

UNIVERSIDAD DE SALAMANCA

Departamento de Estadística

Doctorado en Estadística Multivariante Aplicada

Tesis Doctoral



**CONTRIBUCIONES AL ANÁLISIS BILOT BASADAS
EN SOLUCIONES FACTORIALES DISJUNTAS
Y EN SOLUCIONES SPARSE**

Autora: Mitzi Isabel Cubilla Montilla

Directoras: María Purificación Galindo Villardón
Ana Belén Nieto Librero

2019



Departamento de Estadística
Universidad de Salamanca

Dra. María Purificación Galindo Villardón
Catedrática de Estadística del Departamento de Estadística
de la Universidad de Salamanca.

y

Dra. Ana Belén Nieto Librero
Profesora Ayudante Doctor de Estadística del Departamento de Estadística
de la Universidad de Salamanca

CERTIFICAN:

Que Doña. **Mitzi Isabel Cubilla Montilla** ha realizado en la Universidad de Salamanca, bajo su dirección, el trabajo para optar al título de Doctor en Estadística Multivariante Aplicada, que presenta con el título “*Contribuciones al Análisis Biplot basadas en Soluciones Factoriales Disjuntas y en Soluciones Sparse*”, autorizando expresamente su lectura y defensa.

Y para que conste, firman el presente certificado en Salamanca a ____ de julio de 2019.

M^a Purificación Galindo Villardón

Ana Belén Nieto Librero

**Contribuciones al Análisis Biplot basadas en
Soluciones Factoriales Disjuntas y
en Soluciones Sparse**



Depto. de Estadística
Universidad de Salamanca

Memoria para optar al título de Doctor en
Estadística Multivariante Aplicada
por la Universidad de Salamanca

Presenta:

Mitzi Isabel Cubilla Montilla

Salamanca, 2019

Agradecimientos

En *septiembre del año 2016*, emprendí un viaje con las maletas *cargadas de ilusiones* y con la confianza de empezar un nuevo capítulo en mi vida. Después de un intenso período de tres años, culminó mi tesis doctoral y dedico este apartado de agradecimientos sinceros:

A DIOS: por escuchar mis oraciones, por ayudarme a reponer cada día, por darme fuerza y voluntad para enfrentar este reto y mantenerme firme en los momentos de debilidad.

A PURI: mi directora de tesis, por aventurarse conmigo en la dirección de este proyecto que hoy es una realidad. A ella le debo la disciplina, el compromiso, la excelencia y la pasión por la Estadística. Ha sido una experiencia realmente gratificante y enriquecedora en todos los sentidos.

A ANA BELÉN: codirectora de tesis, una extraordinaria persona que supo guiar mis ideas y estuvo siempre presente; además, de corazón muy noble y con un alma guerrera. Ana, ¡Eres una campeona!

A ISABEL: su apoyo incondicional merece reconocer su alto sentido de profesionalismo y su gran capacidad en la orientación de los artículos; ha sido un verdadero privilegio aprender a su lado.

PURI, ANA, ISABEL sinceramente no he podido tener mejor directora, mejor maestra, mejor guía.

A MIS COMPAÑER@S DOCTORANDO@S Y DOCTORADO@S: junto a ustedes consolidé el verdadero sentido de la amistad, no queda duda de que el valor de una persona es incalculable. Ahora estoy absolutamente convencida de que puedo presumir de amig@s más allá de las fronteras. También quiero agradecer a mis compañer@s panameñ@s: **Estelina, Carmen, Nathalia y Gonzalo** por estar a mi lado en mis altas y bajas. A mis colegas **Adela, Elena y Parker** gracias por su apoyo.

Una tesis es también fruto del valioso apoyo y el coraje que nos ofrece la familia para seguir adelante. **MAMI** este es un logro suyo por educarme en principios y valores. **PAPÁ**, gracias por ser un excelente padre, sé que sufriste en silencio, pero fuiste capaz de sobrellevarlo con serenidad.

A mis **HERMANAS** que, entre otras muchas cosas, con sus “estados de WhatsApp” se sentían orgullosas de mí y me transmitían esa valentía y determinación que necesitaba. Ojalá algún día yo pueda retribuir todo el apoyo brindado. A mis **SOBRINAS**, gracias por hacer de cada cumpleaños, de cada Navidad, de cada día de la madre o del padre...un día especial con sus inventos y locuras.

Viajar con mi familia fue mi mayor reto. El camino no fue fácil, pero valió la pena enfrentar el reto. Por eso, quiero agradecer a mi confidente, mi consejero, mi compañero de vida. **ESPOSO**, *gracias* por valorar mi esfuerzo, por sobrellevar mis estudios con infinita paciencia, con voluntad, con inteligencia; gracias por hacerme compañía cada día o cada larga y agotadora noche de estudio. Mi máximo agradecimiento a mis dos **HIJOS**, quienes durante estos años han tenido que amoldarse a mi vida como doctoranda y entender de primera mano el sentido de la perseverancia. Queridos míos, ustedes tres han sido mis mejores compañeros de viaje; por ello, más que agradecimientos, les quiero presentar mis disculpas por el desajuste que estos años han provocado en nuestras vidas.

En fin...gracias de corazón a todas las personas que confiaron en mí, y también a las que no, ya que de una u otra manera, iluminaron como estrellas mi camino y me dieron fuerzas para remar en una dirección. Espero seguir contando con todos para siempre. ¡**GRACIAS!**

ÍNDICE GENERAL

ÍNDICE GENERAL	i
ÍNDICE DE FIGURAS	iv
ÍNDICE DE TABLAS	vi
INTRODUCCIÓN.....	i
Capítulo 1 LOS MÉTODOS BIPLLOT: DEFINICIÓN Y CONCEPTOS GENERALES	1
1.1 Introducción.....	2
1.2 Conceptos Generales	3
1.2.1 La Matriz de Datos	3
1.2.2 Descomposición en Valores Singulares	4
1.2.3 Formulación Teórica.....	6
1.2.4 Porcentaje de varianza explicada	7
1.2.5 Definición y Representación Gráfica	8
1.2.6 Representación Geométrica del Biplot.....	9
1.2.7 Interpretación Geométrica.....	11
1.3 Contribuciones al Biplot	14
Capítulo 2 BIPLLOT PARA TABLAS DE DOS VÍAS	18
2.1 Introducción	19
2.2 Biplot Clásicos: GH Biplot y JK Biplot	19
2.2.1 GH Biplot	20
2.2.2 JK Biplot.....	22
2.3 HJ-Biplot.....	23
2.4 Biplot Robusto.....	25
2.5 MANOVA-Biplot	28
2.5.1 MANOVA Biplot (una vía).....	31
2.5.2 MANOVA Biplot (de dos vías).....	32
2.5.3 Representación gráfica del MANOVA Biplot	35
2.6 Modelos AMMI GGE Biplot	37
2.6.1 Biplot AMMI	38
2.6.2 GGE Biplot	41
2.7 Biplot para Datos Composicionales.....	45

2.7.1 Principios del análisis	45
2.7.2 Representación biplot en datos composicionales:	47
2.7.3 Propiedades Fundamentales	49
2.8 HJ Biplot Composicional	52
2.9 Biplot Logístico Binario	53
2.9.1 Estructura de datos para realizar un Biplot Logístico	53
2.9.2 Algoritmo General	55
2.9.3 Reglas de interpretación	56
2.9.4 Calidad de Representación de las variables.....	56
2.9.5 Regiones de Predicción	57
2.9.6 Biplot Logístico Externo.....	57
2.10 Biplot Logístico Nominal.....	59
2.11 Biplot Logístico Ordinal	62
2.12 Versión Inferencial del Biplot	64
2.12.1 El Principio “Plug-In”	65
2.12.2 La Muestra Bootstrap.....	66
2.12.3 Estimador Jackknife.....	67
2.12.4 Intervalos de confianza Bootstrap.....	69
Capítulo 3 SOFTWARE SOBRE BIPLLOT	19
3.1 Introducción	72
3.2 Librerías en R para la construcción del Biplot	73
3.2.1 Paquete Stats	74
3.2.2 Paquete “calibrate”	75
3.2.3 Paquete “bpca”	77
3.2.4 Paquete “MultBiplotR”	79
3.2.5 Paquete “NominalLogisticBiplot (NBL)”	80
3.2.6 Paquete “OrdinalLogisticBiplot (OLB)”	82
3.2.7 Paquete “BiplotForCategoricalVariables”	84
3.3 Comparaciones gráficas de las librerías	84
3.4 Interfaz Gráfica de Usuario (GUI)	88
3.4.1 Paquete “BiplotGUI”	89
3.4.2 Paquete “MultiBiplotGUI”	90

3.4.3 Paquete “GGEBiplotGUI”	92
3.4.4 Paquete “dynBiplotGUI”	93
3.4.5 Paquete “biplotbootGUI”	95
3.5 Otros Paquetes.....	98
3.5.1 Paquete “Vegan”	99
3.5.2 Paquete “ade4”	100
3.5.3 Paquete “Ade4TkGUI”	101
3.5.4 Paquete “ca”	101
3.5.5 Paquete “caGUI”	103
Capítulo 4 SOLUCIONES DISJUNTAS Y SPARSE BIPLLOT	104
4.1 Sparse PCA y Soluciones Disjuntas	105
4.2 Sparse Biplot	115
4.3 <i>Ridge</i> HJ Biplot	117
4.4. <i>LASSO</i> HJ Biplot	121
4.5. <i>Elastic Net</i> HJ Biplot	127
4.6 Varianza Total Explicada por las componentes sparse	131
4.7 Paquete en R para Sparse Biplot: “SparseBiplots”	133
4.8 Aplicaciones Prácticas	141
CONCLUSIONES.....	156
ARTÍCULO PUBLICADO	158
ARTÍCULOS SOMETIDOS.....	176
PARTICIPACIÓN EN CONGRESOS NACIONALES E INTERNACIONALES (2015-2019)	177
BIBLIOGRAFÍA.....	178

ÍNDICE DE FIGURAS

<i>Figura 1-1 Esquema de la tabla de datos</i>	4
<i>Figura 1-2 Representación de la matriz de datos</i>	4
<i>Figura 1-3 Transformación de la matriz original en componentes principales</i>	5
<i>Figura 1-4 Esquema de una representación biplot de una matriz (12x5)</i>	8
<i>Figura 1-5 Pasos para la construcción del biplot</i>	10
<i>Figura 1-6 Interpretación geométrica del producto escalar</i>	11
<i>Figura 1-7 Proyección ortogonal de los marcadores fila sobre los marcadores columna</i>	12
<i>Figura 1-8 Reglas básicas de interpretación (a)</i>	13
<i>Figura 1-9 Reglas básicas de interpretación (b)</i>	14
<i>Figura 2-1 Marcadores para filas y columnas en el HJ-Biplot</i>	24
<i>Figura 2-2 Modelo ANOVA de una matriz original $X_{n \times p}$</i>	29
<i>Figura 2-3 Representación de un biplot canónico</i>	36
<i>Figura 2-4 Relación entre ambientes</i>	42
<i>Figura 2-5 Ambiente ideal de prueba</i>	43
<i>Figura 2-6 Rendimiento de un genotipo en los distintos ambientes</i>	44
<i>Figura 2-7 Biplot de forma</i>	50
<i>Figura 2-8 Biplot de covarianza</i>	50
<i>Figura 2-9 Algoritmo general, para la aplicación del Biplot Logístico</i>	55
<i>Figura 2-10 Esquema de representación del Biplot Logístico</i>	55
<i>Figura 2-11 Regiones de predicción del Biplot Logístico</i>	57
<i>Figura 2-12 Algoritmo para la aplicación del Biplot Logístico Externo</i>	59
<i>Figura 2-13 Teselación definida por las regiones de predicción</i>	61
<i>Figura 2-14 Regiones de predicción para una variable ordinal</i>	63
<i>Figura 2-15 Metodología Jackknife</i>	68
<i>Figura 2-16 Intervalos de Confianza Bootstrap para muestras grandes</i>	70
<i>Figura 2-17 Intervalos de confianza Bootstrap para muestras pequeñas</i>	70
<i>Figura 3-1 Comparación del HJ Biplot obtenida de diferentes librerías en R</i>	87
<i>Figura 3-2 Extensiones del Biplot Logístico: Biplot Ordinal y Biplot Nominal</i>	88
<i>Figura 3-3 Biplot obtenido de la Interfaz Gráfica de Usuario "BiplotGUI"</i>	90
<i>Figura 3-4 Ventana principal del paquete "multibiplotGUI"</i>	91
<i>Figura 3-5 Biplot obtenido de la Interfaz Gráfica de Usuario "MultiBiplotGUI"</i>	91
<i>Figura 3-6 Ventana principal del paquete GGEBiplotGUI</i>	92
<i>Figura 3-7 Biplot obtenido de la Interfaz Gráfica de Usuario "GGEBiplot"</i>	93
<i>Figura 3-8 Ventana principal del paquete "dynBiplotGUI"</i>	94

<i>Figura 3-9 Biplot obtenido de la Interfaz Gráfica de Usuario "dynBiplotGUI</i>	95
<i>Figura 3-10 Ventana principal del paquete "biplotbootGUI"</i>	96
<i>Figura 3-11 Biplot obtenido de la Interfaz Gráfica de Usuario "biplotbootGUI"</i>	96
<i>Figura 4-1 CUR: Selección de columnas (C)</i>	109
<i>Figura 4-2 CUR: Selección de Filas (R)</i>	109
<i>Figura 4-3 CUR + HJ Biplot. Datos sobre el Índice para una vida mejor</i>	113
<i>Figura 4-4 Representación del Disjoint Biplot: izquierda (plano 1-2) y derecha (plano 1-3)</i>	114
<i>Figura 4-5 Ilustración de la regla del operador soft-thresholding</i>	123
<i>Figura 4-6 Ilustración de la regla del operador hard-thresholding</i>	124
<i>Figura 4-7. Logaritmo Esquema del Sparse Biplot</i>	131
<i>Figura 4-8 HJ Biplot. Datos sobre el índice para una vida mejor</i>	143
<i>Figura 4-9 Ridge HJ Biplot. Datos sobre el Índice para una vida mejor</i>	143
<i>Figura 4-10 LASSO HJ Biplot. Datos sobre el Índice para una vida mejor</i>	145
<i>Figura 4-11 Elastic Net HJ Biplot. Datos sobre el Índice para una vida mejor</i>	145
<i>Figura 4-12 HJ Biplot. Datos sobre indicadores sociales y dimensiones culturales</i>	150
<i>Figura 4-13 Ridge HJ Biplot. Datos sobre indicadores sociales y dimensiones culturales</i>	150
<i>Figura 4-14 LASSO HJ Biplot con el operador soft Thresholding (izquierda) y con hard thresholding (derecha). Datos sobre indicadores sociales y dimensiones culturales</i>	151
<i>Figura 4-15 Elastic Net HJ Biplot. Datos sobre indicadores sociales y dimensiones culturales</i>	152

ÍNDICE DE TABLAS

<i>Tabla 1: Contribuciones a la metodología Biplot</i>	<i>15</i>
<i>Tabla 2: Paquetes o librerías que realizan a descomposición Biplot.....</i>	<i>73</i>
<i>Tabla 3: Factores y variables relacionadas con el Índice para una vida mejor.....</i>	<i>86</i>
<i>Tabla 4: Interfaces gráficas que realizan descomposición Biplot</i>	<i>89</i>
<i>Tabla 5: Tabla comparativa de los resultados que devuelven las librerías y las GUI que realizan la descomposición Biplot</i>	<i>97</i>
<i>Tabla 6: Paquetes de R que utilizan el término Biplot, pero que no realizan la DVS</i>	<i>98</i>
<i>Tabla 7: Comparación teórica entre la restricción Ridge y Lasso en ACP.....</i>	<i>126</i>
<i>Tabla 8: Comparación de las cargas obtenidas mediante los diferentes métodos (datos del Índice para una vida mejor).....</i>	<i>146</i>
<i>Tabla 9. Comparación de las cargas obtenidas mediante los diferentes métodos (datos sobre indicadores sociales y cultura)</i>	<i>153</i>

INTRODUCCIÓN

En la actualidad, la variedad de información, el rápido crecimiento de los datos y la capacidad de almacenamiento ha originado en la mayoría de las disciplinas y campos de estudio, un incremento de la información. Estos datos son generados por modernas tecnologías que generan o distribuyen datos en tiempo real. Transacciones en línea, correos electrónicos, redes sociales, teléfonos móviles, imágenes, audios, videos, dispositivos GPS, registros de salud, aplicaciones web colaborativas, registros de centros de llamadas y datos generados por sensores, son solo algunos ejemplos de ello.

El flujo de grandes cantidades de datos es recogido en el término "*Big Data*", en referencia a *Datos Masivos* o *Macrodatos*, que se generan a gran velocidad y de forma continua. El término "*Big Data*" se usa principalmente para describir contenido digital masivo, heterogéneo y, a menudo, no estructurado, que es difícil de procesar utilizando herramientas y técnicas de gestión de datos tradicionales (Talía, 2013).

