- Gauch, H. (1988), 'Model selection and validation for yield trials with interaction', *Biometrics* **44**, 705–715.
- Gifi, A. (1990), *Nonlinear Multivariate Analysis*, Wiley, Chichester, UK.
- Gower, J. (1992), 'Generalized biplots', *Biometrika* **79**, 475–493.
- Gower, J., Gardner-Lubbe, S. & Le-Roux, N. (2011), *Understanding Biplots*, Wiley, New York, US.

- Gower, J. & Hand, D. (1996), *Biplots*, Chapman & Hall, London, UK.
- Gower, J. & Harding, S. (1988), 'Nonlinear biplots', *Biometrika* **75**, 445–455.
- Graffelman, J. (2013), *The calibrate R package version 1.7.1: Calibration of scatterplot and biplot axes*, R project.  
\*cran.r-project.org/package=calibrate
- Greenacre, M. J. (1984), *Theory and Application of Correspondence Analysis*, Academic Press, London.
- Greenacre, M. J. (2010), *Biplots in Practice*, Publications of BBVA Foundation, Spain.
- Greenacre, M. J. & Nenadic, O. (2012), *The ca R package version 0.53: simple, multiple and joint correspondence analysis*.  
\*cran.r-project.org/package=ca
- Grosjean, P. (2012), *SciViews-R: A GUI API for R*, MONS, Belgium, www.sciviews.org/SciViews-R.
- Hernández, J. & Vicente-Villardón, J. (2013a), *The NominalLogisticBiplot R package version 0.1: Biplot representations of categorical data*, R project.  
\*cran.r-project.org/web/packages/NominalLogisticBiplot/index.html
- Hernández, J. & Vicente-Villardón, J. (2013b), *The OrdinalLogisticBiplot R package version 0.2: Ordinal logistic biplots*, R project.  
\*cran.r-project.org/web/packages/OrdinalLogisticBiplot/index.html
- Hernández, S. (2005), *Robust Biplot*, PhD Dissertation, University of Salamanca, Spain.
- Holmes, S. (1989), 'Using the bootstrap and the RV coefficient in the multivariate context', *Proceedings of the conference on Data Analysis, Learning Symbolic and Numeric Knowledge* pp. 119–131.
- Jambu, M. (1991), *Exploratory and Multivariate Data Analysis*, Academic Press, Orlando, US.
- Kiers, H. (2004), 'Bootstrap confidence intervals for three-way methods', *Journal of Chemometrics* **18**, 22–36.
- La Grange, A., Le-Roux, N. & Gardner-Lubbe, S. (2009), 'Biplotgui: Interactive biplots in R', *Journal of Statistical Software* **30**, 12–37.
- La Grange, A., Le-Roux, N., Rousseeuw, P., Ruts, I. & Tukey, J. (2013), *The biplotGUI R package version 0.0-7: Interactive biplots*, R project.  
\*cran.r-project.org/package=BiplotGUI
- Lambert, Z., Wildt, A. & Durand, R. (1990), 'Assessing sampling variation relative to number-of-factors criteria', *Educational and Psychological Measurement* **50**, 33–48.



- Lebart, L., Morineau, A. & Piron, M. (1995), *Statistique Exploratoire Multidimensionnelle*, Dunod, Paris, France.
- Leiva, V., Marchant, C., Saulo, H., Aslam, M. & Rojas, F. (2014), 'Capability indices for Birnbaum-Saunders processes applied to electronic and food industries', *Journal of Applied Statistics* **41**, 1881–1902.
- L'Hermier des Plantes, H. (1976), *Structuration Des Tableaux A Trois Indices De La Statistique: Theorie et Application d'une Méthode d'Analyse Conjointe*, Master's thesis, Université Des Sciences et Techniques Du Languedoc, Montpellier.
- Linting, M., Meulman, J. J., Groenen, P. J. F. & Van der Kooij, A. J. (2007), 'Stability of nonlinear principal components analysis. An empirical study using the balanced bootstrap.', *Psychological Methods* **12**(3), 359–379.
- Marcenko, V. & Pastur, L. (1967), 'Distributions of eigenvalues for some sets of random matrices', *Mathematics of the USSR-Sbornik* **1**, 457–483.
- Markos, A. (2012), *The GUI ca R package version 0.1-4: a Tcl/Tk GUI for the functions*, R project.  
\*cran.r-project.org/package=caGUI
- Martín-Rodríguez, J., Galindo, M. & Vicente-Villardón, J. (2002), 'Comparison and integration of subspaces from a biplot perspective', *Journal of Statistical Planning and Inference* **102**, 411–423.
- McKay, B. D. (1981), 'The expected eigenvalue distribution of a large regular graph', *Linear Algebra and Applications* **40**, 203–216.
- Mendes, S., Fernández-Gómez, M., Galindo, M., Morgado, F., Maranhão, P., Azeiteiro, U. & Bacelar-Nicolau, P. (2009), 'The study of bacterioplankton dynamics in the Berlengas archipelago (west coast of Portugal) by applying the HJ-biplot method', *Arquipelago Life and Marine Sciences* **26**, 25–35.
- Meulman, J. J. (1982), *Homogeneity Analysis of Incomplete Data*, DSWO Press, Leiden.
- Milan, L. & Whittaker, J. (1995), 'Application of the parametric bootstrap to models that incorporate a singular value decomposition', *Applied Statistics* **44**, 31–49.
- Nenadic, O. & Greenacre, M. (2007), 'Correspondence analysis in R, with two- and three-dimensional graphics: The ca package', *Journal of Statistical Software* **20**, 1–13.
- Nieto, A., Baccalá, N., Vicente-Galindo, P. & Galindo, M. (2012), *The multi-biplotGUI R package version 0.0-1: Multibiplot analysis*, R project, cran.r-project.org/package=multibiplotGUI.

- 396
- Oksanen, J., Blanchet, F., Kindt, R., Legendre, P., Minchin, P., O'Hara, B., Simpson, G., Solymos, P., Stevens, M. & Wagner, H. (2013), *The vegan R package version 2.0-8: Community ecology*, R project.  
\*cran.r-project.org/package=vegan
- Orfao, A., González, M., San-Miguel, J., Ríos, A., Caballero, M., Sanz, M., Calmuntia, M., Galindo, M. & López-Borrasca, A. (1988), 'Bone marrow histopathologic patterns and immunologic phenotype in B-cell chronic lymphocytic leukaemia', *Blut* **57**, 19–23.
- R-Team (2013), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria., www.R-project.org.
- Ramírez, G., Vásquez, M., Camardiel, A., Pérez, B. & Galindo, M. (2005), 'Graphical detection for the multicollinearity by the h-plot of the inverse matrix of correlations', *Revista Colombiana de Estadística* **28**, 207–219.
- Rivas-Gonzalo, J., Gutiérrez, Y., Polanco, A., Hebrero, E., Vicente-Villardón, J., Galindo, M. & 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**, 302–308.
- Sánchez, L., Leiva, V., Caro-Lopera, F. & Cysneiros, F. (2015), *On matrix-variate Birnbaum-Saunders distributions and their estimation and application*, Brazilian Journal of Probability and Statistics.  
\*http://dx.doi.org/10.1214/14-BJPS247 (in press)
- Sepúlveda, R., Vicente-Villardón, J. & Galindo, M. (2008), 'The biplot as a diagnostic tool of local dependence in latent class models: A medical application', *Statistics in Medicine* **27**, 1855–1869.
- Stewart, G. (1980), 'The efficient generation of random orthogonal matrices with application to condition estimators', *SIAM Journal on Numerical Analysis* **17**, 403–409.
- Ter-Braak, C. (1986), 'Canonical correspondence analysis: A new eigenvector technique for multivariate direct gradient analysis', *Ecology* **5**, 1167–1179.
- Ter-Braak, C. (1990), 'Interpreting canonical correlation analysis through biplot of structure and weights', *Psychometrika* **55**, 519–531.
- Ter-Braak, C. & Looman, C. (1994), 'Biplots in reduced-rank regression', *Biometrical Journal* **36**, 983–1003.
- Thioulouse, J. & Dray, S. (2007), 'Interactive multivariate data analysis in R with the ade4 and ade4TkGUI packages', *Journal of Statistical Software* **22**, 1–14.
- Thioulouse, J. & Dray, S. (2012), *The ade4TkGUI R package version 0.2-6: ade4 Tcl/Tk graphical user interface*, R project.  
\*cran.r-project.org/package=ade4TkGUI