Transformar el *Big Data* en conocimiento va acompañado de complejidad, cuya esencia es captada principalmente por las "tres V's": volumen, velocidad y variedad (Gandomi & Haider, 2015; Hashem et al., 2015; Rodríguez-Mazahua et al., 2016). Las tres características han catalizado el desarrollo de estrategias técnicas y analíticas para hacer frente a los datos (Berman, 2013) y hacen de la visualización una tarea desafiante que se genera exponencialmente (Rodríguez-Mazahua et al., 2016), convirtiéndole en una parte central del conocimiento en muchos campos (Amaratunga & Cabrera, 2016).

Aunque la complejidad de los datos representa oportunidades para los investigadores, también plantea un desafío: el uso y la adopción de nuevas herramientas estadísticas y computacionales que sean capaces de potenciar la información. En tal sentido, el “*Machine Learning*” (Aprendizaje Automático), una subrama de la *Inteligencia Artificial*”, ha sido clave en los avances tecnológicos, diseñando y validando algoritmos en el ámbito del Big Data. Además, ha surgido un nuevo paradigma: “*Data Science*” (Ciencia de Datos), un campo interdisciplinario que se ocupa de la recolección, administración, preparación, análisis y visualización de grandes conjuntos de datos, combinando *Data Mining* (Minería de datos), *Machine Learning* y Estadística para generar perspectivas analíticas a partir del *Big Data*.

Los algoritmos de *Machine Learning* se suelen clasificar en dos áreas: aprendizaje supervisado y no supervisado. Dentro del análisis no supervisado, el enfoque estadístico basado en la reducción de la dimensionalidad ha ganado considerable atención debido al aumento en el volumen de los datos. En este sentido, el Análisis de Componentes Principales ([Pearson, 1901](#)) -una idea de hace más de 100 años- sigue siendo una herramienta esencial para el análisis de datos multivariantes y la reducción de la dimensión. Su objetivo es proyectar los datos originales sobre un subespacio de menor dimensión de tal forma que capture la mayor parte de la variación del conjunto de datos.

Bajo este concepto, los métodos Biplot ([Gabriel, 1971](#); [Galindo, 1986](#); [Gower & Hand, 1995](#)) derivados del Análisis de Componentes Principales (ACP); son, por tanto, métodos de reducción de la dimensionalidad.

Los métodos Biplot han cobrado importancia, durante más de cuatro décadas contribuyendo al avance de la ciencia de manera efectiva. En cualquier campo o área de investigación se ha puesto de relieve el aporte de estos métodos; con los cuales, datos provenientes de las Ciencias Naturales o las Ciencias Sociales, se ven favorecidas en la toma de decisiones a través de esta metodología.

Frente a este escenario, y con el fin de dar un paso importante que contribuya en el proceso de análisis multivariante de grandes y complejos conjuntos de datos; y al mismo tiempo, hacer un novedoso aporte a favor de los métodos Biplot, se ha elaborado este trabajo de investigación que hemos dividido en cuatro capítulos.

El primer capítulo se inicia presentando la definición del Biplot y el desarrollo de los fundamentos teóricos que le sustentan, se presentan luego las propiedades fundamentales de la representación y las reglas básicas de interpretación, y se concluye con una revisión bibliográfica sobre las principales contribuciones a la metodología Biplot.

En el segundo capítulo se presentan en detalle los diferentes métodos biplot, los elementos a considerar para su construcción, sus propiedades y las formas de representación gráfica. Se hace referencia tanto a los métodos biplot para datos cuantitativos, así como también para datos de tipo binario y/o categórico. Al final de este capítulo se presenta la versión inferencial del Biplot, que recoge una explicación sobre los estimadores y los intervalos de confianza.

Posteriormente, en el tercer capítulo se presenta en detalle cada una de las librerías en R y las Interfaces Gráficas de Usuario (GUI) que llevan a la representación del Biplot. En esta sección también se examina la capacidad de utilizar en un contexto específico las diferentes librerías e interfaces sobre Biplot, para lo cual se conformará una matriz de datos reales que permita explorar estas técnicas e ir comparando los resultados.

El cuarto capítulo, apartado central de esta tesis evidencia la necesidad de implementar nuevos modelos para el análisis de datos multivariantes, de cara al desarrollo del *Big Data*. Para adaptarse a datos masivos es necesario la implementación de nuevas técnicas, capaces de reducir la dimensionalidad de los datos y mejorar su interpretación. Como el Biplot basa su fundamento teórico en el ACP, el capítulo inicia con una revisión bibliográfica de los diferentes métodos de Análisis de Componentes Principales encaminados a simplificar la información original, que comprenden desde los métodos de rotación hasta las metodologías SPARSE y las componentes disjuntas. A partir de allí, nos centramos en la construcción de nuevas metodologías Biplot (SPARSE), poniendo especial énfasis en la construcción de componentes principales modificadas, mediante la contracción o anulación de las cargas. Se proponen diferentes metodologías Biplot: *Ridge* HJ Biplot, *LASSO* HJ Biplot y *Elastic Net* HJ Biplot; además de la CUR HJ-Biplot. En cada caso se propone el algoritmo en R para su uso. De esta manera, los métodos biplot enriquecen las técnicas estadísticas multivariantes utilizadas en análisis de datos masivos. Por último, se presentan las principales conclusiones, y seguidamente la bibliografía utilizada.

**Capítulo 1 LOS MÉTODOS BIPLLOT:
DEFINICIÓN Y CONCEPTOS
GENERALES**

1.1 Introducción

El creciente volumen de información que se genera en la actualidad ha irrumpido en todos los niveles; en consecuencia, el análisis de matrices de datos de alta dimensión exige el uso no solo de herramientas tecnológicas, sino de técnicas multivariantes que permitan extraer información no trivial, de manera implícita.

La caracterización de los individuos en función a las variables observadas no puede realizarse en un espacio de más de tres dimensiones; por lo tanto, es necesario reducir la dimensión del problema a un subespacio de dimensión dos (Cárdenas, Noguera, Galindo, & Vicente-Villardón, 2006). Por esto, es fundamental el uso de los métodos de *reducción de la dimensionalidad* que sean capaces de abarcar grandes volúmenes de información y que ayuden a simplificar la descripción del conjunto de datos. La técnica de reducción de la dimensión más popular es la Descomposición en Valores Singulares (DVS) (Eckart & Young, 1936) que genera el Análisis de Componentes Principales (ACP) (Hotelling, 1933; Pearson, 1901). Otros métodos, como el escalamiento multidimensional (Torgerson, 1952), el análisis discriminante (Fisher, 1936), el análisis de correspondencia (Benzécri, 1973; Greenacre, 1984), entre otros, comparten el mismo espíritu de la DVS en el sentido de que la aproximación de bajo rango de la matriz de datos extrae las estructuras ocultas en ellos, capturando a la vez la mayor variación posible. Las diferencias entre las técnicas de reducción de la dimensión dependen del tipo de variable y la métrica impuesta en el espacio de las filas o en el espacio de las columnas.

Los métodos Biplot propuestos por [Gabriel \(1971\)](#) son parte de estas técnicas de reducción de la dimensionalidad y constituyen una de las representaciones gráficas más informativas de datos multivariantes, relacionada fundamentalmente a la DVS de una matriz de datos, y por lo tanto al análisis de componentes principales. La principal diferencia con otras técnicas es que proporciona una representación bidimensional entre marcadores fila (individuos) y marcadores columna (variables). Las coordenadas de las filas son las coordenadas sobre las componentes principales y las coordenadas de las columnas, los vectores propios, que también se pueden ver como la proyección de los ejes unitarios en el espacio p dimensional, asociados a la matriz identidad.

1.2 Conceptos Generales

1.2.1 La Matriz de Datos

En el análisis multivariante; en particular, en los métodos Biplot, la información de base es una estructura de datos en forma de matriz, donde las *filas* (n) identifican a los *individuos* o *unidades estadísticas*, y las *columnas* (p) a cada una de las *variables* o *caracteres* medidas sobre dichos individuos (ver Figura 1-1 y Figura 1-2).

En el término Biplot, el prefijo <bi> debe ser percibido como una interpretación conjunta de filas y columnas; y, la expresión <plot> como gráfico o diagrama.

Un *Biplot* ([Gabriel, 1971](#)) es un término correspondiente a una técnica de análisis multivariante, que representa gráficamente datos de dos o más variables.

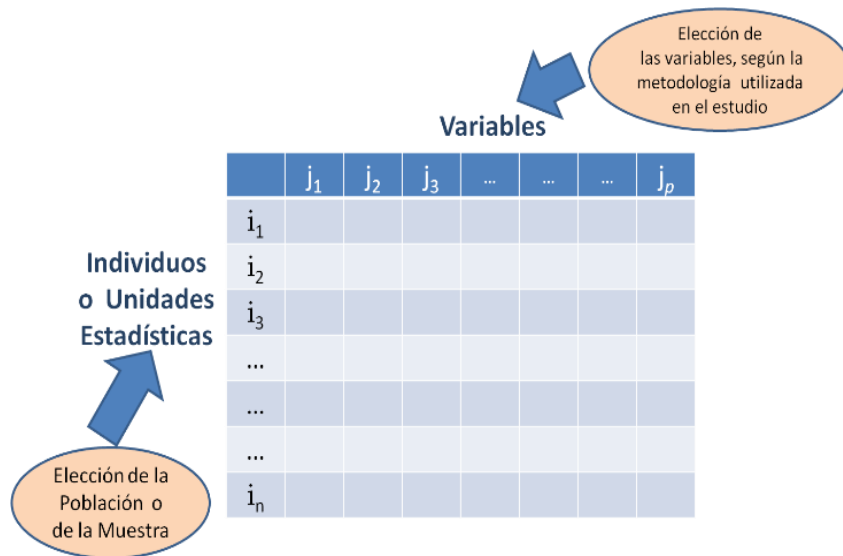


Figura 1-1 Esquema de la tabla de datos

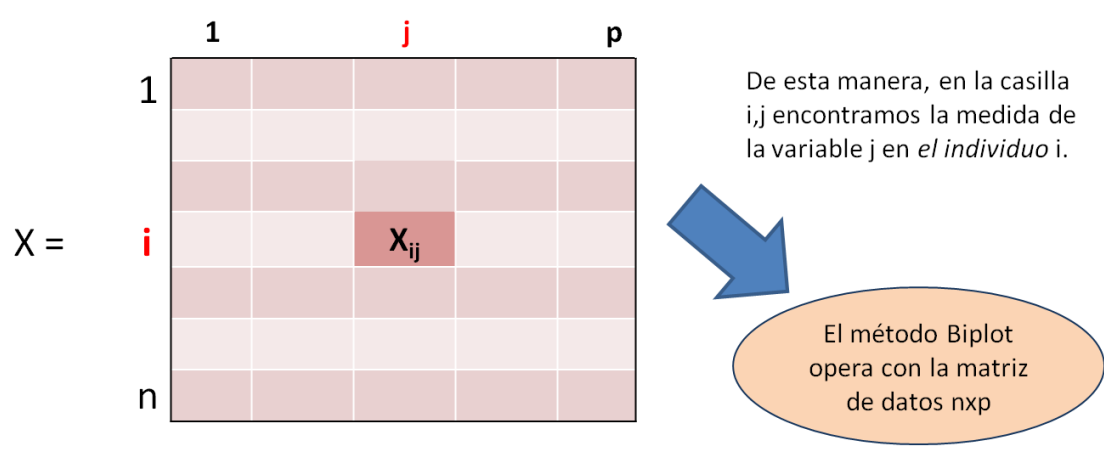


Figura 1-2 Representación de la matriz de datos

1.2.2 Descomposición en Valores Singulares

Generalmente, el biplot se obtiene del Análisis de Componentes Principales (ACP) (Hotelling, 1933; Pearson, 1901) a partir de la Descomposición en Valores Singulares (DVS) de una matriz de datos. Básicamente lo que busca el ACP es reducir la dimensión de un conjunto de variables aleatorias reteniendo la mayor cantidad de información posible. La eficacia de utilizar el

ACP está en eliminar convenientemente la información redundante e identificar variables latentes. Para ello, transforma las variables originales (generalmente correlacionadas), en factores o variables latentes, llamadas componentes principales (CPs) -no correlacionadas- que se corresponden con las direcciones en las que los datos tienen máxima varianza. Estos CPs están ordenadas de manera que los primeros contienen la mayor parte de la información de los datos originales (Figura 1-3).

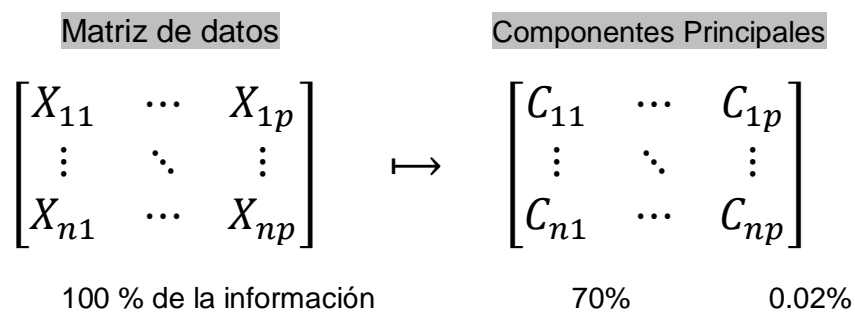


Figura 1-3 Transformación de la matriz original en componentes principales

Las componentes principales obtenidas mediante el ACP permiten visualizar a través del biplot la existencia de patrones y/o relaciones entre variables o entre variables e individuos.

En principio, debemos obtener una matriz X_q que ajuste lo mejor posible a X , a través de una buena aproximación mínimo cuadrática. Esta matriz se plantea en el sentido de que minimice la relación:

$$\sum_i^n \sum_j^p (x_{ij} - x_{(q)ij})^2 = \text{traza} [(X - X_{(q)})(X - X_{(q)})']$$

para todas las matrices $X_{(q)}$ de rango menor o igual a q .

En las siguientes secciones se explica el procedimiento, que concluye definiendo los marcadores para filas y para columnas que permiten la representación biplot.

La teoría matemática ofrece diferentes métodos para aproximar una matriz a bajo rango. La factorización más común se realiza mediante la DVS de la matriz original. A través de este método, formulado por [Eckart & Young \(1936, 1939\)](#), se aproxima la información global contenida en la matriz. Este algoritmo también puede estudiarse en [Young & Householder \(1938\)](#); [Gabriel \(1971\)](#) y [Greenacre \(1984\)](#).

1.2.3 Formulación Teórica

Para la construcción del Biplot se parte de una matriz X de orden $n \times p$ y se realiza una descomposición en valores singulares, considerando los r primeros sumandos; esto es:

$$X_{n \times p} = U_{n \times r} \Sigma_{r \times r} V'_{r \times p}$$

donde,

- X es la matriz de datos, cuyo rango es q , $q \leq \min(I, J)$
- U es la matriz de vectores propios de la matriz XX'
- V es la matriz de vectores propios de la matriz $X'X$ en la cual los vectores columna son ortonormales
- Σ es la matriz diagonal de valores singulares de la matriz X
($\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_r \geq 0$).

Estas matrices comprueban que $U'U = V'V = I$ asegurando la unicidad de la factorización.

En consecuencia, la mejor aproximación de la matriz X en rango q , expresada simbólicamente $X_{(q)}$ está dada por la relación:

$$X_{(q)nxp} = U_{(q)nxq} \Sigma_{(q)q} V'_{(q)qxp} = \sum_{k=1}^q \lambda_k u_k v'_k$$

donde,

$U_{(q)}$ es la matriz proyectada con las k primeras columnas de U

$V_{(q)}$ es la matriz proyectada con las k primeras columnas de V

$\Sigma_{(q)}$ es la matriz diagonal que contiene los q mayores valores propios (λ_k) distintos de cero de la matriz X

El método Biplot pretende aproximar la matriz X mediante $U_{(q)nxq}$, con una adecuada factorización, obteniéndose consecuentemente la expresión:

$$X \cong X_{(q)} \cong U \Sigma V^T \cong E_q G_q^T$$

donde, $E_{(q)}$ y $G_{(q)}$ son matrices de rango completo, definidas como

$$E_{(q)} = U_{(q)} \Sigma_{(q)}^\rho \quad y$$

$$G_{(q)} = V_{(q)} \Sigma_{(q)}^{1-\rho}$$

dependiendo del valor seleccionado para ρ ($0 \leq \rho \leq 1$).

1.2.4 Porcentaje de varianza explicada

En un análisis de componentes principales nos interesa conocer la proporción de varianza explicada por cada una de las componentes principales, para saber cuanta información presente en los datos se pierde por la proyección de las observaciones sobre las primeras componentes; esto es, la varianza contenida en los datos originales que es explicada por cada uno de ellos.