- Tierney, L. (2012), *The tkrplot R package version 0.0-23: TK Rplot*, R project.  
\*cran.r-project.org/package=tkrplot
- Timmerman, M., Kiers, H., Smilde, A. & Stouten, J. (2009), 'Bootstrap confidence intervals in multi-level simultaneous component analysis', *British Journal of Mathematical and Statistical Psychology* **62**, 299–318.
- Tucker, L. (1966), 'Some mathematical notes on three-mode factor analysis', *Psychometrika* **31**, 279–311.
- Vairinhos, V. (2003), Development of a System for Data Mining based on Biplot Methods, PhD Dissertation, University of Salamanca, Spain.
- Vallejo-Arboleda, A., Vicente-Villardón, J. & Galindo, M. (2006), 'Canonical STATIS: Biplot analysis of multi-table group structured data based on STATIS-ACT methodology', *Computational Statistics & Data Analysis* **51**, 4193–4205.
- Vallejo-Arboleda, A., Vicente-Villardón, J., Galindo, M., Fernández, M., Fernández, C. & Bécares, E. (2008), 'Analysis of time evolution for group structured data: Canonical dual statis and doubly multivariate repeated measures model', *Revista Colombiana de Estadística* **31**, 321–340.
- Van Ginkel, J. and Kiers, H. (2011), 'Constructing bootstrap confidence intervals for principal component loadings in the presence of missing data: A multiple-imputation approach', *British Journal of Mathematical and Statistical Psychology* **64**, 498–515.
- Vicente-Villardón, J. (2010), *MULTBILOT: A Package for Multivariate Analysis using Biplots*, Matlab software.  
\*biplot.usal.es/ClassicalBiplot/index.html
- Vicente-Villardón, J., Galindo, M. & Blázquez, A. (2006), *Logistic Biplots*, Chapman & Hall, New York, US.
- Viloria, J., Gil, J., Durango, D. & García, C. (2012), 'Physicochemical characterization of propolis from the region of Bajo Cauca Antioqueño (Antioquia, Colombia)', *Biotecnología en el Sector Agropecuario y Agroindustrial* **10**, 77–86.
- Wachter, K. (1978), 'The strong limits of random matrix spectra for sample matrices of independent elements', *The Annals of Probability* **6**, 1–18.
- Yan, W., Hunt, L., Sheng, Q. & Szlavnic, Z. (2000), 'Cultivar evaluation and mega-environment investigation based on GGE biplot', *Crop Science* **40**, 597–605.
- Yan, W. & Kang, M. (2003), *GGE Biplot Analysis: A Graphical Tool for Breeders, Geneticists, and Agronomists*, CRC Press, Boca Raton, US.



---

# Bibliografía

- H. Abdi, J.P. Dunlop, y L. J. Williams. How to compute reliability estimates and display confidence and tolerance intervals for pattern classifiers using the Bootstrap and 3-way multidimensional scaling (DISTATIS). *Neuroimage*, 45:89–95, 2009.
- H. Abdi, D. Valentin, S. Chollet, y C. Chrea. Analyzing assessors and products in sorting tasks: DISTATIS, theory and applications. *Food Quality Prefer*, 18:627–640, 2007.
- H. Abdi, L. J. Williams, y D. Valentin. Multiple factor analysis: principal component analysis for multitable and multiblock data sets. *Wiley interdisciplinary reviews: Computational Molecular Science*, 5(2):149–179, 2013.
- H. Abdi, L. J. Williams, D. Valentin, y M. Bennani-Dosse. STATIS and DISTATIS: optimum multi-table principal component analysis and three way metric multidimensional scaling. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4, 2012.
- D. Adler y D. Murdoch. The rgl R package version 0.92.894: 3D visualization device system (open GL). [cran.r-project.org/package=rgl](http://cran.r-project.org/package=rgl), 2012.

- M. Alcántara y C. Rivas. Las dimensiones de la polarización partidista en América Latina. *Política y gobierno*, 14(2):349–390, 2007.
- I. R. Amaro, J. L. Vicente-Villardón, y M. P. Galindo. MANOVA biplot for treatment arrays with two factors based on multivariate general linear models. *Interciencia*, 29:26–32, 2004.
- E. Anderson. The irises of the Gaspé Peninsula. *Bulletin of the American Iris Society*, 59:2–5, 1935.
- P. Arabie y L. Hubert. Cluster analysis in marketing research. En R. P. Bagozzi, ed., *Handbook of marketing research*. Blackwell, Oxford, UK, 1994.
- N. Baccalá. *Contribuciones al análisis de matrices de datos multivía: tipología de las variables*. Phd dissertation, Universidad de Salamanca, 2004.
- S. Balbi. On stability in non symmetrical correspondence analysis using bootstrap. *Statistica Applicata*, 4(4):543–552, 1992.
- D. Beaton, C. R. Chin Fatt, y H. Abdi. The DistatisR package version 1.0: DiSTATIS three way metric multidimensional scaling. [cran.r-project.org/web/packages/DistatisR](http://cran.r-project.org/web/packages/DistatisR), 2013.
- D. Beaton, C. R. C. Fatt, y H. Abdi. An ExPosition of multivariate analysis with the singular value decomposition in R. *Computational Statistics and Data Analysis*, 72(0):176–189, 2014.
- M. Bécue-Bertaut y J. Pagès. A principal axes method for comparing multiple contingency tables: Mfact. *Computational Statistics and Data Analysis*, 45:481–503, 2004.

- M. Bécue-Bertaut y J. Pagès. Multiple factor analysis and clustering of a mixture of quantitative, categorical and frequency data. *Computational Statistics and Data Analysis*, 52:3255–3268, 2008.
- J. Benasseni y M. Bennani-Dosse. Analyzing multiset data by the power STATIS-ACT method. *Adv Data Anal Classif*, 6(1):49–65, 2012.
- J. P. Benzècri. *L'analyse des données: Tome 1: La taxinomie. Tome 2: L'analyse des correspondance*. Dunod, Paris, France, 1973.
- J. P. Benzècri. Sur l'analyse des tableaux binaires associés á une correspondance multiple [The analysis of boolean tables associated with a multiple correspondence]. *Les Cahiers de l'Analyse des Données*, 2:55–71, 1977.
- H. H. Bock. Simultaneous clustering of objects and variables. En R. Tomassone, ed., *Analyse des données et informatique*. INRIA, Le Chesnay, France, 1979.
- H. H. Bock. On the interface between cluster analysis, principal component analysis, and multidimensional scaling. En H. Bozdogan y A. K. Gupta, eds., *Multivariate Statistical Modeling and Data Analysis*, págs. 17–34. Reidel, New York, USA, 1987.
- H. H. Bock. Two-way clustering for contingency tables maximizing a dependence measure. En M. Schader, W. Gaul, y M. Vichi, eds., *Between Data Science and Applied Data Analysis*, págs. 143–155. Springer, Heidelberg, 2003.
- U. Böckenholt y I. Böcknholt. Canonical analysis of contingency tables with linear constraints. *Psychometrika*, 55(4):633–639, 1990.
- U. Böckenholt y Y. Takane. Linear constraints in correspondence analysis. En Greenacre M. J. y Blasius J., eds., *Correspondence Analysis in Social Sciences*, págs. 112–127. Academic Press, 1994.

- R. J. Boik. An efficient algorithm for joint correspondence analysis. *Psychometrika*, 61:255–269, 1996.
- D. Bradu y K. R. Gabriel. Simultaneous statistical inference on interactions in two-way analysis of variance. *Journal of the American Statistical Association*, 29:428–436, 1974.
- D. Bradu y K. R. Gabriel. The biplot as a diagnostic tool for models of two-way tables. *Technometrics*, 20:47–68, 1978.
- N. Cadaviz. *Neuropsicología de la construcción de la función ejecutiva*. Phd dissertation, University of Salamanca, Spain, 2008.
- J. Cadima y I. T. Jolliffe. Loadings and correlations in the interpretation of principal components. *Journal of Applied Statistics*, 22:203–214, 1995.
- M. Cadoret y F. Husson. Construction and evaluation of confidence ellipses applied at sensory data. *Food Quality and Preference*, 28:106–115, 2013.
- O. Cárdenas y M. P. Galindo. *Biplot with external information based on generalized bilinear models*. Council of Scientific and Humanistic Development of the Central University of Venezuela, Caracas, Venezuela, 2003.
- O. Cárdenas, M. P. Galindo, y J. L. Vicente-Villardón. Biplot methods: evolution and applications. *Revista Venezolana de Análisis de Coyuntura*, 13:279–303, 2007.
- A. Carlier y P. M. Kroonenberg. Decompositions and biplots in three-way correspondence analysis. *Psychometrika*, 61:355–373, 1996.



- E. Castela y M. P. Galindo. Ecological Inference for the characterization of electoral turnout: The Portuguese Case. *CIEO-Research Centre for Spatial and Organizational Dynamics, University of Algarve*, 2010.
- S. Chatterjee. Variance estimation in factor analysis: an application of the bootstrap. *British Journal of Mathematical and Statistical Psychology*, 37:252–262, 1984.
- D. Chessel, A. B. Dufour, S. Dray, T. Jombart, J. R. Lobry, S. Ollier, y J. Thioulouse. The ade4 R package version 1.5-2: analysis of ecological data: exploratory and Euclidean methods in environmental sciences. [cran.r-project.org/package=ade4](http://cran.r-project.org/package=ade4), 2013.
- D. Chessel, A. B. Dufour, y J. Thioulouse. The ade4 R package-I: one-table methods. *R Journal*, 4(1):5–10, 2004.
- D. Chessel y M. Hanafi. Analyse de la co-inertie de K nuages de points. *Revue de Statistique Appliquée*, 44:35–60, 1996.
- D. Chessel y P. Mercier. Couplage de triplets statistiques et liaisons espèces-environment. En J. D. Lebreton y B. Asselain, eds., *Biométrie et Environment*, págs. 15–44. Mason: Paris, 1993.
- C. R. Chin Fatt, D. Beaton, y H. Abdi. The MExPosition R package version 2.0.3: multi-table ExPosition. [cran.r-project.org/web/packages/MExPosition](http://cran.r-project.org/web/packages/MExPosition), 2013.
- V. Choulakian. Generalized bilinear models. *Psychometrika*, 61:271–283, 1996.
- D. Corrales y O. Rodríguez. INTERSTATIS: The STATIS method for interval valued data. *Revista de Matemática: Teoría y Aplicaciones*, 21(1):73–83, 2014.

- G. Correa, L. L. Lavalett, M. P. Galindo, y L. Afanador. Uso de métodos multivariantes para la agrupación de aislamientos de *Colletotrichum spp.* con base en características morfológicas y culturales. *Revista Facultad Nacional de Agronomía de Medellín*, 60(1):3671–3690, 2007.
- P. Cortez, A. Cerdeira, F. Almeida, T. Matos, y J. Reis. Modeling wine preferences by data mining from physicochemical properties. En *Decision Support Systems*, tomo 47, págs. 547–553. Elsevier, 2009.
- T. F. Cox y M. A. A. Cox. *Multidimensional Scaling*. Chapman & Hall, London, UK, 1994.
- A. C. Culhane, J. Thioulouse, G. Perriere, y D. G. Higgins. MADE4: an R package for multivariate analysis of gene expression data. *Bioinformatics*, 21(11):2789–2790, 2005.
- A. D'Aspremont, L. El Ghaoui, M. I. Jordan, y G. R. G. Lanckriet. A direct formulation for sparse pca using semidefinite programming. *SIAM*, 49(3):434–448, 2007.
- L. D'Ambra y E. J. Beh. Non symmetrical correspondence analysis with concatenation and linear constraints. *Australian and New Zealand Journal of Statistics*, 52(1):27–44, 2010.
- L. D'Ambra y N. Lauro. Non symmetrical analysis of three-way contingency table. En R. Coppi y S. Bolasco, eds., *Multway Data Analysis*, págs. 301–315. Elsevier, Amsterdam, 1989.
- L. D'Ambra y N. Lauro. Exploratory non symmetric data analysis. *Italian J. Appl. Statist.*, 4:511–529, 1992.

- L. D'Ambrá y N. Lauro. Normalized non symmetrical correspondence analysis for three-way data sets. En *Bulletin of the ISI. Contributed papers 49th session*, tomo 1, págs. 301–302. Rome, Italy: International Statistical Society, 1993.
- L. D'Ambrá, B. Simonetti, y E. J. Beh. A dimensional reduction method for ordinal three-way contingency table. En A. Rizzi y M. Vichi, eds., *Compstat 2006 - Proceedings in Computational Statistics*, págs. 271–283. Physica-Verlag HD, 2006.
- J. Daudin, C. DUBY, y P. Trécourt. Stability of principal components studied by the bootstrap method. *Statistics*, 19:241–258, 1988.
- J. De Leeuw y P. Mair. Simple and canonical correspondence analysis using the R package anacor. *Journal of Statistical Software*, 31(5):1–18, 2009.
- G. De Soete y J. D. Carroll. k-means clustering in a low-dimensional Euclidean space. En Diday E., Y. Lechevallier, M. Schader, P. Bertrand, y B. Burtschy, eds., *New Approaches in Classification and Data Analysis*, págs. 212–219. Springer, Berlin Heidelberg, 1994.
- C. Dehlholm, P. B. Brockhoff, y W. L. P. Bredie. Confidence ellipses: a variation based on parametric bootstrapping applicable on multiple factor analysis results for rapid graphical evaluation. *Food Quality and Preference*, 26:278–280, 2012.
- M. A. Del Ferraro, H. A. L. Kiers, y P. Giordani. The Three-Way R package version 1.1.1: three-way component analysis. [cran.r-project.org/package=ThreeWay](http://cran.r-project.org/package=ThreeWay), 2013.
- J. R. Demey, J. L. Vicente-Villardón, M. P. Galindo, y A. Y. Zambrano.

- Identifying molecular markers associated with classifications of genotypes by external logistic biplot. *Bioinformatics*, 24:28–32, 2008.
- J. B. Denis. Ajustements de modèles lineaires et bilineaires sous contraintes lineaires avec données manquantes. *Statistique Appliquée*, 39:5–24, 1991.
- W. S. DeSarbo. GENNCLUS: new models for general non-hierarchical clustering analysis. *Psychometrika*, 47:449–475, 1982.
- W. S. DeSarbo y W. J. Heiser. A latent class unfolding model for analyzing single stimulus preference ratings. *Psychometrika*, 58:545–565, 1993.
- W. S. DeSarbo, D. J. Howard, y K. Jedidi. MULTICLUS: a new method for simultaneous performing multidimensional scaling and clustering. *Psychometrika*, 56:121–136, 1990a.
- W. S. DeSarbo, K. Jedidi, K. Cool, y D. Schendel. Simultaneous multidimensional unfolding and cluster analysis: an investigation of strategic groups. *Marketing Letters*, 2:129–146, 1990b.
- P. Diaconis y B. Efron. Computer intensive methods in statistics. *Scientific American*, 248:116–130, 1983.
- A. A. Díaz-Faes, B. González-Albo, M. P. Galindo, y M. Bordons. HJ-biplot as tool of matrix inspection for bibliometrical data. *Revista Española de Documentación Científica*, 36:1–16, 2013.
- S. Dolédec y D. Chessel. Co-inertia analysis: an alternative method for studying species environment relationships. *Freshwater Biology*, 31:277–294, 1994.
- S. Dolédec, D. Chessel, C. J. F. Ter Braak, y S. Champely. Matching species traits to environmental variables: a new three-table ordination method. *Environmental and Ecological Statistics*, 3:143–166, 1996.