El porcentaje de varianza explicada es visto como una medida de bondad de ajuste del modelo.

$$\frac{\sum_{k=1}^q \lambda_k}{\sum_{i=1}^p \lambda_i}$$

1.2.5 Definición y Representación Gráfica

Sea $X_{n \times p}$ la matriz de datos que se desea representar mediante un gráfico Biplot. Como se ha mencionado, las filas corresponden a $\langle n \rangle$ individuos o unidades estadísticas y las columnas a $\langle p \rangle$ variables medidas sobre estos mismos individuos.

Un Biplot para la matriz definida es una representación gráfica (ver Figura 1-4) mediante marcadores, denominados: e_1, e_2, \dots, e_n para las filas de X y g_1, g_2, \dots, g_p para las columnas de X .

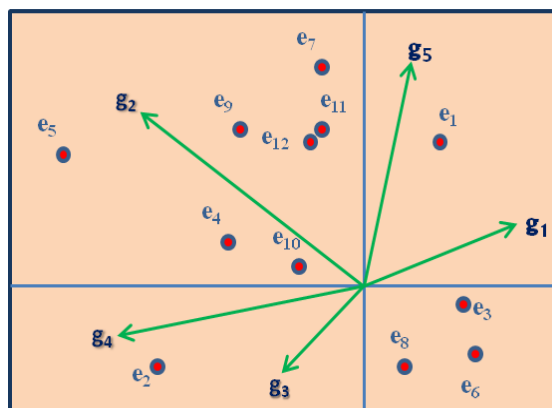


Figura 1-4 Esquema de una representación biplot de una matriz (12x5)

De esta manera, el producto escalar $e_i^T g_j$ aproxima como sea posible, al elemento x_{ij} de la matriz inicial de datos.

1.2.6 Representación Geométrica del Biplot

Teniendo en cuenta las matrices \mathbf{E} y \mathbf{G} , constituidas respectivamente por los marcadores fila $e_i(e_1, e_2, \dots, e_n)$ y los marcadores columna $g_i(g_1, g_2, \dots, g_p)$, podemos formular la relación:

$$X \cong \mathbf{E}\mathbf{G}^T$$

Ambos marcadores están representados en un espacio de dimensión $q \leq r$, en donde q simboliza el número de ejes retenidos y r el rango de la matriz de datos. La factorización en matrices de marcadores filas y de marcadores columna garantiza la representación Biplot aproximada de la matriz, puesto que cada x_{ij} puede expresarse en la forma:

$$x_{ij} = e_i^T g_j$$

Esta relación que establece una *forma bilineal* (Gollob, 1968) determina que la matriz inicial es similar al producto escalar $e_i^T g_j$ permitiendo una sistematización gráfica por medio de la proyección ortogonal de los marcadores fila sobre los marcadores columna y viceversa.

No obstante, Bradu & Gabriel (1978) y Gabriel (1998) demostraron que la interacción entre filas y columnas en una tabla de doble entrada, a partir de las características geométricas de un gráfico Biplot, posibilita el ajuste de modelos *bilineales* de tipo *multiplicativo*. También se tienen contribuciones de Denis (1991), Van Eeuwijk (1995), Choulakian (1996) y De Falguerolles (1996), quienes describen la interacción entre filas y columnas en modelos *bilineales aditivos* y *multiplicativos*.

En esta línea de pensamiento, diversos autores han estudiado los Biplot para describir la interacción de filas y columnas a través de modelos bilineales generalizados. [Cárdenas, Noguera, Galindo, & Vicente-Villardón \(2003\)](#) presentan un estudio en profundidad de esta alternativa, a partir de los métodos Biplot.

Resumiendo (ver Figura 1-5), el diseño general para la construcción del biplot se inicia a partir de una matriz de datos $X_{n \times p}$. Usualmente, estos datos iniciales necesitan transformarse (por ejemplo, centrando o estandarizando las variables, en otras alternativas). Luego, una descomposición en valores singulares extrae las dimensiones que comunican la mayor parte de la variabilidad de la información de la matriz original; para finalmente aproximar la matriz inicial mediante una apropiada factorización y obtener los puntos para materializarlos en el biplot.

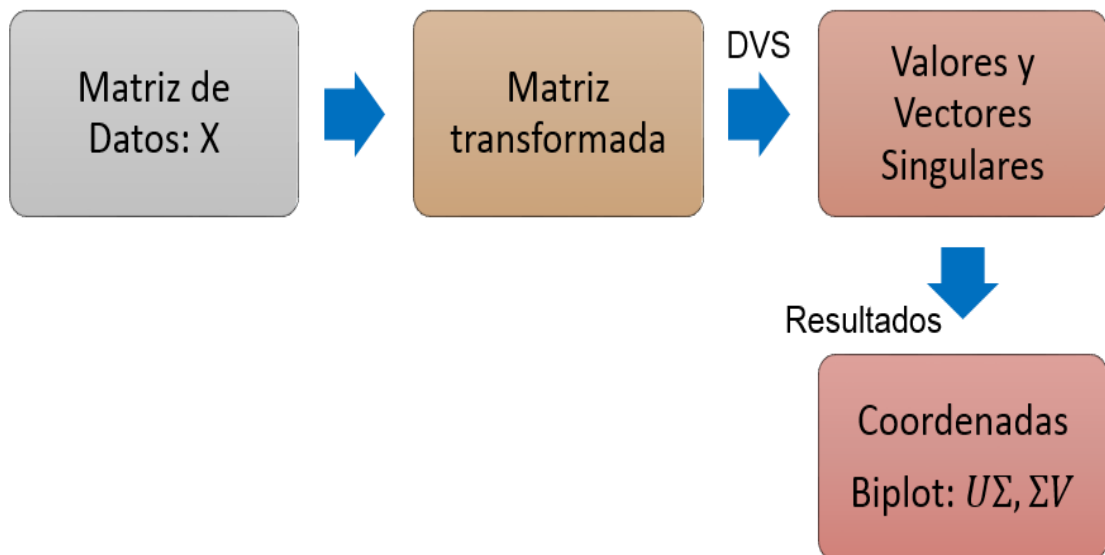


Figura 1-5 Pasos para la construcción del biplot

1.2.7 Interpretación Geométrica

En el biplot, la proyección de los marcadores fila sobre los marcadores columnas permite estudiar las relaciones entre individuos y variables. La interpretación del Biplot se basa, de forma general en las siguientes ideas (Figura 1-6):

- La similitud entre individuos (filas) es una función inversa de la distancia entre los mismos.
- Las longitudes y los ángulos de los vectores que representan a las variables se interpretan en términos de *variabilidad* y *covariación* respectivamente.
- Las relaciones entre filas y columnas se interpretan en términos de *producto escalar*, es decir, en términos de las proyecciones de los puntos “fila” sobre los vectores “columna”.

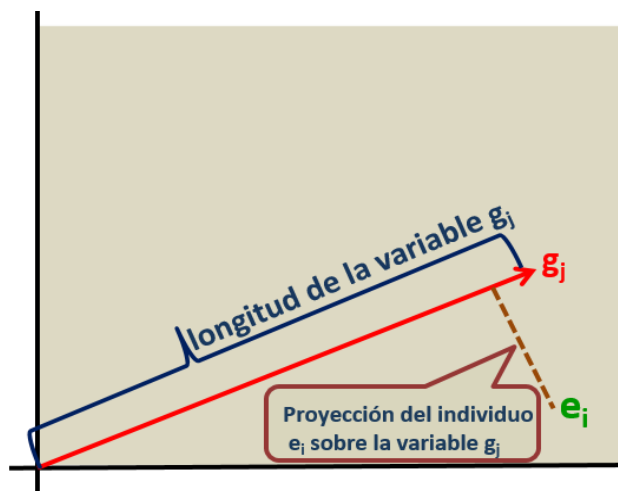


Figura 1-6 Interpretación geométrica del producto escalar

De esta manera, $\left\| \frac{\text{Proy } e_i}{g_j} \right\| \cong \text{Distancia de la proyección de } e_i \text{ sobre } g_j$

$\| g_j \| \cong \text{Distancia entre el origen y el extremo del vector } g_j$

En consecuencia, $x_{ij} \cong e_i^T g_j \cong \|\text{Proy}(e_i / g_j)\| (\text{signo}) \|g_j\|$

El orden de las proyecciones de cada uno de los marcadores fila, sobre un marcador columna (Figura 1-7), reproduce el orden de los elementos de la matriz inicial. Si examinamos las posiciones de las proyecciones de los marcadores fila (individuos) sobre cada marcador columna (que representa a cada variable), es posible ordenar los individuos en función del valor que toman en esa variable. Estas proyecciones se pueden hacer para cada una de las variables en estudio.

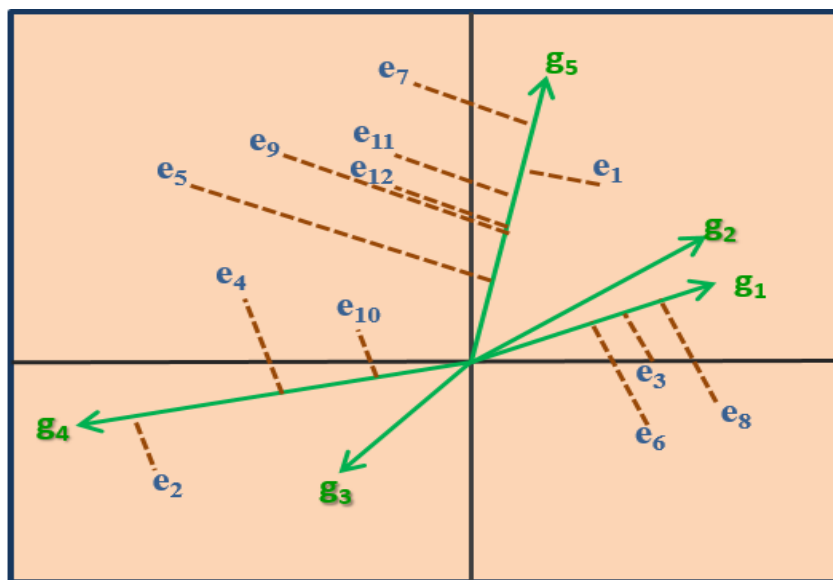


Figura 1-7 Proyección ortogonal de los marcadores fila sobre los marcadores columna

La representación Biplot sobre un sistema cartesiano y las propiedades geométricas, permiten interpretar el gráfico, en base a conceptos matemáticos elementales de ángulos, distancia, y longitud de un vector (ver Figura 1-8).

Los puntos representan los marcadores filas (individuos o unidades estadísticas). Los vectores representan los marcadores columnas, es decir las variables.

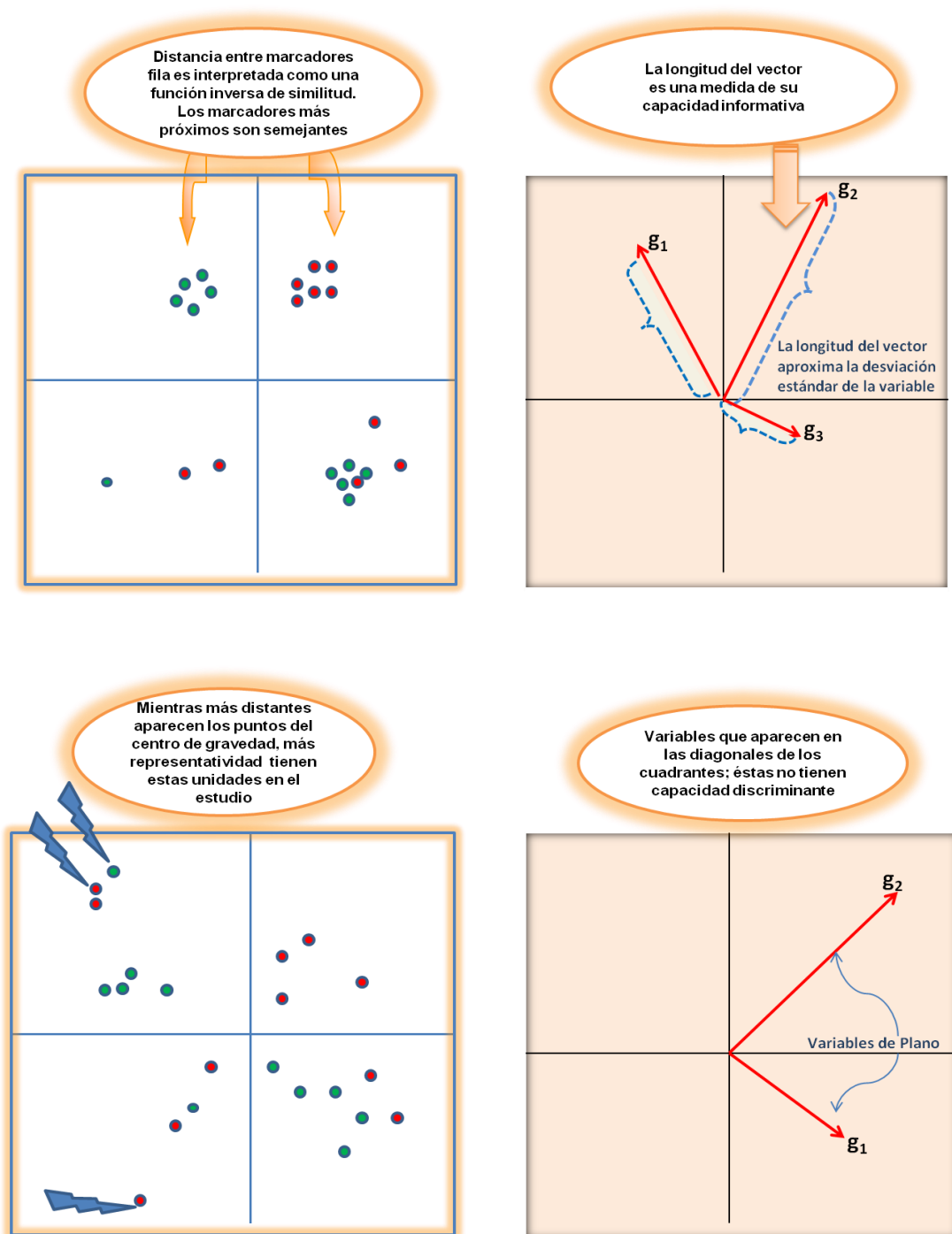


Figura 1-8 Reglas básicas de interpretación (a)

Los cosenos de los ángulos entre los vectores columna aproximan la correlación entre las variables (ver Figura 1-9 b), de forma tal que:

- Ángulos pequeños (agudos) entre variables indican correlaciones positivas altas entre estas variables
- Ángulos obtusos entre variables se asocian a correlaciones negativas altas entre estas variables.
- Ángulos rectos señalan que las variables no tienen relación y las mismas se comportan de forma independiente.

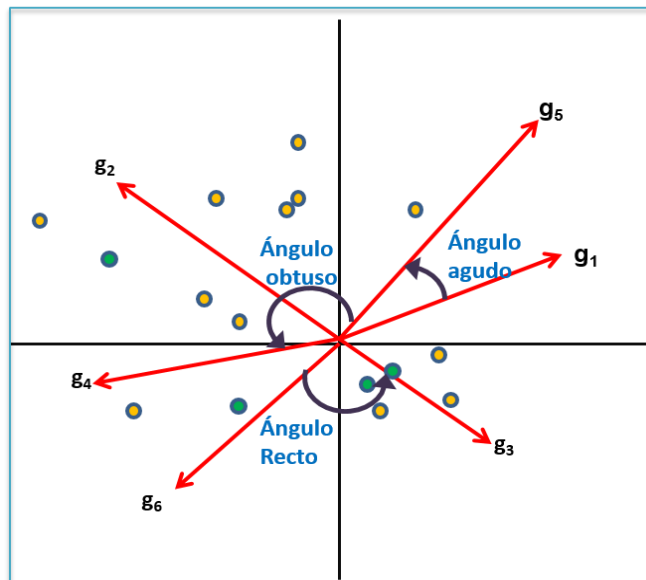


Figura 1-9 Reglas básicas de interpretación (b)

1.3 Contribuciones al Biplot

Desde sus orígenes, los métodos Biplot (Gabriel, 1971) se han constituido en métodos de representación gráfica por excelencia. De esta manera, en la Estadística Multivariante, se suman cada día nuevas investigaciones, que contribuyen aún más al desarrollo y aplicación de los métodos Biplot, con sus diversas variantes.

El Biplot se ha utilizado tradicionalmente con fines descriptivos y también en la diagnosis de modelos (Bradu & Gabriel, 1974, 1978). En la actualidad, los investigadores continúan realizando estudios utilizando la metodología Biplot y combinando ésta con otras técnicas de naturaleza clásica; como por ejemplo, el análisis de varianza, el análisis de componentes principales, entre otros.