- A. Dorado, S. Vicente, A. Blazquez, y J. Martin. Análisis HJ Biplot de la evolución de la productividad agraria de la comunidad de Castilla y León a lo largo del quinquenio 1991-1995. *Investigación agraria. Producción y protección vegetales*, 14(3):515–530, 1999.
- S. Dray, D. Chessel, y J. Thioulouse. Procrustean co-inertia analysis for the linking of multivariate datasets. *Ecoscience*, 10(1):110–119, 2003.
- S. Dray y A. B. Dufour. The ade4 package: implementing the duality diagram for ecologists. *Journal of Statistical Software*, 22(4):1–20, 2007.
- S. Dray, A. B. Dufour, y D. Chessel. The ade4 package-II: Two-table and K-table methods. *R Journal*, 7(2):47–52, 2007.
- S. Dray, N. Pettorelli, y D. Chessel. Matching data sets from two different spatial samplings. *Journal of Vegetation Science*, 13:867–874, 2002.
- P. Drineas, R. Kannan, y M. W. Mahoney. Fast Monte-Carlo algorithms for matrices III: Computing a compressed approximate matrix decomposition. *SIAM Journal on Computing*, 36:184–206, 2006.
- P. Drineas, M. W. Mahoney, y S. Muthukrishnan. Relative-error cur matrix decompositions. *Journal on Matrix Analysis*, 30:844–881, 2008.
- I. L. Dryden. The shapes R package version 1.1-10: statistical shape analysis. <http://CRAN.R-project.org/package=shapes>, 2014.
- G. Eckart y G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1:211–218, 1936.
- B. Efron. Bootstrap methods: another look at the jackknife. *Annals of Statistics*, 7:1–26, 1979.

- B. Efron. Nonparametric standard errors and confidence intervals (with discussions). *Canadian Journal of Statistics*, 82:171–185, 1981.
- B. Efron. *The jackknife, the bootstrap, and other resampling plans*. SIAM, Philadelphia, US, 1982.
- B. Efron. Better bootstrap confidence intervals (with discussions). *Journal of the American Statistical Association*, 82(397):171–185, 1987.
- B. Efron y R. J. Tibshirani. *An introduction into the bootstrap*. Chapman and Hall, New York, US, 1993.
- J. Egido. The dynBiplotGUI R package version 1.0.1: full interactive GUI for dynamic biplot. [cran.r-project.org/web/packages/dynBiplotGUI](http://cran.r-project.org/web/packages/dynBiplotGUI), 2014.
- B. Escoufier y J. Pagès. *Analyse factorielle multiple*. Cahiers du BURO, 2, ISUP, Paris, 1984.
- B. Escoufier y J. Pagès. *Analyses factorielles simples et multiples: objectifs, méthodes et interprétation*. Dunod, Paris, 1990.
- Y. Escoufier. The duality diagram: a means for better practical applications. En Pierre Legendre y Louis Legendre, eds., *Develoments in Numerical Ecology*, tomo 14 de *NATO ASI Series*, págs. 139–156. Springer Berlin Heidelberg, 1987.
- V. Esposito. Un ´analisi non simmetrica comparativa con osservazioni stratificate. En *S.I.S. Conference La Statistica per le Imprese*. 1997.
- A. Falguerolles. *Generalized bilinear models and generalized biplots: some examples*. Publications du Laboratoire de Statistique et Probabilités. Université Paul Sabatier, Toulouse, Francia, 1995.

- J. C. Faria y C. G. B. Demetrio. The bpca R package version 1.0-10: biplot of multivariate data based on principal component analysis. [cran.r-project.org/package=bpca](http://cran.r-project.org/package=bpca), 2012.
- J. A. Felício y M. P. Galindo. Governance Mechanisms and Performance of Publicly Traded Companies. *International Journal of Business and Management*, 9(12):1–15, 2014.
- A. Fisher. The bootSVD R package version 0.5: Fast, exact bootstrap principal component analysis for high dimensional data. [cran.r-project.org/package=bootSVD](http://cran.r-project.org/package=bootSVD), 2015.
- A. Fisher, B. Caffo, B. Schwartz, y V. Zipunnikov. Fast, exact bootstrap principal component analysis for  $p > 1$  million. 2014. URL <http://arxiv.org/abs/1405.0922v3>.
- A. Frieze, R. Kannan, y S. Vempala. Fast monte-carlo algorithms for finding low-rank approximations. *Journal of the ACM*, 51:1025–1041, 2004.
- E. Frutos y M. P. Galindo. The GGEBiplotGUI R package version 1.0-6: interactive GGE biplots in R. [cran.r-project.org/package=GGEBiplotGUI](http://cran.r-project.org/package=GGEBiplotGUI), 2013.
- E. Frutos, M. P. Galindo, y V. Leiva. An interactive biplot implementation in R for modeling genotype-by-environment interaction. *Stochastic Environmental Research and Risk Assessment*, 28(7):1629–1641, 2014.
- K. R. Gabriel. The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58:453–467, 1971.
- K. R. Gabriel. MANOVA Biplots for two-way contingency tables. En W. J.

- Krzanowski, ed., *Recent Advances in Descriptive Multivariate Analysis*, págs. 268–277. Oxford Science Publications, Oxford, 1995.
- K. R. Gabriel, M. P. Galindo, y J. L. Vicente-Villardón. Use of biplots to diagnose independence models in three-way contingency tables. En J. Blasius y M. Grenacre, eds., *Visualization of Categorical Data*, págs. 391–404. Academic Press, London, UK, 1998.
- K. R. Gabriel y S. Zamir. Lower rank approximation of matrices by least squares with any choice of weights. *Technometrics*, 21:489–498, 1979.
- E. Galante, M. García-Román, I. Barrera, y M. P. Galindo. Comparison of spatial distribution patterns of dung-feeding scarabs (Coleoptera: Scarabaeidae, Geotrupidae) in wooded and open pastureland in the Mediterranean “dehesa” area of the Iberian peninsula. *Environmental Entomology*, 20(1):90–97, 1991.
- M. P. Galindo. An alternative for simultaneous representation: HJ-Biplot. *Questúo*, 10:12–23, 1986.
- M. P. Galindo, I. Barrera, M. J. Fernández, y A. Martín. Estudio comparativo de ordenación de comunidades ecológicas basado en técnicas factoriales. *Mediterranea*, Serie de estudios biológicos:55–61, 1996.
- I. Gallego-Álvarez, H. Formigoni, y M. T. Pompa. Corporate social responsibility practices at brazilian firms. *Revista de Administración de Empresas*, 54(1):12–27, 2014a.
- I. Gallego-Álvarez, M. P. Galindo, y M. Rodríguez-Rosa. Analysis of the sustainable society index worldwide: a study from the biplot perspective. *Social Indicators Research*, págs. 1–37, 2014b.



- I. Gallego-Álvarez, L. Rodríguez-Domínguez, y R. García-Rubio. Analysis of environmental issues worldwide: a study from the biplot perspective. *Journal of Cleaner Production*, 42:19–30, 2013.
- I. Gallego-Álvarez, M. P. Vicente, M. P. Galindo, y M. Rodríguez-Rosa. Environmental performance in countries worldwide: determinant factors and multivariate analysis. *Sustainability*, 6:7807–7832, 2014c.
- I. Gallego-Álvarez y J. L. Vicente-Villardón. Analysis of environmental indicators in international companies by applying the logistic biplot. *Ecological Indicators*, 23:250–261, 2012.
- J. J. García, G. A. Correa, y S. C. Pardo-Carrasco. Phytoplankton and periphyton in ponds with Nile tilapia (*Oreochromis niloticus*) and bocachico (*Prochilodus magdalenae*). *Revista Colombiana de Ciencias Pecuarias*, 25:603–614, 2012.
- I. M. García-Sánchez, J. V. Frías-Aceituno, y L. Rodríguez-Domínguez. Determinants of corporate social disclosure in Spanish local governments. *Journal of Cleaner Production*, 39:60–72, 2013.
- J. García-Talegón, M. A. Vicente, E. Molina-Ballesteros, y S. Vicente-Tavera. Determination of the origin and evolution of building stones as a function of their chemical composition using the inertia criterion based on an HJ-biplot. *Chemical Geology*, 153:37–51, 1999.
- H. G. Gauch. Model selection and validation for yield trials with interaction. *Biometrics*, 44:705–715, 1988.
- A. Gifi. *Nonlinear multivariate analysis*. Wiley, Chichester, UK, 1990.

- C. Gimaret-Carpentier, D. Chessel, y J. P. Pascal. Non-symmetric correspondence analysis: an alternative for species occurrences data. *Plant Ecology*, 138(1):97–112, 1998.
- J. M. González, M. R. Fidalgo, M. J. Martín, y S. Vicente. Study of the evolution of air pollution in Salamanca (Spain) along a five-year period (1994–1998) using HJ-Biplot simultaneous representation analysis. *Environmental Modelling and Software*, 21:61–68, 2006.
- S. A. Goreinov y E. E. Tyrtshnikov. The maximum-volume concept in approximation by low-rank matrices. *Contemporary Mathematics*, 280:47–51, 2001.
- J. C. Gower. Generalized biplots. *Biometrika*, 79:475–493, 1992.
- J. C. Gower, S. Gardner-Lubbe, y N. J. Le Roux. *Understanding biplots*. Wiley, New York, US, 2011.
- J. C. Gower y D. Hand. *Biplots*. Chapman & Hall, London, UK, 1996.
- J. C. Gower y S. A. Harding. Nonlinear biplots. *Biometrika*, 75:445–455, 1988.
- J. Graffelman. The calibrate R package version 1.7.1: calibration of scatterplot and biplot axes. [cran.r-project.org/package=calibrate](http://cran.r-project.org/package=calibrate), 2012.
- P. E. Green, F. Carmone, y S. M. Smith. *Multidimensional Scaling: Concept and Applications*. Allyn & Bacon, Boston, USA, 1989.
- M. J. Greenacre. *Theory and application of correspondence analysis*. Academic Press, London, UK, 1984.
- M. J. Greenacre. Correspondence analysis of multivariate categorical data by weighted least-squares. *Biometrika*, 75:457–467, 1988.

- M. J. Greenacre. Some limitations of correspondence analysis. *Computational Statistics Quarterly*, 3:249–256, 1990.
- M. J. Greenacre. Interpreting multiple correspondence analysis. *Applied Stochastic Models and Data Analysis*, 7(2):195–210, 1991.
- M. J. Greenacre. *Biplots in practice*. BBVA Foundation, Spain, 2010.
- M. J. Greenacre y O. Nenadic. The ca R package version 0.53: simple, multiple and joint correspondence analysis. [cran.r-project.org/package=ca](http://cran.r-project.org/package=ca), 2012.
- P. Grosjean. *SciViews-R: a GUI API for R*. MONS, Belgium, 2012. URL [www.sciviews.org/SciViews-R](http://www.sciviews.org/SciViews-R).
- J. F. Hair, R. E. Anderson, y W. C. Tatham, R. L. Black. *Multivariate Data Analysis*. Prentice Hall Inc, 1998.
- J. A. Hartigan. Direct clustering of a data matrix. *Journal of American Statistical Association*, 67:123–129, 1972.
- J. A. Hartigan. *Clustering Algorithms*. Wiley, New York, US, 1975.
- W. J. Heiser. Clustering in low-dimensional space. En O. Opitz, B. Lausen, y R. Klar, eds., *Information and Classification: Concepts, Methods and Applications*, págs. 162–173. Springer, Berlin Heidelberg, 1993.
- W. J. Heiser y P. J. F. Groenen. Cluster differences scaling with a within-clusters loss component and a fuzzy successive approximation strategy to avoid local minima. *Psychometrika*, 62:63–83, 1997.
- W. J. Heiser y J. J. Meulman. Constrained multidimensional scaling including confirmation. *Applied Psychological Measurement*, 7:381–404, 1983.

- J. C. Hernández y J. L. Vicente-Villardón. The NominalLogisticBiplot R package version 0.1: biplot representations of categorical data. [cran.r-project.org/web/packages/NominalLogisticBiplot/index.html](http://cran.r-project.org/web/packages/NominalLogisticBiplot/index.html), 2013a.
- J. C. Hernández y J. L. Vicente-Villardón. The OrdinalLogisticBiplot R package version 0.2: ordinal logistic biplots. [cran.r-project.org/web/packages/OrdinalLogisticBiplot/index.html](http://cran.r-project.org/web/packages/OrdinalLogisticBiplot/index.html), 2013b.
- S. Hernández. *Robust Biplot*. Phd dissertation, Universidad de Salamanca, Spain, 2005.
- M. O. Hill. Decorana - a fortran program for detrended correspondence analysis an reciprocal averaging. 1979a.
- M. O. Hill. Twinspan - a fortran programme for arranging multivariate data in an ordered two-way table by classification of individuals and attributes. 1979b.
- M. O. Hill y H. G. Gauch. Detrended correspondence analysis: an improved ordination technique. *Vegetatio*, 42:47–58, 1980.
- S. Holmes. *Outils informatiques pour l'évaluation de la pertinence d'un résultat en analyse des données*. USTL, Montpellier, Francia, 1985.
- S. Holmes. Using the bootstrap and the RV coefficient in the multivariate context. En E. Diday, ed., *Proceedings of the conference on Data Analysis, Learning Symbolic and Numeric Knowledge*, págs. 119–131. Nova Science, New York, US, 1989.
- R. Horn y C. Johnson. *Matrix Analysis*. University Press, Cambridge, 1985.

- F. Husson, J. Josse, S. Lê, y J. Mazet. The FactoMineR package version 1.20: factor analysis and data mining with R. <http://CRAN.R-project.org/package=FactoMineR>, 2012.
- F. Husson, S. Le Dien, y J. Pagès. Confidence ellipse for the sensory profiles obtained by principal component analysis. *Food Quality and Preference*, 16:245–250, 2005.
- H. Hwang y Y. Takane. Generalized constrained multiple correspondence analysis. *Psychometrika*, 67:215–228, 2002.
- A. C. Iñigo, F. J. López-Moro, S. Vicente-Tavera, y V. Rives. Monitoring of origin and evolution of building stones through their major components. *Journal of Materials in Civil Engineering*, 17(4):440–446, 2005.
- P. A. Jaffrenou. *Sur l'analyse des familles finies de variables vectorielles: bases algébriques et applications à la description statistique*. Phd dissertation, University of Lyon, France, 1978.
- M. Jambu. *Exploratory and multivariate data analysis*. Academic Press, Orlando, US, 1991.
- G. Jefferis. The dendroextras R package version 0.2.1: Extra functions to cut, label and colour dendrogram clusters. [cran.r-project.org/package=dendroextras](http://cran.r-project.org/package=dendroextras), 2014.
- I. T. Jolliffe. Rotation of principal components: choice of normalization constraints. *Journal of Applied Statistics*, 22:29–35, 1995.
- I. T. Jolliffe, N. T. Trendafilov, y M. Uddin. A modified principal component technique based on the lasso. *Journal of Computational and Graphical Statistics*, 12:531–547, 2003.

- H. A. L. Kiers. Comparison of “Anglo-Saxo” and “French” three-mode methods. *Statistique et Analyse des Données*, 13:14–32, 1988.
- H. A. L. Kiers. Hierarchical relations among three-way methods. *Psychometrika*, 56:449–470, 1991.
- H. A. L. Kiers. Bootstrap confidence intervals for three-way methods. *Journal of Chemometrics*, 18:22–36, 2004.
- H. A. L. Kiers, D. Vicari, y M. Vichi. Simultaneous classification and multidimensional scaling with external information. *Psychometrika*, 70(3):433–460, 2005.
- R. G. Knox y R. K. Peet. Bootstrapped ordination: a method for estimating samplings effects in indirect gradient analysis. *Vegetatio*, 80:153–165, 1989.
- P. M. Kroonenberg y R. Lombardo. Nonsymmetric correspondence analysis: a tool for analysing contingency tables with a dependence structure. *Multivariate Behavioral Research*, 34(3):367–396, 1999.
- A. M. La Grange, N. J. Le Roux, y S. Gardner-Lubbe. BiplotGUI: interactive biplots in R. *Journal of Statistical Software*, 30(12):1–37, 2009.
- A. M. La Grange, N. J. Le Roux, P. J. Rousseeuw, I. Ruts, y J. W. Tukey. The biplotGUI R package version 0.0-7: interactive biplots. [cran.r-project.org/package=BiplotGUI](http://cran.r-project.org/package=BiplotGUI), 2013.
- R. Lafosse y M. Hanafi. Concordance d’un tableau avec k tableaux: définition de k+1 uples synthétiques. *Revue de Statistique Appliquée*, 45(4):111–126, 1997.