Desde el punto de vista teórico, como práctico, surgen nuevas técnicas de análisis; en consecuencia, nuevas líneas de investigación cuyos resultados quedan recogidos en un Biplot. Paralelamente a este desarrollo, surgen nuevas alternativas de análisis de datos, a través de aplicaciones informáticas variadas. A continuación, se presenta una tabla resumen de las contribuciones más relevantes a la metodología biplot, transcurridas más de cuatro décadas desde su planteamiento.

Tabla 1: Contribuciones a la metodología Biplot

Referencia	Contribución
(Gabriel, 1971)	<ul style="list-style-type: none"> • Creador del Biplot Clásico • Establece dos tipos de modelos: GH-Biplot y JK-Biplot
(Greenacre, 1984)	<ul style="list-style-type: none"> • Propone una nueva forma de denominar el GH Biplot y JK Biplot • Les denomina CMP y RMP respectivamente
(Kempton, 1984)	<ul style="list-style-type: none"> • Demuestra que los métodos biplot aumentan la información obtenida con otros métodos.
(Galindo, 1986)	<ul style="list-style-type: none"> • Demuestra la conveniencia de representar filas y columnas sobre un mismo sistema de coordenadas • Define el HJ- Biplot

(Ter Braak, 1986, 1990)	<ul style="list-style-type: none"> • Utiliza los Biplot en el contexto del Análisis Directo del Gradiente • Muestra la utilidad del Biplot en la relación entre el análisis de redundancia y la correlación canónica
(Blázquez, 1998)	<ul style="list-style-type: none"> • Incluye información externa de filas y columnas • Utiliza modelos lineales generalizados alternos
(Gauch, 1988)	<ul style="list-style-type: none"> • Utiliza el biplot para la validación de modelos y estudia la relación genotipo ambiente.
(Gower & Harding, 1988)	<ul style="list-style-type: none"> • Definen los biplot no lineales
(Vicente-Tavera, 1992)	<ul style="list-style-type: none"> • Realiza una clasificación jerárquica ascendente utilizando el HJ Biplot • Basa esta relación en el concepto de inercia
(Gower, 1992)	<ul style="list-style-type: none"> • Señala que en el biplot la factorización no está basada en DVS
(Braak & Looman, 1994)	<ul style="list-style-type: none"> • Incorpora el biplot de la matriz de coeficientes de la regresión y el biplot basado en regresión de rango reducido.
(Vicente-Villardón, 1992)	<ul style="list-style-type: none"> • Define los biplot generalizados
(Fernández-Gómez, 1995)	<ul style="list-style-type: none"> • Propone el HJBiplot como una alternativa del Análisis Canónico de correspondencia
(Gower & Hand, 1996)	<ul style="list-style-type: none"> • Definen los biplot de interpolación y predicción
(Martín-Rodríguez, 1996)	<ul style="list-style-type: none"> • Desarrolla los Meta-biplot • Define una estructura consenso para estudiar la relación
(Carlier & Kroonenberg, 1996)	<ul style="list-style-type: none"> • En base a los modelos de Tucker y Tuckals3 proponen el biplot interactivo y conjunto para tablas de 3 vías.

(Cárdenas & Galindo, 2004)	<ul style="list-style-type: none"> • Estudian el biplot desde el punto de vista inferencial bajo la metodología de los modelos bilineales generalizados
(Amaro, 2001)	<ul style="list-style-type: none"> • Generaliza el MANOVA Biplot de una vía al caso de dos factores de variación
(Crossa, Cornelius, & Yan, 2002; Yan, Cornelius, Crossa, & Hunt, 2001; Yan & Hunt, 2002; Yan & Kang, 2002)	<ul style="list-style-type: none"> • Utiliza el biplot como modelos bilineales para el análisis de la interacción genotipo ambiente
(Baccalá, 2004)	<ul style="list-style-type: none"> • Desarrolla y hace una propuesta sobre el biplot de múltiples vías
(Vairinhos, 2003)	<ul style="list-style-type: none"> • Desarrolla un sistema de minería de datos a través de los métodos biplot
(Hernández & Galindo-Villardón, 2006)	<ul style="list-style-type: none"> • Profundiza en el estudio de la metodología biplot en presencia de valores atípicos
(Vicente-Villardón, Galindo-Villardón, & Blázquez-Zavallos, 2006)	<ul style="list-style-type: none"> • Desarrolla el biplot logístico
(Demey, 2008)	<ul style="list-style-type: none"> • Desarrolla la teoría del biplot logístico externo
(Egido, 2014)	<ul style="list-style-type: none"> • Desarrolla el Biplot dinámico
(Nieto-Librero, 2015; Nieto-Librero, Sierra, Vicente-Galindo, Ruíz-Barzola, & Galindo-Villardón, 2017)	<ul style="list-style-type: none"> • Implementa el Clustering Disjoint Biplot
(Nieto-Librero, 2015)	<ul style="list-style-type: none"> • Realiza el análisis biplot de forma inferencial, combinándolo con los métodos Bootstrap.
(Hernández Sánchez, 2016)	<ul style="list-style-type: none"> • Desarrolla la teoría del biplot para datos de tipo nominal y ordinal

Capítulo 2 **BIPLOT PARA TABLAS DE DOS VÍAS**

2.1 Introducción

Hemos visto que la descomposición en valores singulares es un método apropiado para aproximar una matriz a bajo rango. Como resultado de esta descomposición, procede entonces la elección de los marcadores tanto de filas como de columnas, acogiendo una métrica que conlleva el estudio de las propiedades del Biplot resultante.

2.2 Biplot Clásicos: GH Biplot y JK Biplot

Desde sus orígenes, los métodos Biplot, (Gabriel, 1971) se han constituido en métodos de representación gráfica por excelencia. Estas contribuciones, denominadas GH Biplot y JK Biplot deben su nombre a la notación utilizada para ambos marcadores. Además de los nombres, estos métodos varían entre sí, porque presentan diferencia en la bondad de ajuste.

El GH Biplot, debe su nombre a que Gabriel adaptó la notación G para simbolizar la matriz de los marcadores fila y H para la matriz de los marcadores columna.

$$G = U \quad H = V \Sigma$$

Por su parte, en el JK Biplot, Gabriel representó los marcadores fila con la notación J y los marcadores columna con H .

$$J = U \Sigma \quad K = V$$

Si se tiene la métrica identidad, los marcadores se eligen de acuerdo a los valores indicados para ρ al realizar la descomposición:

$$E = U \Sigma^\rho \quad G = V \Sigma^{1-\rho}$$

Si optamos por $\rho = 0$ se obtiene $E = U$ y $G = V\Sigma$ verificándose que $G'G = I$ con lo cual se obtiene el *GH Biplot*, que preserva la métrica solo para las *columnas*. De esta manera, en el *GH Biplot*, las columnas tienen una alta calidad de representación.

Si optamos por $\rho = 1$ se obtiene $E = U\Sigma$ y $G = V$ verificándose que $E'E = I$ con lo que obtenemos el *JK Biplot*, que preserva la métrica solo para las *filas*. De allí que, en el *JK Biplot*, las filas tienen una alta calidad de representación.

2.2.1 GH Biplot

Bajo el supuesto de que los datos de la matriz X están centrados, entonces la matriz simétrica $(X'X)$ es proporcional a la matriz de varianzas-covarianzas y el producto interno (g_j, g_k) expresa las covarianzas. Al definir un factor de

escala apropiado, la matriz $\frac{1}{n-1}X'X$ coincide con la matriz de covarianzas y se

establecen los marcadores para filas y columnas como: $E = \sqrt{n-1} U$ y $G =$

$$\frac{1}{n-1} (V\Sigma)$$

Sin perder de vista que la métrica utilizada en el *GH Biplot* es $E'E = I$ destacamos sus propiedades más importantes:

<p>Propiedad 1: El producto escalar de las columnas de la matriz X, coincide con el producto escalar de los marcadores columna, G</p>	$\begin{aligned} X'X &= (EG')'(EG') \\ &= GE'EG' \\ &= GU'UG' \\ &= GG' \end{aligned}$
--	--

<p>Al mismo tiempo, la aproximación del producto interno (varianza-covarianza, simbolizada S), en dimensión reducida, es óptima en el sentido de los mínimos cuadrados.</p>	$S = \frac{1}{n-1} X'X$ $= \frac{1}{n-1} GG'$ $= \frac{1}{n-1} V\Sigma^2V'$
<p>La mejor aproximación de la matriz de covarianzas en rango s es</p>	$S \cong S_{(s)} = \frac{1}{n-1} V_{(s)}\Sigma_{(s)}\Sigma_s V'_{(s)}$ $= G_{(s)}G'_{s(s)}$
<p><u>Propiedad 2:</u></p> <p>La distancia de Mahalanobis entre dos filas (i y j) de la matriz X es aproximada por la distancia euclídea entre dos marcadores fila.</p>	
$(X_i - X_j)'S^{-1}(X_i - X_j) = (Ge_i - Ge_j)'S^{-1}(Ge_i - Ge_j)$ $= (e_i - e_j)' G' S^{-1} G(e_i - e_j)$ $= \frac{1}{n-1} (e_i - e_j)' \Sigma V' S^{-1} V \Sigma (e_i - e_j)$ $= \frac{1}{n-1} (e_i - e_j)' \Sigma V' (n-1) (V\Sigma^{-2}V'V) \Sigma (e_i - e_j)$ $= (e_i - e_j)'(e_i - e_j)$	
<p><u>Propiedad 3:</u></p> <p>En el GH-Biplot existe una mejor aproximación para las varianzas covarianzas.</p> <p>Este método logra buenas propiedades para las variables, columnas de la matriz X, pues éstas aparecen bien representadas.</p>	
<p>La calidad de representación global se obtiene a través del cociente entre la suma de cuadrados de $X_{(q)}$ y la suma de cuadrados de los elementos de X. λ_k denotan los autovalores de $X^T X$.</p>	$\left(\lambda_1^2 + \lambda_2^2 / \sum_{k=1}^r \lambda_k^2 \right) \times 100$

2.2.2 JK Biplot

Seguimos bajo el supuesto de que los datos de la matriz X están centrados; pero, la métrica a utilizar es ahora $G'G = I$. Recordando lo pactado para los marcadores filas y los marcadores columna: $E = U \Sigma$ y $G = V$ respectivamente, se procede a destacar las propiedades más importantes para el JK Biplot.

<p>Propiedad 1: El producto escalar de las filas de la matriz X, concuerda con el producto escalar de los marcadores fila, E</p>	$\begin{aligned} XX' &= (EG')(EG')' \\ &= EG'GE' \\ &= EV'VE' \\ &= EE' \end{aligned}$
<p>Propiedad 2: Los marcadores para las filas coinciden con las coordenadas de los individuos en el espacio de las componentes principales.</p>	$\begin{aligned} XV &= U\Sigma V'V \\ &= U\Sigma \\ &= E \end{aligned}$
<p>Esta propiedad conlleva la posibilidad de estudiar las similitudes entre individuos, siempre y cuando, la distancia euclídea sea la adecuada.</p>	
<p>Propiedad 3: La calidad de representación global es mejor para filas que para columnas</p>	$\left(\frac{\lambda_1^2 + \lambda_2^2}{\sum_{k=1}^r \lambda_k^2} \right) \times 100$

M. J. Greenacre (1984) establece una nueva terminología para el GH - Biplot, denominada CMP Biplot (Column Metric Preserving) en torno al hecho de que preserva la métrica para las columnas; y, al JK - Biplot, le denomina RMP - Biplot (Row Metric Preserving), en función de que éste preserva la métrica para las filas.

2.3 HJ-Biplot

Como una alternativa para optimizar los métodos Biplot descritos por [Gabriel, \(1971\)](#), [Galindo \(1986\)](#) plantea la técnica multivariante denominada HJ-Biplot; contribución que logra el objetivo de representar con máxima calidad de representación ([Galindo & Cuadras, 1986](#)) y en forma simultánea las filas y columnas, sobre un mismo sistema de coordenadas. De esta manera, es posible interpretar conjuntamente las relaciones entre individuos y variables.

El HJ Biplot facilita la interpretación de las posiciones de filas, columnas y relaciones filas-columnas, a través de los ejes; como lo hace el Análisis Factorial de Correspondencias ([Benzécri, 1973](#); [Greenacre, 1984](#)).

En el HJ-Biplot, siguiendo la misma notación hasta ahora utilizada en los apartados anteriores, se parte de la descomposición en valores y vectores singulares de la matriz $X_{n \times p}$ definida previamente:

$$X \cong U \Sigma V^T$$

Bajo el supuesto de que los datos de la matriz X (filas y columnas) están centrados, los marcadores para las columnas en el HJ Biplot, se hacen coincidir con los marcadores columnas del GH Biplot; a su vez, los marcadores para las filas se hacen convenir con los marcadores filas del JK Biplot. Esto es,

$$E = U \Sigma \text{ y } G = V \Sigma$$

Por lo tanto, en el HJ -Biplot, las coordenadas para las filas coinciden con los marcadores para las filas en un JK Biplot; y las coordenadas para las columnas

coinciden con los marcadores para las columnas en un *GH* Biplot, respecto a los ejes factoriales (ver Figura 2-1).

En consecuencia, la bondad de ajuste en el HJ Biplot, expresada como calidad de representación global, se puede expresar como

$$\left(\lambda_1^2 + \lambda_2^2 / \sum_{k=1}^r \lambda_k^2 \right) \times 100$$

Teniendo en cuenta las relaciones entre U y V :

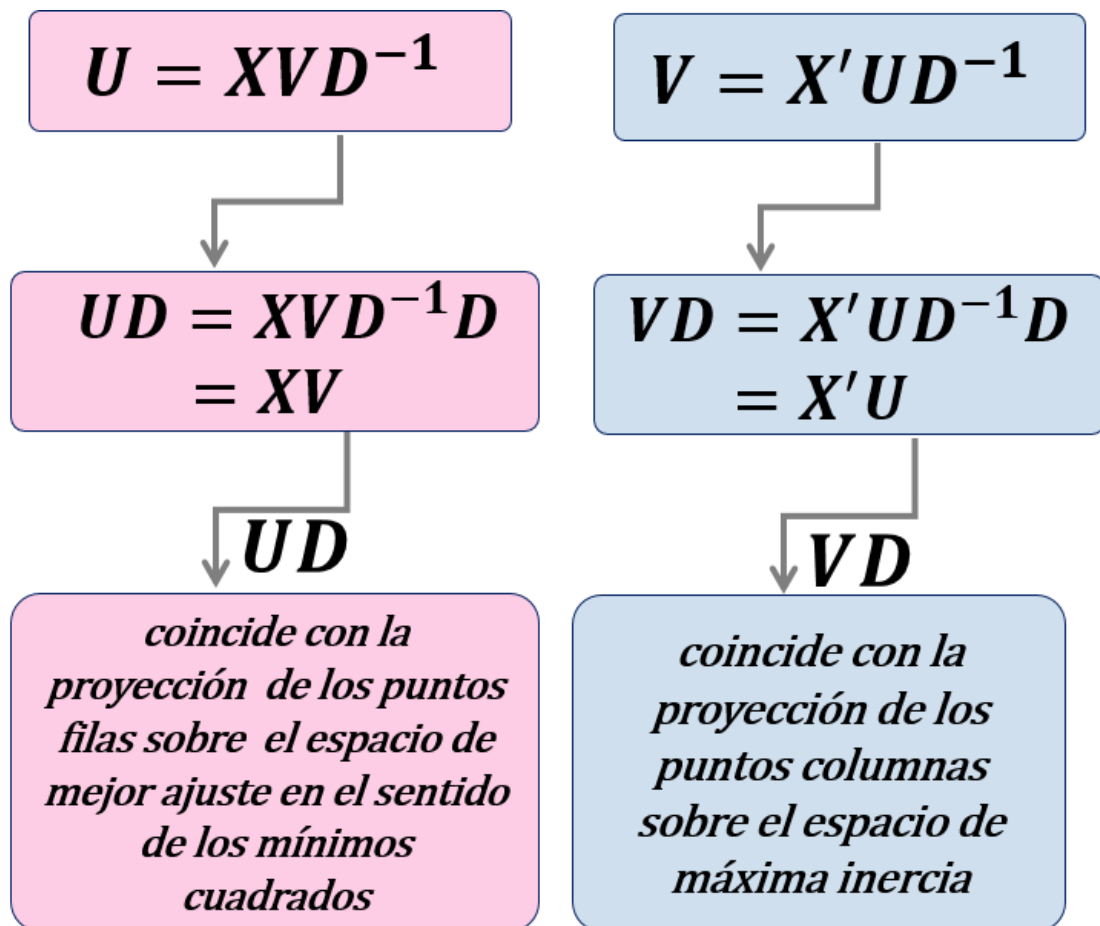


Figura 2-1 Marcadores para filas y columnas en el HJ-Biplot

2.4 Biplot Robusto

Tradicionalmente, la descomposición en valores y vectores singulares se aproxima mediante el Análisis de Componentes Principales de la matriz $X'X$. No obstante, [Hernández \(2005\)](#) plantea que ambos procedimientos, tanto la descomposición en valores singulares, como el análisis de componentes principales, son susceptibles a la presencia de valores atípicos (outliers) y propone una aproximación de los métodos Biplot a través de modelos bilineales, al cual ha denominado Biplot Robusto.