- Z. V. Lambert, A. R. Wildt, y R. M. Durand. Assessing sampling variation relative to number-of-factors criteria. *Educational and Psychological Measurement*, 50:33–48, 1990.
- Z. V. Lambert, A. R. Wildt, y R. M. Durand. Approximating confidence intervals for factor loadings. *Multivariate Behavioral Research*, 26:421–434, 1991.
- A. G. Larsen. The soc.ca R package version 0.7.1: specific correspondence analysis for the social sciences. <http://CRAN.R-project.org/package=soc.ca>, 2014.
- N. Lauro y L. D'Ambrà. L'analyse non symétrique des correspondances. En E. L. Diday, ed., *Data Analysis and Informatics III*, págs. 433–446. Amsterdam: North-Holland, 1984.
- S. Le y F. Husson. SensoMineR: a package for sensory data analysis. *Journal of Sensory Studies*, 23(1):14–25, 2008.
- S. Lê, J. Josse, y F. Husson. FactomineR: an R package for multivariate analysis. *Journal of Statistical Software*, 25(1):1–18, 2008.
- S. Lê y J. Pagès. DMFA: dual multiple factor analysis. En *Proceedings of the 12th International Conference on Applied Stochastic Models and Data Analysis*. 2007.
- S. Le Dien y J. Pagès. Hierarchical multiple factor analysis: application to the comparison of sensory profiles. *Food Quality and Preference*, 14(5-6):397–403, 2003.
- L. Lebart, A. Morineau, y M. P. Piron. *Statistique exploratoire multidimensionnelle*. Dunod, Paris, Francia, 1995.

- H. L'Hermier des Plantes. *Structuration Des Tableaux A Trois Indices De La Statistique: theorie et application d'une méthode d'analyse conjointe*. Proyecto Fin de Carrera, Université Des Sciences et Techniques Du Languedoc, Montpellier, 1976.
- M. Linting, J. J. Meulman, P. J. F. Groenen, y A. J. Van der Kooij. Stability of nonlinear principal components analysis. An empirical study using the balanced bootstrap. *Psychological Methods*, 12:359–379, 2007.
- Q. Liu. Variation partitioning by partial redundancy analysis (rda). *Environmetrics*, 8(2):75–85, 1997.
- R. Lombardo y E. J. Beh. The CAvariants R package version 1.0: correspondence analysis variants. <http://CRAN.R-project.org/package=CAvariants>, 2014.
- R. Lombardo, E. J. Beh, y L. D'Ambra. Non-symmetric correspondence analysis with ordinal variables using orthogonal polynomials. *Computational Statistics and Data Analysis*, 52:566–577, 2007.
- R. Lombardo, A. Carlier, y L. D'Ambra. Nonsymmetric correspondence analysis for three-way contingency tables. *Methodologica*, 4:59–80, 1996.
- R. Lombardo y T. Ringrose. Bootstrap confidence regions in non-symmetrical correspondence analysis. *Electronic Journal of Applied Statistical Analysis*, 5:413–417, 2012.
- R. Lombardo, T. Ringrose, y E. J. Beh. Bootstrap confidence regions in classical and ordered multiple correspondence analysis. En D. Vicari, A. Okada, y G. Ragozino, eds., *Book of Short Paper Analysis and Modeling of Complex*



- Data in Behavioural and Social Sciences-Cladag2012 Capri*, págs. 53–55. Cleup Padova, 2012.
- E. Macedo y A. Freitas. The alternating least-squares algorithm for CDPCA. *Communications in Computer and Information Science (CCIS)*, Springer Verlag (to appear), 2015.
- M. Maechler, P. Rousseeuw, A. Struyf, M. Hubert, y K. Hornik. *cluster: Cluster Analysis Basics and Extensions*, 2015. R package version 2.0.1 — For new features, see the 'Changelog' file (in the package source).
- M. W. Mahoney y P. Drineas. CUR matrix decompositions for improved data analysis. *PNAS*, 106:697–702, 2009.
- A. Markos. The GUI ca R package version 0.1-4: a Tcl/Tk GUI for the functions. [cran.r-project.org/package=caGUI](http://cran.r-project.org/package=caGUI), 2012.
- M. Markus y R. A. Visser. Bootstrapping and related techniques. En *Lecture Notes in Economics and Mathematical Systems*, tomo 376, págs. 71–75. Springer Berlin Heidelberg, 1992.
- A. Marreiros, G. Castela, E. Rebelo, y M. P. Galindo. The pathological-numeric codification of public hospitals in Portugal: implementation of mechanisms to support the assessment process of hospital clinical records and their relationship with funding. *CIEO-Research Centre for Spatial and Organizational Dynamics*, University of Algarve, 2010.
- J. Martín-Rodríguez, M. P. Galindo, y J. L. Vicente-Villardón. Comparison and integration of subspaces from a biplot perspective. *Journal of Statistical Planning and Inference*, 102:411–423, 2002.

- C. Martínez-Ruiz, B. Fernández-Santos, Putwain P. D., y M. J. Fernández-Gómez. Natural and man-induced revegetation on mining wastes: Changes in the floristic composition during early succession. *Ecological Engineering*, 30:286–294, 2007.
- G. McCabe. Principal variables. *Technometrics*, 26:137–144, 1984.
- S. Mendes. *Métodos multivariantes para evaluar patrones de estabilidad y cambio desde una perspectiva BIPLLOT*. Phd dissertation, Universidad de Salamanca, Spain, 2011.
- S. Mendes, M. J. Fernández-Gómez, M. P. Galindo, F. Morgado, P. Maranhão, U. Azeiteiro, y P. Bacelar-Nicolau. The study of bacterioplankton dynamics in the Berlengas Archipelago (West coast of Portugal) by applying the HJ-Biplot method. *Arquipelago Life and Marine Sciences*, 26:25–35, 2009.
- C. Meng, B. Kuster, A. C. Culhane, y A. M. Gholami. A multivariate approach to the integration of multi-omics datasets. *BMC bioinformatics*, 15(162), 2014.
- J. J. Meulman. *Homogeneity analysis of incomplete data*. DSWO Press, Netherland, 1982.
- L. Milan y J. Whittaker. Application of the parametric bootstrap to models that incorporate a singular value decomposition. *Applied Statistics*, 44:31–49, 1995.
- F. Morillo, A. A. Díaz-Faes, B. González-Albo, y L. Moreno. Do networking centres perform better? An exploratory analysis in Psychiatry and Gastroenterology/Hepatology in Spain. *Scientometrics*, 98(2):1401–1416, 2014.
- O. Nenadic y M. J. Greenacre. Correspondence analysis in R, with two- and

- three-dimensional graphics: the ca package. *Journal of Statistical Software*, 20(3):1–13, 2007.
- A. B. Nieto, N. Baccalá, P. Vicente-Galindo, y M. P. Galindo. The multibiplotGUI R package version 0.0-1: multibiplot analysis. [cran.r-project.org/package=multibiplotGUI](http://cran.r-project.org/package=multibiplotGUI), 2012.
- A. B. Nieto, M. P. Galindo, V. Leiva, y P. Vicente-Galindo. A methodology for biplots based on bootstrapping with R. *Revista Colombiana de Estadística*, 37(2):367–397, 2014.
- S. Nishisato. Forced classification: a simple application of a quantitative technique. *Psychometrika*, 49:25–36, 1984.
- G. F. Ochoa, E. Martínez, R. Ramírez, y G. Correa. Growth and Development of Lime (*Citrus latifolia* Tanaka), cv. Tahiti, in Soils with Limitations by Effective Depth in a Tropical Dry Forest. *Revista Facultad Nacional de Agronomía de Medellín*, 65(2):6567–6578, 2012.
- J. Oksanen, F. G. Blanchet, R. Kindt, P. Legendre, P. R. Minchin, B. O’Hara, G. L. Simpson, P. Solymos, M. H. H. Stevens, y H. Wagner. The vegan R package version 2.0-8: community ecology. [cran.r-project.org/package=vegan](http://cran.r-project.org/package=vegan), 2013.
- A. Orfao, M. González, J. San Miguel, M. C. Cañizo, M. P. Galindo, M. D. Caballero, R. Jiménez, y A. López-Borrasca. Clinical and Immunological Findings in Large B-Cell Chronic Lymphocytic Leukemia . *Clinical Immunology and Immunopathology*, 46:177–185, 1988a.
- A. Orfao, M. González, J. San Miguel, A. Ríos, M. D. Caballero, M. Sanz, M. J. Calmuntia, M. P. Galindo, y A. López-Borrasca. Bone marrow