Este método, se constituye en una alternativa al Biplot clásico, bajo la presencia de valores atípicos que conlleven un comportamiento discordante.

La idea del Biplot Robusto ([Hernández & Galindo-Villardón, 2006](#)) es obtener estimadores para la matriz de marcadores, así como también para la matriz de cargas a partir de un modelo bilineal ([Gollob, 1968](#)) de la forma:

$$x_{ij} = \mu + \alpha_i + \beta_j + \sum_{l=1}^K \lambda_{jl} f_{il}$$

donde,

μ representa el efecto global común

α_i representa el efecto fila (individuos)

β_j representa el efecto columna (variables)

$\sum_{l=1}^K \lambda_{jl} f_{il}$ simboliza la interacción entre las filas y las columnas (producto escalar entre el vector de cargas y un vector de marcadores), donde λ_j es el vector de cargas y f_{il} es el vector de marcadores (filas y columnas)

Bajo este principio, se desarrolla el método, tomando valores iniciales arbitrarios para A ó B.

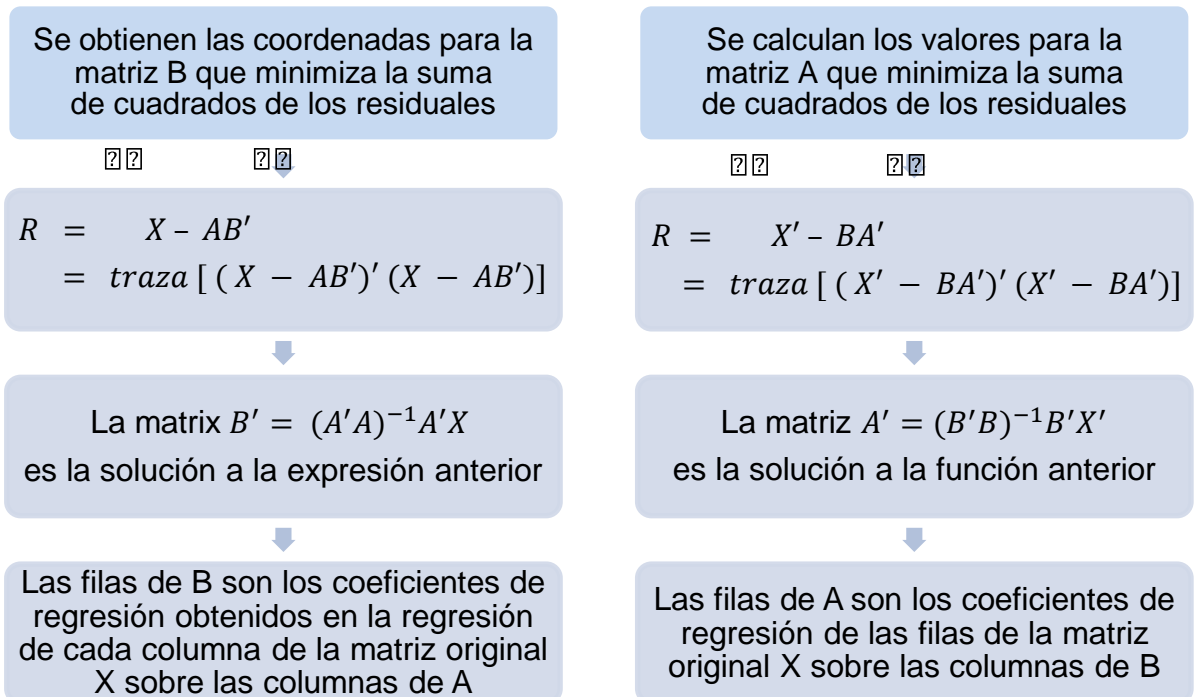
Escribiendo:

$$X = AB' + E'$$

y fijando las coordenadas para las columnas de A, tenemos que:

$$X' = BA' + E'$$

y fijando las coordenadas de B, tenemos que:



Estableciendo valores arbitrarios para la matriz A ó la matriz B, se logran estimadores para A (matriz de marcadores) y para B (matriz de cargas), de la misma forma que con la descomposición en valores singulares.

Si escogemos la primera columna de la matriz original, como inicio de la matriz de marcadores, es posible aplicar una regresión con lo cual estimamos la primera columna de la matriz B, esto es:

$$\hat{B} = (A'A)^{-1}X'A \text{ y se normaliza } \hat{B} = \frac{\hat{B}}{\text{norma } \hat{B}}$$

A través de la estimación de B , se realiza una regresión para estimar la primera columna de la matriz de A , en consecuencia:

$$\hat{A} = (\hat{B}'\hat{B})^{-1}X\hat{B}$$

Se procede a evaluar la suma de las desviaciones absolutas de los residuales, verificando si esta suma es mínima, con lo cual el proceso converge. El nivel de tolerancia propuesto es $1 * e^{-12}$. Con los valores de las estimaciones obtenidas de la primera columna de A y de B , se procede a estimar la matriz original:

$$\hat{X} = X - \hat{A}\hat{B}'$$

A continuación, se estima la segunda columna de la matriz A y de la matriz B , bajo el mismo procedimiento: escogemos la segunda columna de la matriz original, como inicio de la matriz de marcadores, y se calcula una regresión con lo cual estimamos la segunda columna de la matriz B . El proceso se repite hasta lograr las k (número de componentes principales) columnas, de ambas matrices, A y B . Finalmente, se procede a construir la representación Biplot, con las matrices A y B , donde se establece como marcadores para las filas a la matriz A y marcadores para las columnas a la matriz B .

En síntesis, el Biplot Robusto presenta una mejora con respecto a los métodos Biplot, ya que resiste la presencia de valores extremos o atípicos, con lo cual se puede realizar un mejor análisis exploratorio de los datos. No obstante, el uso del Biplot Robusto es limitado, a continuación citamos algunas de sus aplicaciones.

2.5 MANOVA-Biplot

Las diferentes técnicas multivariantes presentan algún grado de dificultad, para la interpretación de resultados de experimentos cuyos diseños buscan estudiar las diferencias entre los grupos y en los cuales se cuenta con muchas variables. En algunos casos se recurre a analizar las diferencias para cada una de las variables por separado; con lo cual podrían cometerse errores y quedar estos enmascarados al momento de la interpretación de los resultados obtenidos.

En este sentido, el Análisis Multivariante de la varianza (MANOVA) es una alternativa que intenta caracterizar las diferencias entre los vectores de medias de los grupos; una generalización del ANOVA, en términos multivariantes. Bajo este principio, surgen nuevos métodos que permiten estudiar las principales diferencias entre grupos, y a la vez determinar las variables comprometidas con dichas diferencias.

El MANOVA Biplot de una vía ([Gabriel, 1972](#)), así como también el llamado Biplot Canónico ([Vicente-Villardón, 1992](#)) son propuestas orientadas a la obtención de representaciones ponderadas de la matriz de medias de los grupos, basadas en el Biplot.

El Biplot Canónico se concibe como una representación gráfica del análisis discriminante que incorpora información simultánea sobre los grupos y las variables; y, en consecuencia, separa los grupos con máximo poder discriminante a partir de las variables originales.

Amaro et al. (2004) generalizan el MANOVA Biplot de dos vías (Gabriel, 1972), en donde la matriz de parámetros del modelo MANOVA, se apoya en cuatro representaciones biplot que recogen las diferentes fuentes de variación:



Cada uno de éstos persigue un objetivo particular, con cierto grado de diferencia en la forma de escoger las matrices en la descomposición inicial. Esta generalización sienta sus bases en el modelo ANOVA de la matriz original $X_{n \times p}$ (ver Figura 2-2).

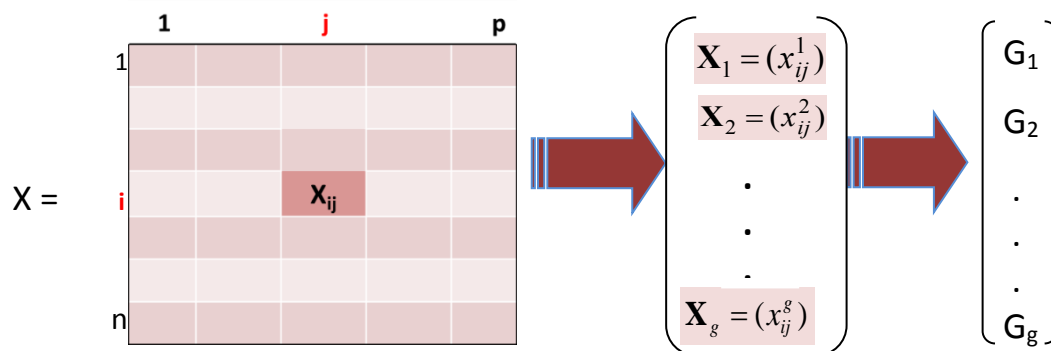


Figura 2-2 Modelo ANOVA de una matriz original $X_{n \times p}$

donde las filas de la matriz original X se dividen en grupos y la matriz está centrada previamente por columnas:

$$\begin{aligned} \bar{X}_1 &= (\bar{x}_{11}, \dots, \bar{x}_{1p}) \\ \bar{X}_2 &= (\bar{x}_{21}, \dots, \bar{x}_{2p}) \\ &\vdots \\ &\vdots \\ \bar{X}_g &= (\bar{x}_{g1}, \dots, \bar{x}_{gp}) \end{aligned}$$

para definir la matriz de medias de los grupos (g):

$$\bar{X} = \begin{pmatrix} \bar{X}_{11} & \bar{X}_{12} & \dots & \bar{X}_{1p} \\ \bar{X}_{21} & \bar{X}_{22} & \dots & \bar{X}_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \bar{X}_{g1} & \dots & \dots & \bar{X}_{gp} \end{pmatrix} = \begin{pmatrix} \bar{X}_1 \\ \vdots \\ \bar{X}_g \end{pmatrix}$$

El manejo de esta estructura nos lleva a plantear el modelo MANOVA, en la forma:

$$X = A\beta + \varepsilon$$

donde,

$X_{n \times p}$ representa la matriz original de los valores observados,

$A_{n \times p}$ es la matriz de q variables independientes observadas en cada individuo, con valores 1 y 0 para representar la pertenencia del individuo a un determinado grupo,

$\beta_{q \times p}$ es la matriz de parámetros de regresión,

$\varepsilon_{n \times p}$ es la matriz que recoge las fuentes de variación aleatorias o errores aleatorios

La estimación de los parámetros es similar al caso univariado. \hat{B} es el estimador de mínimos cuadrados de β que minimiza

$$\|\varepsilon'\varepsilon\| = \text{tr} [(X - A\beta)'(X - A\beta)]$$

$$\hat{B} = (A'A)^{-1}A'X$$

En el MANOVA, la hipótesis general multivariante se define en la forma:

$$H_0: C\beta = 0$$

donde la matriz C de dimensión $g \times q$ y rango g ($g \leq r$) puede ser elegida de formas diferentes; r es el rango de la matriz A . La matriz A se forma por los coeficientes de contrastes para las medias.

Así mismo, es posible construir el Biplot para las distintas hipótesis. [Mardia, Kent, & Bibby, \(1979\)](#) definen las filas de la matriz C como el efecto de las combinaciones lineales de las variables independientes.

Sea $D = C\beta$ un estimador de D definido por $\hat{D} = C\hat{\beta} = C(A'A)^{-1}A'X$

entonces la matriz de suma de cuadrados y productos “entre” grupos está dada por:

$$H = \underbrace{X'A(A'A)^{-1}C'}_{\hat{D}'} \underbrace{[C(A'A)^{-1}C']}_{R^{-1}} \underbrace{C(A'A)^{-1}A'X}_{\hat{D}}$$

y la matriz de suma de cuadrados y productos “dentro” de grupos, está definida por:

$$E = X'[I - A(A'A)^{-1}A']X$$

siendo $R = C(A'A)^{-1}C'$

2.5.1 MANOVA Biplot (una vía)

Al realizar la Descomposición en Valores Singulares Generalizada (DVSG), se construye la representación Biplot para la matriz de estimadores de los parámetros \hat{D} con $R^{-1/2}$ y $E^{-1/2}$ como métricas:

$$R^{-1/2}\hat{D}E^{-1/2} = U\Sigma V' \quad \text{por lo tanto, } \hat{D} = R^{1/2}U\Sigma V'E^{1/2}$$

Tomando:

- Como marcadores filas (grupos) a la matriz:

$$P = R^{1/2}U\Sigma$$

- Como marcadores para las columnas (variables) a la matriz:

$$Q = E^{1/2}EV'$$

La aproximación de la matriz de parámetros está dada por:

$$\begin{aligned} PQ' &= RU\Sigma V'E^{1/2} \\ &= R^{1/2}R^{-1/2}\widehat{D}E^{-1/2}E^{1/2} \\ &= \widehat{D} \end{aligned}$$

lográndose máxima separación entre grupos.

2.5.2 MANOVA Biplot (de dos vías)

Como se mencionó, [Amaro et al. \(2004\)](#) establecen cuatro tipos de Biplot en un diseño de dos vías, para representar resultados de un MANOVA.

Biplot Total:

Equivale a tomar todos los tratamientos (combinaciones de niveles de los factores) como MANOVA unidimensional ([Gabriel, 1995](#)).

$$\begin{aligned} R &= (A'A)^{-1} \\ \widehat{D} &= (A'A)^{-1}A'X \\ C &= I \end{aligned}$$

Estos resultados combinan los efectos principales (de filas y columnas) y la interacción, lo cual no es recomendable, ni mucho menos concluyente. No obstante, facilita el estudio de los grupos, en cuanto a sus diferencias y similitudes; reconoce las variables que influyen al contrastar los grupos y como un valor adicional, da las regiones de confianza para medias.

El Biplot de Filas:

En éste:

$$\begin{aligned} R &= C_1(A'_1A_1)^{-1}C'_1(A'A)^{-1} \\ \widehat{D} &= C_1(A'_1A_1)^{-1}A'_1X \end{aligned}$$

Denótese la matriz de medias de las filas como

$$\bar{X}_{fila} = (A'_1 A_1)^{-1} A'_1$$

Los marcadores para los grupos son:

$$\bar{P}_{fila} = \bar{X}_{fila} E^{-1/2} V ;$$

Los marcadores para las columnas son

$$Q = E^{-1/2} V$$

En el Biplot de filas, las propiedades son:

- Los ejes del Biplot de filas son combinaciones lineales de las variables que maximizan la F de Snedecor para el efecto del factor fila.
- Permite estudiar similitudes y diferencias entre los grupos, cuando el efecto es significativo.
- El producto escalar entre los marcadores fila para \hat{D} aproxima los productos escalares para las métricas, que coincide con la distancia de Mahalanobis.
- Sobre los centroides de los grupos (puntos estrellas), es posible superponer círculos de confianza y medir el efecto del factor fila para analizar la significación de las medias de los grupos.
- La importancia de las variables en la separación de los grupos se mide de la misma forma como en el MANOVA Biplot de una vía.

Biplot de Columnas:

El Biplot de columnas se da mediante la sustitución de las matrices:

$$R = C_2 (A'_2 A_2)^{-1} C'_2$$

$$\hat{D} = C_2 (A'_2 A_2)^{-1} A'_2 X$$

Denótese la matriz de medias para las columnas

$$\bar{X}_{col} = (A'_2 A_2)^{-1} A'_2 X$$

Los marcadores para los grupos son

$$\bar{P}_{col} = \bar{X}_{col} E^{-1/2} V$$

y, los marcadores para las variables (columnas) como

$$Q = E^{1/2} V$$

El Biplot de columnas maneja propiedades similares a las propiedades del Biplot fila; además, el producto escalar entre marcadores de columnas para \hat{D} aproxima la matriz de suma de cuadrados y productos dentro de los grupos,

$$QQ' = E^{1/2} VV'E^{1/2} = E$$

Biplot de Interacción:

Para la interacción se sustituyen las matrices:

$$R = C_3(A'_3 A_3)^{-1} C'_3$$

$$\hat{D} = C_3(A'_3 A_3)^{-1} A'_3 X$$

Los marcadores para los centroides de los grupos son

$$\bar{P} = \bar{X}_{int} E^{-1/2} V$$

Los marcadores para las variables son

$$Q = E^{-1/2} V$$

y, los marcadores para el conjunto de contrastes son

$$P = R^{1/2} U \Sigma$$

En este caso, las propiedades son:

- Los ejes son combinaciones lineales de variables que maximizan la F de Snedecor para la interacción.
- La bondad de ajuste para la representación de los contrastes es

$$\frac{\sum_{j=1}^d \lambda_j^2}{\sum_{j=1}^r \lambda_j^2}$$

- La longitud al cuadrado del marcador aproxima la distancia de Mahalanobis al vector nulo.
- En la proyección de las medias, la longitud al cuadrado del marcador que representa a una media aproxima la distancia de Mahalanobis al vector nulo. Se suele interpretar como la parte de la distancia explicada por la interacción:

$$\bar{P} \bar{P}' = \bar{X}_{int} E^{-1/2} V V' E^{-1/2} \bar{X}_{int} = \bar{X}_{int} E^{-1} \bar{X}_{int}$$

- La interacción será máxima para aquellos vectores paralelos al vector de interacción resultante, y será mínima para los vectores perpendiculares.

2.5.3 Representación gráfica del MANOVA Biplot

Los resultados del MANOVA Biplot son representados en el plano, donde los vectores constituyen los marcadores columnas (variables), y los grupos son incorporados como puntos estrellas rodeados de círculos (puntos-círculos), estos círculos representan las regiones de confianza de las medias poblacionales para las variables con un nivel de confianza preestablecido (ver Figura 2-3). Los radios de los círculos suelen ser diferentes, ya que cada grupo se pondera en función del tamaño de la muestra.

En términos generales, el MANOVA Biplot se interpreta siguiendo los siguientes criterios:

- Dos marcadores columnas que se encuentren próximos, indican que ambas variables están correlacionadas.
- Al proyectar los grupos sobre los vectores o variables, podemos aproximar una prueba estadística de diferencia de grupos similar a la prueba t de Student para dos muestras. Esto es, si las proyecciones logran cruzarse, entonces se infiere que no existe diferencia significativa entre estos dos grupos. Por el contrario, si las proyecciones de los grupos no se cruzan, o bien no logran intersecciones, se deduce que existen diferencias significativas entre estos dos grupos.

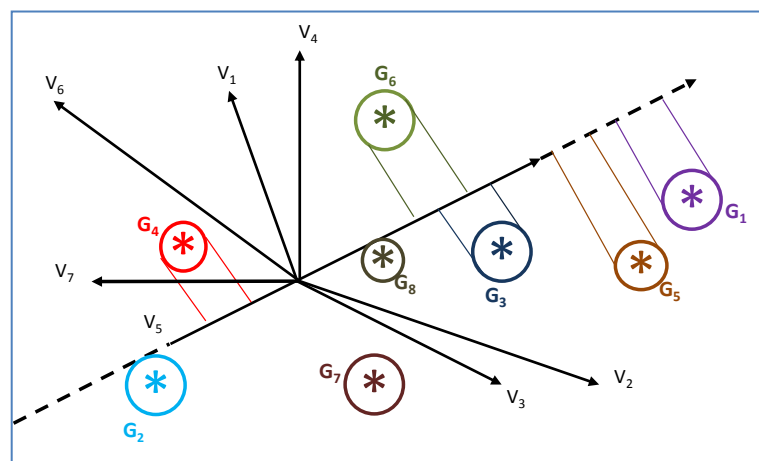


Figura 2-3 Representación de un biplot canónico

- La proyección de un grupo que se estime cercano al extremo del vector de un marcador columna revela que el grupo toma valores altos en esa variable.
- El coseno del ángulo de los vectores mide el grado de relación de las variables. Dos vectores que se encuentren próximos, explica que esas variables están correlacionadas.

2.6 Modelos AMMI GGE Biplot

Usualmente, un gran número de genotipos se someten a pruebas, en diferentes ambientes, lo que hace difícil determinar el patrón de respuesta genotípica, sin la ayuda de una visualización gráfica (Yan, Cornelius, Crossa, & Hunt, 2001).

Esta respuesta o reacción de los genotipos en condiciones ambientales diferentes es el resultado de la relación entre el genotipo (G) y el ambiente (E), y es denominada interacción genotipo ambiente (GEI).

Para evaluar genotipos probados en diferentes ambientes, se realizan análisis de *adaptabilidad* y *estabilidad*. La *adaptabilidad* es la capacidad del genotipo de responder positivamente a los efectos ambientales, garantizando un alto nivel de productividad (Becker, 1981). La *estabilidad* está relacionada con el sostenimiento o subsistencia de la productividad en diferentes entornos.

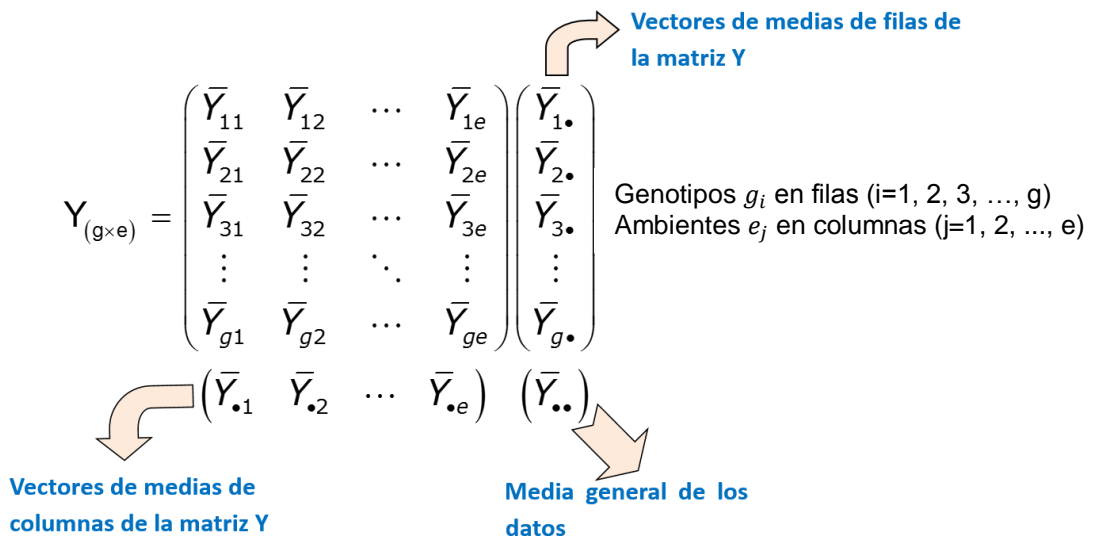
Existen diversas metodologías para estudiar la adaptabilidad y estabilidad fenotípica, las cuales difieren en cuanto a sus conceptos y procedimientos biométricos de estimación.

Dentro de las metodologías planteadas, podemos destacar el análisis AMMI (Additive Main Effects and Multiplicative Interaction) propuesto por Mandel (1971) y el GGE Biplot (Yan, Hunt, Sheng, & Szlavnic, 2000). Ambas metodologías, se basan en los gráficos Biplot, por tanto, permiten representar gráficamente una matriz de datos.

2.6.1 Biplot AMMI

El modelo AMMI (Efectos Principales Aditivos y Análisis de Interacción Multiplicativa), estudiado también por [Zobel, Wright, & Gauch \(1988\)](#) y validado por [Gauch & Zobel \(1989\)](#) y [Crossa, Gauch, & Zobel \(1990\)](#), combina el análisis de regresión lineal con el análisis por componentes principales (PCA), bajo los supuestos de que los efectos principales (genotipo y ambiente) son de naturaleza aditiva y la interacción genotipo x ambiente es de naturaleza multiplicativa.

Bajo el supuesto de que un conjunto de genotipos (g) han sido probados experimentalmente en diferentes ambientes (e); la media de cada combinación de genotipo y ambiente puede ser presentada en una matriz de dimensión $g \times e$.



El modelo AMMI está representado por la ecuación ([Zobel et al., 1988](#)):

$$Y_{ij} = \mu + g_i + e_j + \sum_{k=1}^p \lambda_k \gamma_{ik} \alpha_{jk} \varepsilon_{ij}$$

donde:

- Y_{ij} es el rendimiento del genotipo i en el ambiente j
- μ es la media general
- g_i es el efecto del genotipo i
- e_j es el efecto del ambiente j
- λ_k es el valor propio del componente principal K
- γ_{ik} son los vectores propios unitarios genotípicos asociados a λ_k
- α_{jk} son los vectores propios unitarios ambientales asociados a λ_k
- ε_{ij} es el error experimental
- k es el número de ejes de componentes principales considerados en el modelo AMMI

El modelo AMMI primero realiza un Análisis de Varianza (ANOVA) con dos factores a partir de la matriz de medias para calcular los principales efectos *aditivos* de genotipo y ambiente. Luego, realiza un Análisis de Componentes Principales (ACP) sobre los residuos de este modelo aditivo para estudiar los efectos multiplicativos de la interacción (Gauch, 1988). Los residuos obtenidos a partir de la matriz de medias constituyen la matriz de *interacciones* representada por:

$$GE_{(gxe)} = [(\hat{g}_e)_{ij}]$$

donde $(\hat{g}_e)_{ij} = \bar{Y}_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y}_{\cdot\cdot}$

Los términos de la interacción multiplicativa son estimados por medio de la descomposición en valores singulares a partir de la matriz GE luego de ajustar por mínimos cuadrados el modelo de efectos principales.

Los componentes principales se extraen de la matriz de covarianzas, ya que el ACP realizado de esta manera para los genotipos y para los ambientes permite obtener así las coordenadas ambientales y genotípicas respectivamente.

El modelo AMMI genera un gráfico de dos dimensiones (BIPLOT) en el que se pueden observar las diferencias entre ambientes (columnas), el grado de interacción de los genotipos (filas) con el ambiente, la estabilidad y las adaptaciones específicas de algunos genotipos a determinados ambientes.

El efecto de interacción entre un genotipo y un ambiente está dado por la proyección ortogonal del vector del genotipo sobre la dirección determinada por el vector del ambiente. Aquellos vectores de los ambientes que poseen la misma dirección que los vectores del genotipo se dice que tienen *interacción positiva*, es decir dichos ambientes son *favorables* para esos genotipos; por el contrario, los vectores en *direcciones opuestas* tienen *interacción negativa*, son ambientes desfavorables.

Podemos también, interpretar a la distancia entre dos genotipos. El coseno del ángulo entre los vectores de dos genotipos o ambientes indica la correlación entre ellos con respecto a su interacción. Así, ángulos *agudos* entre los vectores indican correlación positiva entre genotipos. Cuando los ángulos son *obtusos*, esto indica que existe correlación negativa, y si las direcciones son *opuestas* la correlación es inversa. Por otro lado, direcciones *perpendiculares* entre vectores indican correlación nula; y, vectores próximos hacen suponer que existe alta correlación entre dichos vectores.

Otro enfoque para estudios de interacción genotipo ambiente, es propuesto por Yan et al.(2000), denominado GGE Biplot.

2.6.2 GGE Biplot

El Biplot GGE es un modelo similar al Modelo AMMI, pero los términos lineales de genotipos no se consideran individualmente y se adicionan al término multiplicativo de la interacción genotipo x ambiente. El GGE Biplot está basado en el modelo SREG (Modelo de Regresión de Sitios), por lo tanto, el modelo para el GGE Biplot está dado por:

$$Y_{ij} = \mu + e_j + \underbrace{\sum_{k=1}^p \lambda_k \gamma_{ik} \alpha_{jk}}_{\text{Efecto Multiplicativo}} + \varepsilon_{ij}$$

El modelo del GGE Biplot conserva unidos G y GE, y particiona GGE en dos términos multiplicativos.

$$Y_{ij} = \mu + e_j + \lambda_1 \gamma_{i1} \alpha_{j1} + \lambda_2 \gamma_{i2} \alpha_{j2}$$

donde:

- Y_{ij} Es el rendimiento del genotipo i en el ambiente j
- μ Es la media global
- e_j Es el efecto del ambiente j
- λ_1 Es el valor propio de la primera componente principal
- λ_2 Es el valor propio de la segunda componente principal
- γ_{i1} Son los vectores propios del genotipo i asociados a la primera componente
- γ_{i2} Son los vectores propios del genotipo i asociados a la segunda componente
- α_{j1} Son los vectores propios ambientales j para la primera componente
- α_{j2} Son los vectores propios ambientales j para la segunda componente
- ε_{ij} Es el error experimental

La descomposición en valores singulares en el análisis GGE Biplot permite calcular los componentes principales, y también proporciona una medida de la variabilidad capturada por cada una de las componentes (Yan, 2002; Yan & Hunt, 2002; Yan & Tinker, 2006). La primera componente, se encuentra altamente correlacionada con el efecto principal del genotipo; por tanto, representa la proporción del rendimiento que se debe solo a las características del *genotipo*. La segunda componente representa la parte del rendimiento debido a la interacción *genotipo-ambiente*.

El Biplot GGE permite examinar, mediante la interacción genotipo ambiente, la capacidad de discriminar y la representatividad de los ambientes de prueba como una medida conveniente para definir mega-ambientes (ver Figura 2-4) homogéneos (Yan & Hunt, 2002; Yan et al., 2000; Yan & Kang, 2002). En la siguiente figura se pueden apreciar dos mega-ambientes claramente definidos (elipses grises).

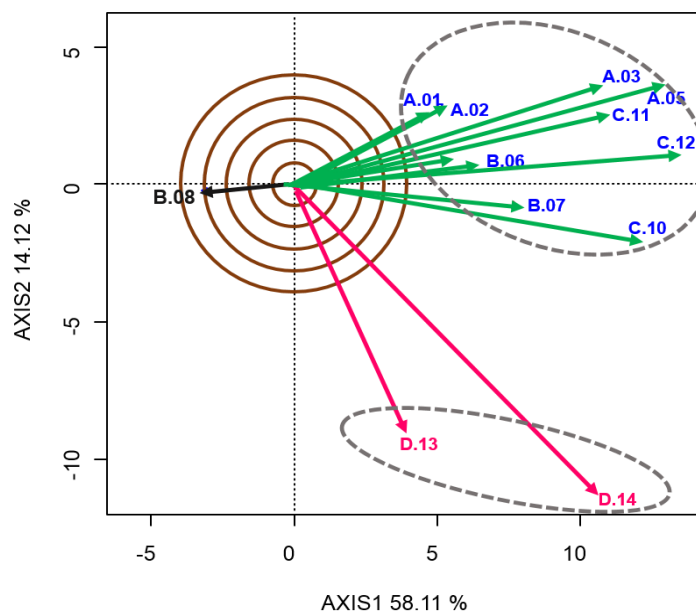


Figura 2-4 Relación entre ambientes

Para la interpretación gráfica, se siguen las mismas reglas de ángulos, distancia y longitud de marcadores fila y marcadores columna señalados anteriormente para el HJ-Biplot, sólo que el GGE Biplot se hace la interpretación en términos de genotipo (filas) – ambiente (columna).

Así, es posible identificar el genotipo ideal (Yan, 2001) - ver Figura 2-5- como aquel con alta puntuación en el primer eje del componente principal que está asociado a altos rendimientos; y, las puntuaciones cercanas a cero en el segundo eje del componente principal, relacionado con buena estabilidad del genotipo a través de los ambientes contrastantes.

Dentro de un mega-ambiente el entorno de prueba ideal es el más discriminativo (informativo) y el más representativo.

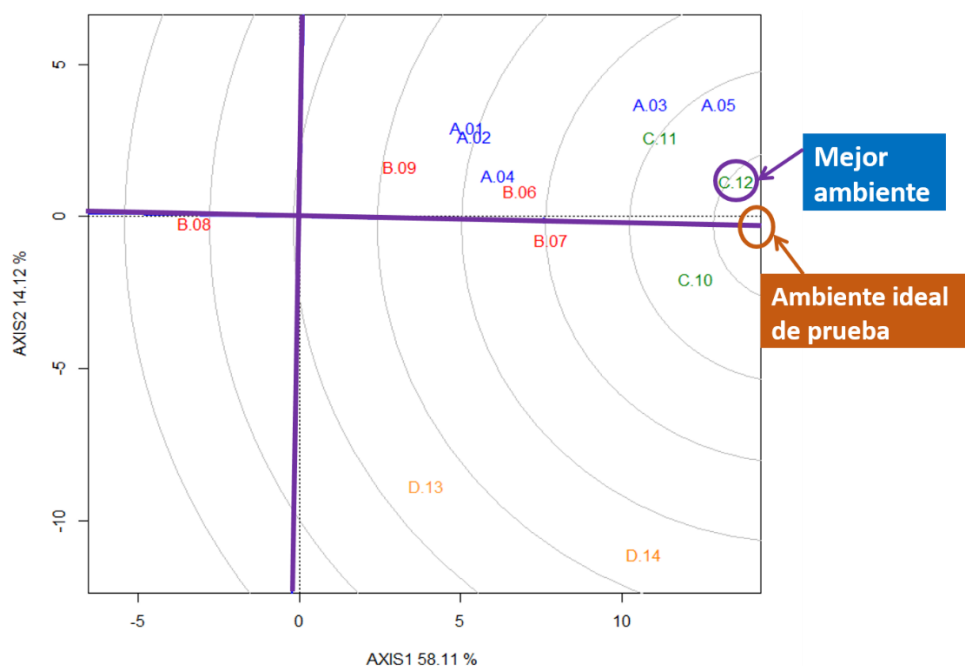


Figura 2-5 Ambiente ideal de prueba

El modelo GGE Biplot también permite examinar gráficamente cada genotipo y conocer el rendimiento de éste en función de los diferentes ambientes de prueba (ver Figura 2-6). Los ambientes están ordenados en la dirección del eje genotipo en términos de rendimiento. La recta perpendicular separa los ambientes en los que el rendimiento del genotipo está por encima de la media de aquellos en los que está por debajo de la media.