- histopathologic patterns and immunologic phenotype in B-cell chronic lymphocytic leukaemia. *Blut*, 57:19–23, 1988b.
- J. Pagès. Eléments de comparaison entre l'analyse factorielle multiple et la méthode STATIS. *Revue de Statistique Appliquée*, 44(4):81–95, 1996.
- J. Pagès y F. Husson. Multiple factor analysis with confidence ellipses: a methodology to study the relationships between sensory and instrumental data. *Journal of Chemometrics*, 19:138–144, 2005.
- M. C. Patino, M. P. Vicente, y M. P. Galindo. Multivariate profile of the domestic workers. *Cuadernos de Relaciones Laborales*, 29(2):13–44, 2011.
- M. E. Pinto. *Suicidio juvenil: Sociología de una realidad social*. Phd dissertation, Universidad Complutense de Madrid, Spain, 2006.
- M. Quenouille. Approximation tests of correlation in time series. *Journal of the Royal Statistical Society*, 11:68–84, 1949.
- R-Team. *R: a language and environment for statistical computing*. R foundation for statistical computing. Vienna, Austria, 2014. URL [www.R-project.org](http://www.R-project.org).
- G. Ramírez, M. Vásquez, A. Camardiel, B. Pérez, y M. P. Galindo. Graphical detection for the multicollinearity by the h-plot of the inverse matrix of correlations. *Revista Colombiana de Estadística*, 28:207–219, 2005.
- R. Rao. The use and interpretation of principal component analysis in applied research. *Sankhya: The Indian Journal of Statistics: Series A*, 26:329–358, 1964.
- T. Raykov y T. D. Little. A note on procrustean rotation in exploratory factor analysis: a computer intensive approach to goodness-of-fit evaluation. *Educational and Psychological Measurement*, 59:47–57, 1999.

- J. Reiczigel. Bootstrap tests in correspondence analysis. *Applied Stochastic Models and Data Analysis*, 12(2):107–117, 1996.
- T. J. Ringrose. Bootstrap confidence regions for correspondence analysis. *Journal of Statistical Computation and Simulation*, 82(10):1397–1413, 2012.
- T. J. Ringrose. The cabootcrs R package version 1.0: bootstrap confidence regions for correspondence analysis. <http://CRAN.R-project.org/package=cabootcrs>, 2013.
- J. C. Rivas-Gonzalo, Y. Gutiérrez, A. M. Polanco, E. Hebrero, J. L. Vicente-Villardón, M. P. Galindo, y C. Santos-Buelga. Biplot analysis applied to enological parameters in the geographical classification of young red wines. *American Journal of Enology and Viticulture*, 44:302–308, 1993.
- R. Rocci, S. A. Gattone, y M. Vichi. A new dimension reduction method: Factor discriminant k-means. *Journal of Classification*, 28:210–226, 2011.
- R. Sabatier y M. Vivien. A new linear method for analyzing four-way multiblocks tables: STATIS-4. *J. Chem*, 22:299–407, 2008.
- C. Santos, S. S. Muñoz, Y. Gutiérrez, E. Hebrero, J. L. Vicente, M. P. Galindo, y J. C. Rivas. Characterization of young red wines by application of HJ Biplot analysis to anthocyanin profiles. *Journal of Agricultural and Food Chemistry*, 39:1086–1090, 1991.
- L. Sauzay, M. Hanafi, E. M. Qannari, y P. Schlich. Analyse de  $K + 1$  tableaux a l'aide de la methode STATIS: application en evaluation sensorielle. En *9-ieme Journees Europeennes Agro-industrie et Methodes Statistiques*. 2006.
- R. Sepúlveda, J. L. Vicente-Villardón, y M. P. Galindo. The biplot as a diagnostic

- tool of local dependence in latent class models: a medical application. *Statistics in Medicine*, 27:1855–1869, 2008.
- A. Serafim, R. Company, B. Lopes, J. Rosa, A. Cavaco, G. Castela, E. Castela, N. Olea, y M. J. Bebianno. Assessment of essential and nonessential metals and different metal exposure biomarkers in the human placenta in a population from the south of Portugal. *Journal of Toxicology and Environmental Health*, 42:867–877, 2012.
- R. Siciliano, A. B. Mooijart, y P. G. M. Van der Heijden. A probabilistic model for nonsymmetric correspondence analysis and prediction in contingency tables. *Journal of the Italian Statistical Society*, 2(1):85–106, 1993.
- M. Simier, L. Blanc, F. Pellegrin, y D. Nandris. Approche simultanée de K couples de tableaux: application à l'étude des relations pathologie végétale-environnement. *Revue de Statistique Appliquée*, 47:31–46, 1999.
- B. Simonetti y M. Gallo. Alternative interpretations to the non symmetrical correspondence analysis. *Caribbean Journal of Mathematical and Computing Sciences*, 12:18–22, 2002.
- G. L. Simpson. The cocorresp R package version 0.2-1: co-correspondence analysis ordination methods. <http://CRAN.R-project.org/package=cocorresp>, 2009.
- D. F. Stauffer, E. O. Garton, y R. K. Steinhorst. A comparison of principal component from real and random data. *Ecology*, 66:1693–1698, 1985.
- G. W. Stewart. Four algorithms for the efficient computation of truncated qr approximations to a sparse matrix. *Numerische Mathematik*, 83:313–323, 1999.

- S. Takane, Y. and Jung. Regularized partial and/or constrained redundancy analysis. *Psychometrika*, 73(4):671–690, 2008.
- Y. Takane y H. Hwang. Regularized multiple correspondence analysis. En M. J. Greenacre y J. Blasius, eds., *Multiple Correspondence Analysis and Related Methods*, págs. 259–279. London: Chapman & Hall, 2006.
- Y. Takane y H. Hwang. Regularized linear and kernel redundancy analysis. *Computational Statistics and Data Analysis*, 52:394–405, 2007.
- Y. Takane y S. Jung. Regularized non symmetrical correspondence analysis. *Computational Statistics and Data Analysis*, 53(8):3159–3170, 2009.
- Y. Takane, H. Yanai, y S. Mayekawa. Relationships among several methods of linearly constrained correspondence analysis. *Psychometrika*, 56:667–684, 1991.
- K. Tateneni y M. W. Browne. A non-iterative method of joint correspondence analysis. *Psychometrika*, 65:157–165, 2000.
- A. P. Ter Braak, C. J. F. and Schaffers. Co-correspondence analysis: a new ordination method to relate two community compositions. *Ecology*, 85(3):834–846, 2004.
- C. J. F. Ter Braak. Canoco a fortran program for canonical community ordination by [partial] [detrended] [canonical] correspondence analysis, principal component analysis and redundancy analysis (version 2.1). technical report lwa-88-02. 1985a.
- C. J. F. Ter Braak. Canoco a fortran program for canonical correspondence analysis and detrended correspondence analysis. 1985b.

- C. J. F. Ter Braak. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology*, 5:1167–1179, 1986.
- C. J. F. Ter Braak. Canoco an extension of decorana to analyze species-environment relationships. *Vegetatio*, 75(3):159–160, 1988a.
- C. J. F. Ter Braak. Partial canonical correspondence analysis. En Bock H. H., ed., *Classification and Related Methods of Data Analysis*, págs. 551–558. North Holland Press, 1988b.
- C. J. F. Ter Braak. Interpreting canonical correlation analysis through biplot of structure and weights. *Psychometrika*, 55:519–531, 1990.
- C. J. F. Ter Braak y C. W. Looman. Biplots in reduced-rank regression. *Biometrical Journal*, 36:983–1003, 1994.
- C. J. F. Ter Braak y P. F. M. Verdonschot. Canonical correspondence analysis and related multivariate methods in aquatic ecology. *Aquatic Sciences*, 57:255–289, 1995.
- J. Thioulouse. Simultaneous analysis of a sequence of paired ecological tables: a comparison of several methods. *The Annals of Applied Statistics*, 5:2300–2325, 2011.
- J. Thioulouse, D. Chessel, S. Dolédec, y J. M. Olivier. Ade-4: a multivariate analysis and graphical display software. *Statistics and Computing*, 7:75–83, 1997.
- J. Thioulouse y S. Dray. Interactive multivariate data analysis in R with the ade4 and ade4TkGUI packages. *Journal of Statistical Software*, 22(5):1–14, 2007.