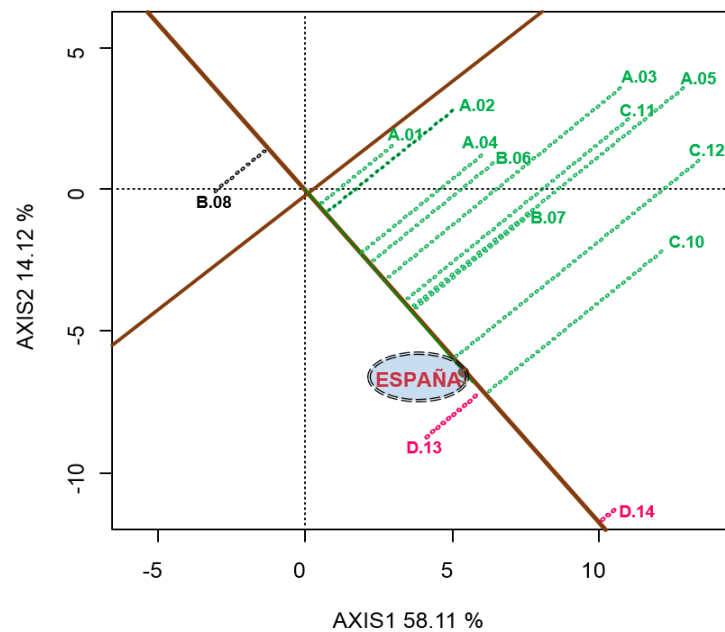


Figura 2-6 Rendimiento de un genotipo en los distintos ambientes

En síntesis, la gráfica del GGE Biplot permite visualizar tres aspectos importantes:

1. Determina el genotipo con mejor comportamiento en un ambiente específico.
2. Identifica el ambiente más apropiado para un genotipo específico.
3. Compara pares de genotipos en un ambiente y diferencia mega-ambientes.

2.7 Biplot para Datos Composicionales

Gran parte de los datos provenientes de diversas ramas de la ciencia, y datos que surgen como resultado de muchos experimentos, tienen carácter composicional; esto es, describen cuantitativamente las partes que forman un todo y revelan información sobre la variación relativa entre sus componentes. Bajo este modelo, ni el tratamiento de los datos en valor absoluto, ni el total representan interés especial.

Los datos composicionales están presentes en diferentes campos y/o disciplinas ([Cortés-Rodríguez & Sánchez-Barba, 2013](#)) como por ejemplo, en Biología para caracterizar la fauna existente en un determinado hábitat; en Economía, para identificar partidas de inversión de los recursos económicos; en Psicología, partiendo de la inteligencia global, es posible analizar cierto tipo de inteligencia y sus correspondientes variables.

En este sentido, en la década del 80, [Aitchison \(1982\)](#) propuso una metodología para el análisis de datos composicionales, trabajo que ha recibido valiosos aportes y reformulaciones de otros autores intentando afrontar las dificultades de interpretación de las magnitudes absolutas de las partes.

2.7.1 Principios del análisis

[Aitchison \(1997\)](#) revela algunos principios básicos a los que debe responder el análisis de datos composicionales. Estos principios son: invariancia por escala y coherencia subcomposicional.

Invariancia por escala: refiriéndose a que los vectores de componentes positivas proporcionales representan la misma composición. Dicho de otra forma, "cualquier función aplicada sobre datos composicionales debe poder expresarse en términos de *cocientes* entre sus partes o componentes".

Es lógico pensar que los resultados de un análisis referente a un subconjunto de las partes de una subcomposición, no dependa de cuáles fueron las partes restantes de las composiciones medidas. Esto da lugar al principio de **coherencia subcomposicional**, que exige que la información no varíe cuando se extrae una subcomposición.

El biplot para datos composicionales consiste en un análisis de componentes principales de la matriz de datos transformados, convenientemente centrados y usando la descomposición de la matriz de covarianzas.

A partir del año 2000 se promueven diversos aportes en los aspectos formales del análisis ([Aitchison et al. 2002](#); [Billheimer et al. 2001](#); [Pawlowsky-Glahn & Egozcue, 2001](#)), que han permitido un mejor tratamiento de los métodos ya propuestos.

Básicamente, el análisis de datos composicionales puede resumirse en tres pasos:

- La transformación de los datos a coordenadas de tipo log-cociente.
- El análisis estadístico (usual) de dichas coordenadas como variables reales.

- La interpretación de los modelos obtenidos en las propias coordenadas o volviendo a expresar los resultados en términos de composiciones.

2.7.2 Representación biplot en datos composicionales:

Sea $X = (x_{ij})$ una matriz de datos composicionales de dimensión $n \times D$. La *representación log-cociente centrada* clr define una nueva matriz Z_{ij}^* :

$$\begin{aligned} Z_{ij}^* &= \text{clr}(X) \\ &= \ln \left(\frac{x_1}{\sqrt[D]{\prod_{j=1}^D x_{ij}}}, \frac{x_2}{\sqrt[D]{\prod_{j=1}^D x_{ij}}}, \dots, \frac{x_D}{\sqrt[D]{\prod_{j=1}^D x_{ij}}} \right) \\ &= \ln \frac{x_{ij}}{\sqrt[D]{\prod_{j=1}^D x_{ij}}} \\ &= \underbrace{\ln(x_{ij})}_{l_{ij}} - \underbrace{\frac{1}{D} \sum_{j=1}^D \ln(x_{ij})}_{l_i} \end{aligned}$$

denotando l_{ij} l_i se tiene $Z_{ij}^* = l_{ij} - l_i$.

$$\text{por tanto } \sum_{j=1}^D Z_{ij}^* = \frac{1}{D} \sum_{j=1}^D Z_{ij}^* = 0$$

Con esto, la media de los elementos de una fila es igual a cero.

Al centrar la matriz con respecto a las medias de las columnas se obtiene la matriz:

$$\begin{aligned} Z_{ij} &= l_{ij} - l_i - \frac{1}{n} \sum_{i=1}^n (l_{ij} - l_i) \\ &= l_{ij} - l_i - l_j + l_{..} \end{aligned}$$

$$\text{en consecuencia, } \frac{1}{D} \sum_{i=1}^n Z_{ij}^* = 0$$

Considerando que Z es una matriz doblemente centrada (por filas y columnas), los vectores singulares (U y V) también están centrados. El rango de la matriz Z es $D - 1$.

Aitchison & Greenacre (2002) llamaron al biplot derivado de la matriz de datos composiciones, "*biplot de variación relativa*" en vista de que este constituye la variación en todas las relaciones de los componentes. La geometría de Aitchison es una geometría euclídea. Permitiendo utilizar todas las propiedades y herramientas de estos espacios que nos son familiares. Es decir, disponemos de productos escalares para hacer proyecciones ortogonales y de ejes ortogonales que permiten trabajar en coordenadas.

El biplot composicional muestra gráficamente la aproximación en rango dos de la matriz, dada por la descomposición en valores singulares; consta de:

- Un origen O que representa el centro del conjunto de datos composicionales,
- Un vértice en la posición h_j para cada una de las D partes, y
- Un r caso en la posición g_i para cada una de las n muestras o casos.

Se determina que la unión de O a un vértice h_j se denomina "**rayo**" y la unión de dos vértices h_j y h_k se denomina "**link**".

Éstas constituyen las características básicas de un biplot de variación relativa, con ciertas propiedades para la interpretación de la variabilidad composicional.

Cada rayo representa una variable; y su longitud, la varianza asociada explicada en la proyección.

La interpretación geométrica de un biplot para datos composicionales se da en términos de la representación de los individuos, denominado “*biplot de forma*” (ver Figura 2-7); y en base a la representación de las variables, llamada “*biplot de covarianza*” (ver Figura 2-8).

2.7.3 Propiedades Fundamentales

Las propiedades de los biplot de forma en datos composicionales son las siguientes:

- Las distancias entre los puntos fila son aproximaciones de las distancias entre los individuos, calculados a partir de la matriz de log-ratios centrados.
- La longitud de un rayo se aproxima a la calidad de representación de la variable que representa.
- La proyección de un punto fila sobre el vector columna es aproximadamente el elemento transformado de la matriz de datos de partida.

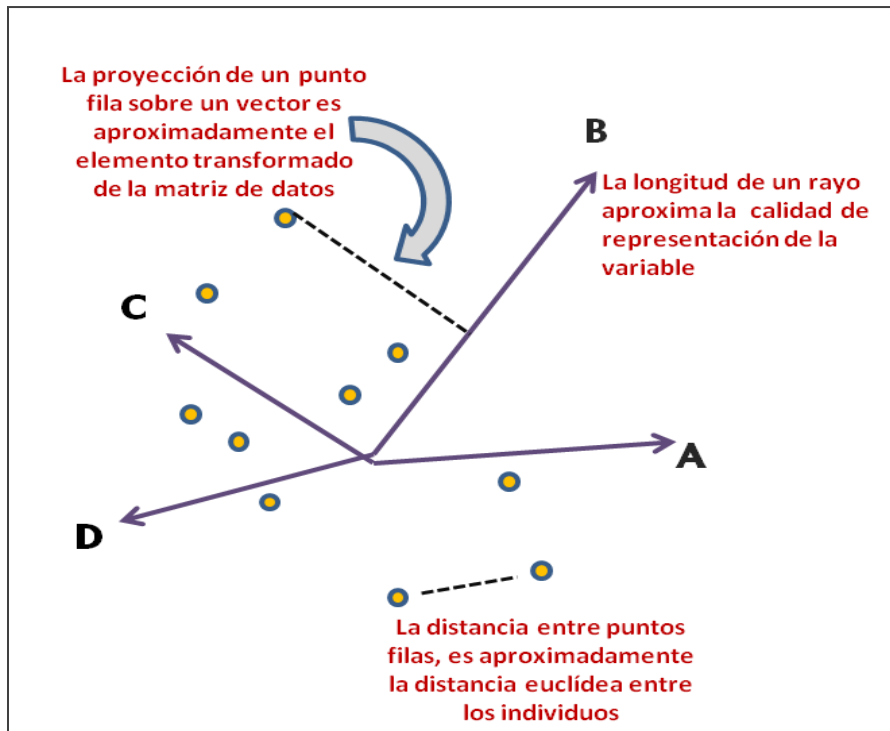


Figura 2-7 Biplot de forma

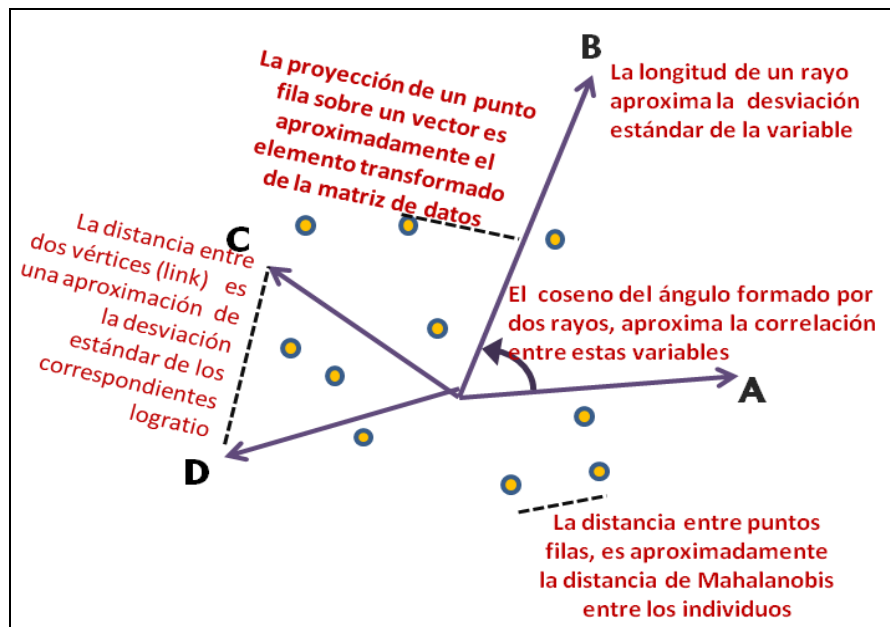


Figura 2-8 Biplot de covarianza

Las propiedades de los biplot de covarianza en datos composicionales son:

- Propiedad 1: {
 - La longitud de un rayo $\|h_j\|$ asociado a la variable Z_j es aproximadamente la desviación estándar de la variable
- Propiedad 2: {
 - La distancia entre dos vértices (link) es una aproximación de la desviación estándar de los correspondientes log-ratio
- Propiedad 3: {
 - El coseno del ángulo que forman dos rayos, se aproxima a la correlación entre las variables correspondientes
- Propiedad 4: {
 - La distancia de Aitchison entre dos rayos, coincide con la distancia entre los respectivos marcadores columna transformados
- Propiedad 5: {
 - El coseno de los ángulos del link h_j-h_k entre las variables Z_j y Z_k y el link $h_j - h_m$ entre las variables Z_j y Z_m es aproximadamente la correlación entre los correspondientes log-ratio
- Propiedad 6: {
 - Las distancias entre los puntos de fila (individuos) son aproximaciones de las distancias de Mahalanobis entre los individuos, calculados a partir de la matriz de log-ratio centrados

Las técnicas que hemos abordado hasta aquí son adecuadas si se quiere representar datos cuantitativos continuos y métricas lineales. Sin embargo, ante la presencia de información cualitativa las técnicas mencionadas pierden poder de análisis y síntesis, pero surgen otros enfoques que trabajan con información de este tipo. Básicamente, son técnicas de análisis estadístico multivariante que persiguen el mismo objetivo de los métodos biplot: la representación gráfica simultánea de una matriz de datos, pero con ciertas características particulares y que explicamos a continuación.

2.8 HJ Biplot Composicional

La metodología explicada en el punto anterior hace referencia solamente a los Biplots GH y JK para datos composicionales. Como una alternativa a estos modelos y con el propósito de facilitar un nuevo procedimiento para obtener el HJ Biplot en datos composicionales, [Hernández Suárez et al. \(2016\)](#) presentan el método denominado “HJ Biplot Composicional”.

Para aplicar el análisis estadístico, los datos composicionales en bruto se someten a una transformación “clr”. Esta transformación es simétrica con respecto a las partes y mantiene el mismo número de componentes que el número de partes en la composición. La transformación clr viene dada por la expresión:

$$clr(x) = \left[\ln \frac{x_1}{g_m(x)}, \ln \frac{x_2}{g_m(x)}, \dots, \ln \frac{x_D}{g_m(x)} \right]$$

donde $x = (x_1, x_2, \dots, x_D)$ es un vector de datos composicionales, $g_m(x)$ es la media geométrica de ese vector de datos composicionales y x_i representa las variables.

Por lo tanto, la matriz X está formada por datos transformados mediante la expresión clr, y luego es doblemente centrada a través de la DVS para asegurar que los componentes se analicen en una escala de relación en las dimensiones apropiadas.

El HJ Biplot composicional se basa en los mismos conceptos explicados arriba de vértices, rayos y marcadores relacionados con cada observación individual.

2.9 Biplot Logístico Binario

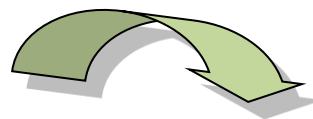
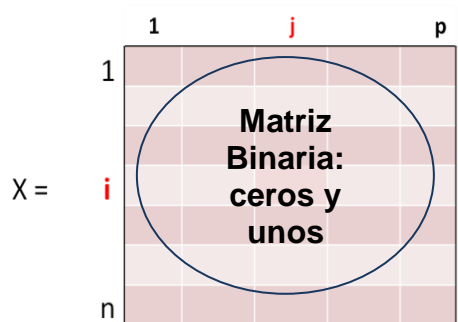
Como un procedimiento alternativo ante escenarios en los que se tienen datos binarios, Vicente-Villardón (2001) y Vicente-Villardón et al. (2006) proponen un nuevo enfoque biplot, denominado *Biplot Logístico*, basado en un modelo en el cual las coordenadas de los individuos y las variables están previstas para obtener respuestas de tipo logístico (Vicente-Villardón et al. 2004). En la formulación de esta metodología, se define una matriz de datos $X_{n \times p}$, en la cual las filas corresponden a los individuos, y las columnas miden atributos o datos cualitativos que se asocian a variables binarias (*presencia o ausencia de una determinada característica*).