- J. Thioulouse y S. Dray. The ade4TkGUI R package version 0.2-6: ade4 Tcl/Tk graphical user interface. [cran.r-project.org/package=ade4TkGUI](http://cran.r-project.org/package=ade4TkGUI), 2012.
- J. Thioulouse, M. Simier, y D. Chessel. Simultaneous analysis of a sequence of paired ecological tables with the statico method. *Ecology*, 85:272–283, 2004.
- R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1996.
- L. Tierney. The tkrplot R package version 0.0-23: TK Rplot. [cran.r-project.org/package=tkrplot](http://cran.r-project.org/package=tkrplot), 2012.
- M. E. Timmerman, H. A. L. Kiers, A. K. Smilde, E. Ceulemans, y J. Stouten. Bootstrap confidence intervals in multi-level simultaneous component analysis. *British Journal of Mathematical and Statistical Psychology*, 62:299–318, 2009.
- N. T. Trendafilov. From simple structure to sparse components: a review. *Computational Statistics*, 29(3-4):431–454, 2014.
- D. Tu y L. Zhang. Jackknife approximations for some nonparametric confidence intervals of functional parameters based on normalizing transformations. *Computational Statistics*, 7:3–15, 1992.
- L. R. Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31:279–311, 1966.
- J. Tukey. Bias and confidence in not quite large samples. *Annals of Mathematical Statistics*, 29:614, 1958.
- V. Vairinhos. *Development of a system for data mining based on biplot methods*. Phd dissertation, Universidad de Salamanca, Spain, 2003.

- A. Vallejo-Arboleda, J. L. Vicente-Villardón, y M. P. Galindo. Canonical STATIS: biplot analysis of multi-table group structured data based on STATIS-ACT methodology. *Computational Statistics and Data Analysis*, 51:4193–4205, 2006.
- A. Vallejo-Arboleda, J. L. Vicente-Villardón, M. P. Galindo, M. Fernández, C. Fernández, y E. Bécares. Analysis of time evolution for group structured data: canonical dual STATIS and doubly multivariate repeated measures model. *Revista Colombiana de Estadística*, 31:321–340, 2008.
- S. Van Buuren y J. De Leeuw. Equality constraints in multiple correspondence analysis. *Multivariate Behavioral Research*, 27:567–583, 1992.
- A. L. Van den Wollenberg. Redundancy analysis an alternative for canonical correlation analysis. *Psychometrika*, 42(2):207–219, 1977.
- I. Van Mechelen, H. H. Bock, y P. De Boeck. Two-mode clustering methods: a structured overview. *Statistical Methods in Medical Research*, 13(5):363–394, 2004.
- J. P. Vázquez, M. P. Vicente, y M. P. Galindo. Variables que inciden en la seguridad de las escuelas públicas de los Estados Unidos. *Revista Pedagogía*, 44(1):141–165, 2011.
- B. R. Vázquez-de Aldana, A. García-Ciudad, B. García-Criado, S. Vicente-Tavera, y I. Zabalgogeoazcoa. Fungal endophyte (epichloë festucae) alters the nutrient content of festuca rubra regardless of water availability. *Plos One*, 8(12), 2013.
- P. Vicente, E. Vaz, y T. Noronha. How corporations deal with reporting

- sustainability: Assessment using the multicriteria logistic biplot approach. *Systems*, 3:6–26, 2015.
- M. P. Vicente-Galindo. *Análisis de tablas de tres vías: recientes desarrollos del STATIS*. Tesis Doctoral, Universidad de Salamanca, Spain, 2013.
- J. L. Vicente-Villardón. MULTBILOT: a package for multivariate analysis using biplots. [biplot.usal.es/ClassicalBiplot/index.html](http://biplot.usal.es/ClassicalBiplot/index.html), 2003.
- J. L. Vicente-Villardón, M. P. Galindo, y A. Blázquez. Logistic biplots. En M. Grenacre y J. Blasius, eds., *Multiple Correspondence Analysis and Related Methods*, págs. 491–509. Chapman & Hall, New York, US, 2006.
- M. Vichi. Double k-means clustering for simultaneous classification of objects and variables. En S. Borra, R. Rocci, M. Vichi, y M. Schader, eds., *Advances in Classification and Data Analysis*. Springer, Berlin, 2000.
- M. Vichi y H. A. L. Kiers. Factorial k-means analysis for two-way data. *Computational Statistics and Data Analysis*, 37:49–64, 2001.
- M. Vichi y G. Saporta. Clustering and disjoint principal component analysis. *Computational Statistics and Data Analysis*, 53:3194–3208, 2009.
- E. Vigneau y E. M. Qannari. Clustering of variables around latent component application to sensory analysis. *Communications in Statistics, Simulation and Computation*, 32(4):1131–1150, 2004.
- J. D. Vilorio, J. H. Gil, D. L. Durango, y C. M. García. Physicochemical characterization of propolis from the region of Bajo Cauca Antioqueño (Antioquia, Colombia). *Bioteología en el Sector Agropecuario y Agroindustrial*, 10:77–86, 2012.

- S. Vines. Simple principal components. *Applied Statistics*, 49:441–451, 2000.
- R. H. Whittaker. Gradient analysis of vegetation. *Biological Reviews*, 42(2):207–264, 1967.
- P. Willems y M. P. Galindo. Canonical non-symmetrical correspondence analysis: an alternative in constrained ordination. *SORT*, 32(1):93–112, 2008.
- D. M. Witten y R. Tibshirani. A framework for feature selection in clustering. *Journal of the American Statistical Association*, 105(490):713–726, 2010.
- W. Yan, L. A. Hunt, Q. Sheng, y Z. Szlavnic. Cultivar evaluation and mega-environment investigation based on GGE biplot. *Crop Science*, 40:597–605, 2000.
- W. Yan y M. S. Kang. *GGE biplot analysis: a graphical tool for breeders, geneticists, and agronomists*. CRC Press, Boca Raton, US, 2003.
- H. Yanai. Some generalizations of correspondence analysis in terms of projection operators. En E. Diday, Y. Escoufier, L. Lebart, J. P. Pagès, Y. Schektman, y R. Thomassone, eds., *Data Analysis and Informatics IV*, págs. 193–207. North Holland, 1986.
- H. Yanai. Partial multiple correspondence analysis and its properties. En C. Hayashi, M. Jambu, E. Diday, y N. Ohsumi, eds., *Recent Development in Clustering and Data Analysis*, págs. 259–266. Academic Press, 1987.
- H. Yanai. Generalized canonical correlation analysis with linear constraints. En C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H. H. Bock, y Y. Baba, eds., *Data Science, Classification, and Related Methods*, págs. 539–546. Springer-Verlag, 1998.

- H. Yanai y T. Maeda. Partial multiple correspondence analysis. En S. Nishisato, Y. Baba, H. Bozdogan, y K. Kanefuji, eds., *Measurement and Multivariate Analysis*, págs. 57–68. Springer Japan, 2002.
- A. S. Young, G. and Householder. Discussion of a set of points in terms of their mutual distances. *Psychometrika*, 3:19–22, 1938.
- A. Zárraga y B. Goitisoló. Simultaneous analysis in S-PLUS: the SimultAn package. *Journal of Statistical Software*, 70(11):1–22, 2011.
- A. Zárraga y B. Goitisoló. The SimultAnR package version 1.1: correspondence and simultaneous analysis. <http://CRAN.R-project.org/package=SimultAnR>, 2013.
- Z. Zhang, H. Zha, y H Simon. Low rank approximations with sparse factors i: basic algorithms and error analysis. *SIAM journal on matrix analysis and its applications*, 23:706–727, 2002.
- Z. Zhang, H. Zha, y H Simon. Low rank approximations with sparse factors ii: penalized methods with discrete newton like iterations. *SIAM journal on matrix analysis and its applications*, 25:901–920, 2004.
- H. Zou y T. Hastie. Regression shrinkage and selection via the elastic net, with applications to microarrays. *Technical report, Department of Statistics, Stanford University*, 2003. URL <http://www-stat.stanford.edu/~hastie/pub.htm>.
- H. Zou, T. Hastie, y R. Tibshirani. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):262–286, 2006.



---

## Apéndice A