2.9.1 Estructura de datos para realizar un Biplot Logístico

El biplot logístico, se corresponde con la regresión logística, persigue el mismo objetivo de los métodos biplot y puede ser formulado como:

$$\pi_{ij} = \frac{e^{b_{j0} + \sum_{s=1}^k b_{js} a_{is}}}{1 + e^{b_{j0} + \sum_{s=1}^k b_{js} a_{is}}}$$

donde, a_{is} y b_{js} ($i = 1, \dots, n; j = 1, \dots, p; s = 1, \dots, k$) son los parámetros del modelo usados como marcadores filas y columnas respectivamente, representados en el plano.



$\pi_{ij} = E(x_{ij})$ la probabilidad esperada que asocia el valor cero (0) si la característica está ausente y el valor uno (1) si la característica está presente.

La transformación logit (regresión logística), como función de vínculo, es equivalente al modelo lineal generalizado.

$$\begin{aligned} \text{logit}(\pi_{ij}) &= \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) \\ &= b_{j0} + \sum_{s=1}^k b_{js} a_{is} \\ &= b_{j0} + a'_i b_j \end{aligned}$$

donde $a_i = (a_{i1}, \dots, a_{is})'$ y $b_j = (b_{j1}, \dots, b_{js})'$

En forma matricial,

$$\underbrace{\text{logit}(\pi)} = 1b'_0 + AB'$$

donde, π es la matriz de probabilidades esperadas
 1 es un vector de unos
 b_0 es el término que contiene las constantes
 A y B son las matrices que contienen los marcadores para las filas y columnas de la matriz X

En el modelo anterior, los ejes de ordenación son considerados como variables latentes que explican la asociación entre las variables observadas.

Al suponer que los individuos responden de forma independiente a las variables, y que las variables son independientes para valores dados de rasgos latentes; la función de máximo verosimilitud está dada por:

$$\text{Prob}(x_{ij} / b_0, A, B) = \prod_{i=1}^n \prod_{j=1}^p \pi_{ij}^{x_{ij}} (1 - \pi_{ij})^{1-x_{ij}}$$

Aplicando el logaritmo a esta función, se obtiene:

$$L = \log \text{Prob}(x_{ij} / b_0, A, B) = \sum_{i=1}^n \sum_{j=1}^p [x_{ij} \log(\pi_{ij}) + (1 - x_{ij}) \log(1 - \pi_{ij})]$$

Tomando las derivadas de L con respecto a los parámetros, igualando a cero y resolviendo $3p+2n$ ecuaciones simultáneas, se obtienen las estimaciones.

Vicente-Villardón et al. (2006) fijan los parámetros del biplot logístico en un esquema iterativo que alterna A y B. Un conjunto de parámetros es introducido, mientras que el otro se mantiene fijo, y este procedimiento se repite hasta que la probabilidad converge a un grado de precisión deseado (Figura 2-9).

2.9.2 Algoritmo General



Figura 2-9 Algoritmo general, para la aplicación del Biplot Logístico

En el biplot logístico, la representación de variables e individuos no se refiere a las puntuaciones del individuo en la variable, sino a la *probabilidad* de que el individuo tenga la característica que define la variable. Por tanto, la exclusividad de su resultado se da en términos de la interpretación de “*áreas de probabilidad*” en el plano resultante (Figura 2-10).

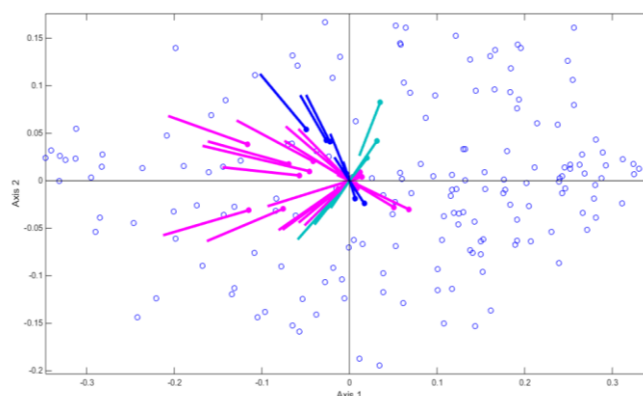


Figura 2-10 Esquema de representación del Biplot Logístico

La proyección de cada uno de los individuos sobre el segmento que representa cada variable logra la probabilidad estimada de presencia de la característica (Figura 2-11). No todas las variables estarán asociadas a la clasificación; en consecuencia, solo se suelen proyectar aquellas que presentan mejor calidad de representación.

2.9.3 Reglas de interpretación

En la representación gráfica del Biplot Logístico, hay conceptos básicos a tomar en cuenta al momento de la interpretación:

- La longitud del vector indica el poder discriminante de la variable; a menor longitud, mayor poder discriminante, y viceversa
- La dirección a la que apunta el vector indica la zona positiva del gradiente; pudiendo determinarse que variables se asocian con cada eje.
- El ángulo entre vectores indica el grado de asociación entre esas variables.
- El ángulo entre cada vector y el eje señala el grado de relación entre el vector y el eje (gradiente).

2.9.4 Calidad de Representación de las variables

Una vez obtenidos los resultados del Biplot Logístico, la calidad de representación de cada variable se puede medir con tres elementos fundamentales:

- (a) El p-valor a partir del cual determinamos qué variables son significativas en el modelo.
- (b) r^2 de Nagelkerke para conocer la capacidad explicativa del modelo logístico.

(c) El porcentaje de variables bien clasificadas obtenidas de las probabilidades esperadas, tomando de referencia un percentil de 0.5 como indicador de corte para la predicción de presencia y ausencia (menor de 0.5 indica ausencia).

2.9.5 Regiones de Predicción

Además, para cada variable es posible obtener un diagrama de ordenación, en el cual la variable es dividida en dos regiones que predicen presencia o ausencia del atributo (Vicente-Villardón, 2010). Ambas regiones quedan separadas por una línea que es perpendicular al vector que representa a la variable y corta al vector en el punto 0,5. El origen del cada vector corresponde con la probabilidad de 0.5 indicador de presencia de la característica; y el otro extremo o punta de la flecha corresponde con la probabilidad de 0.75.

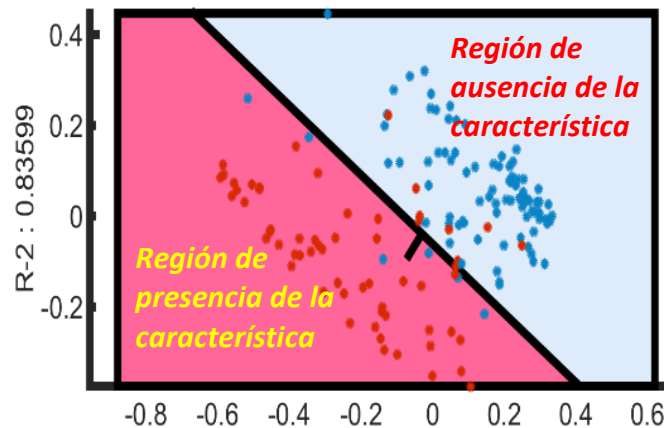


Figura 2-11 Regiones de predicción del Biplot Logístico

2.9.6 Biplot Logístico Externo

Siguiendo la línea de pensamiento de Vicente-Villardón et al. (2006), surge una variante al biplot logístico; Demey (2008) combina en un mismo algoritmo, el

análisis de coordenadas principales y la regresión logística, para construir la técnica conocida como *Biplot Logístico Externo*. La propuesta se basa en el hecho de que la regresión en el procedimiento alternativo para datos binarios no es más que una regresión logística que se puede articular convenientemente, a la configuración obtenida a partir del Análisis de Coordenadas Principales.

Se parte de una matriz $X_{n \times p}$ de datos, obtenida de observaciones en las que se registran atributos o características que se han asociado a variables binarias, donde el valor cero indica que el atributo está ausente y el 1 que está presente. Seguidamente, se define una matriz $S = (s_{ij})$ que contiene las similitudes entre filas (individuos) obtenida de la matriz de datos binarios X . A partir de estas matrices, se inicia el algoritmo con un Análisis de Coordenadas Principales (ACoP) ordenando los individuos en un espacio Euclideo de baja dimensión, de tal manera que la distancia entre cualesquiera dos puntos aproxime tanto como sea posible, la disimilitud entre individuos representados por estos puntos. Para determinar las variables asociadas a la ordenación obtenida en PCoA, buscamos las direcciones en el diagrama de ordenación que mejor predicen la probabilidad de presencia de cada variable. Una vez obtenida la ordenación, se procede con la metodología planteada por Villardón, sobre el biplot logístico.

Esta propuesta sobre el Biplot Logístico Externo, obedece al hecho de que las coordenadas de los individuos son calculadas en un *procedimiento externo* (PCoA). En un *Biplot Logístico Externo*, se genera un gráfico, en el cual los individuos son representados como puntos, y los coeficientes de regresión

como vectores, determinando la dirección de los ejes. A continuación, se presentan los pasos a seguir para su construcción (Figura 2-12).

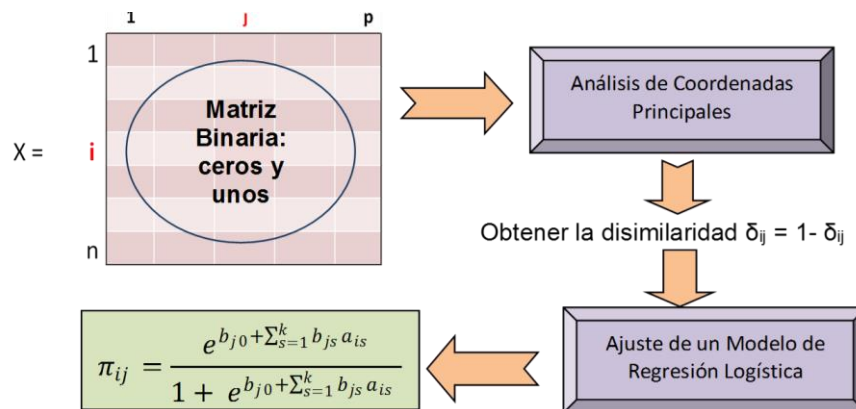


Figura 2-12 Algoritmo para la aplicación del Biplot Logístico Externo

Los resultados del biplot logístico externo son utilizados, además, para evaluar la separación y formación de grupos, en función del análisis de cluster.

2.10 Biplot Logístico Nominal

Recientemente [Hernández Sánchez \(2016\)](#) amplía el concepto del biplot logístico, desarrollando un método para el análisis de datos categóricos, denominado *Biplot Logístico Nominal* (BLN). Bajo este enfoque se representan los datos en un espacio de dimensión reducida y se explica la correlación entre las variables nominales, a través de regiones de predicción divididas en categorías. La capacidad del análisis del BLN permite hacer una valoración de los resultados en términos de distancias, de forma tal que la categoría que se predice en una variable cualquiera para cada individuo va a ser la más cercana a éste.

En la formulación de esta metodología, se define una matriz de datos $X_{I \times J}$, en la cual las filas corresponden a los I individuos, y las columnas a las J variables

nominales, cada una asociada a K_j ($j = 1, \dots, J$) categorías. Además, se define una matriz indicadora $\mathbf{G}_{I \times L}$ en la cual $L = \sum_j K_j$ columnas.

Sea $\pi_{ij}(k) = E(x_{ij})$ la probabilidad esperada que asocia la categoría k (de la variable j) al individuo, si efectivamente está presente. En la construcción del modelo logístico multinomial de respuesta latente con S rasgos latentes, la probabilidad pueden ser formulada como:

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

Para cada variable, se puede usar como categoría base, la primera o la última categoría. Tomando como base la última categoría, y asumiendo que en el modelo logístico multinomial de respuesta latente el logaritmo de los odds para cada respuesta sigue un modelo lineal, se tiene que:

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

donde, a_{is} y $b_{j(k)s}$ son los parámetros del modelo

$$i = 1, \dots, I; j = 1, \dots, J; k = 1, \dots, K_j - 1; s = 1, \dots, S$$

Matricialmente la expresión: $\mathbf{O} = \mathbf{1}_I \mathbf{b}'_0 + \mathbf{A} \mathbf{B}'$ define el biplot para los odds, donde $\mathbf{O}_{I \times (L-J)}$ es la matriz que recoge los odds esperados. Por lo tanto, las

probabilidades predichas (para cada variable) son interpretadas como “regiones de predicción” o “puntos categoría”.

En el biplot logístico nominal, las regiones de predicción son polígonos convexos que dividen el espacio de representación en tantas regiones como categorías tenga la variable (teselación) y configurada por un Diagrama de Voronoi (ver *Figura 2-13*). De esta manera, cada región convexa predice una categoría, siendo la probabilidad para cada categoría tan alta como la probabilidad asignada al resto de las categorías.

La representación obtenida se interpreta en términos de distancias, en el sentido de que la categoría predicha para cada individuo está definida por los *puntos de la categoría* más cercana.

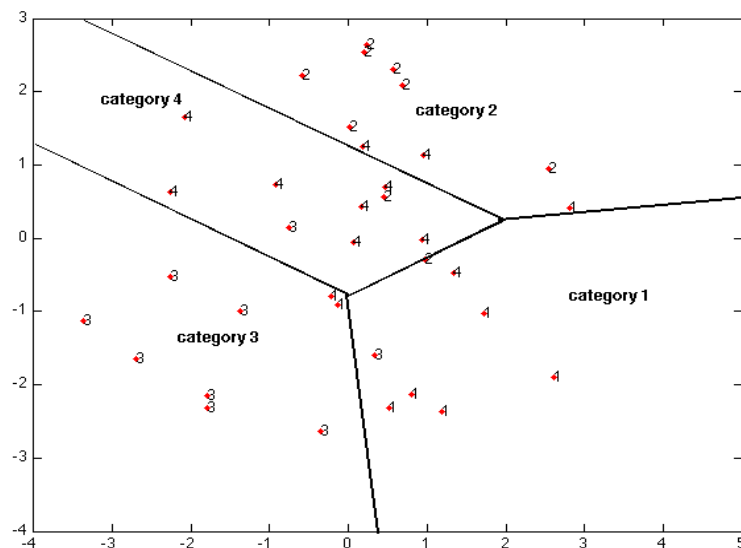


Figura 2-13 Teselación definida por las regiones de predicción.

Tomado de [Hernández Sánchez \(2016\)](#), Tesis doctoral, Universidad de Salamanca

2.11 Biplot Logístico Ordinal

Simultáneamente al modelo anterior, [Hernández Sánchez \(2016\)](#) amplía el concepto del biplot logístico al estudio de variables que conllevan un orden en sus categorías, denominándole Biplot Logístico Ordinal (BLO).

En la formulación de esta metodología, se define una matriz de datos X_{Ixj} , en la cual las filas corresponden a los I individuos, y las columnas a las J variables ordinales, cada una asociada a K_j ($j = 1, \dots, J$) categorías ordenadas. Además, se define una matriz indicadora P_{IxL} con $L = \sum_j K_j$ columnas para cada variable categórica P_j ($P = P_1, \dots, P_j$). P representa la probabilidad asociada a cada categoría de cada una de las variables.

Sea $\pi_{ij(k)}^* = P(x_{ij} \leq k)$ la probabilidad esperada de que el individuo i asuma un valor menor o igual que k en la j -ésima variable ordinal.

Sea $\pi_{ij(k)} = P(x_{ij} = k)$ la probabilidad esperada de que el individuo i tome el k -ésimo valor de la j -ésima variable ordinal

En la construcción del modelo logístico multinomial de respuesta latente, la probabilidad puede ser formulada como:

$$\begin{aligned} \pi_{ij(k)}^* &= \frac{1}{1 + e^{-(d_{jk} + \sum_{s=1}^S a_{is} b_{js})}} \\ &= \frac{1}{1 + e^{-(d_{jk} + a_i' b_j)}} \end{aligned}$$

donde,

$a_{is} = (a_{i1}, \dots, a_{iS})'$ es el vector de puntuaciones de la respuesta latente para el i -ésimo individuo

$b_j = (b_{j1}, \dots, b_{jS})'$ es el parámetro para cada variable

d_{jk} es el parámetro para cada categoría

Las ecuaciones definen un biplot en la escala logit siguiendo la geometría del modelo binario, uno para cada categoría. Las puntuaciones a_i se usan para identificar grupos con características similares; y, los parámetros b_j señalan las direcciones que predicen las probabilidades de las categorías y ayudan a encontrar las variables encargadas de las diferencias entre los individuos.

El subespacio de representación queda dividido en regiones de predicción para cada categoría, delimitadas por líneas rectas paralelas (ver *Figura 2-14*).

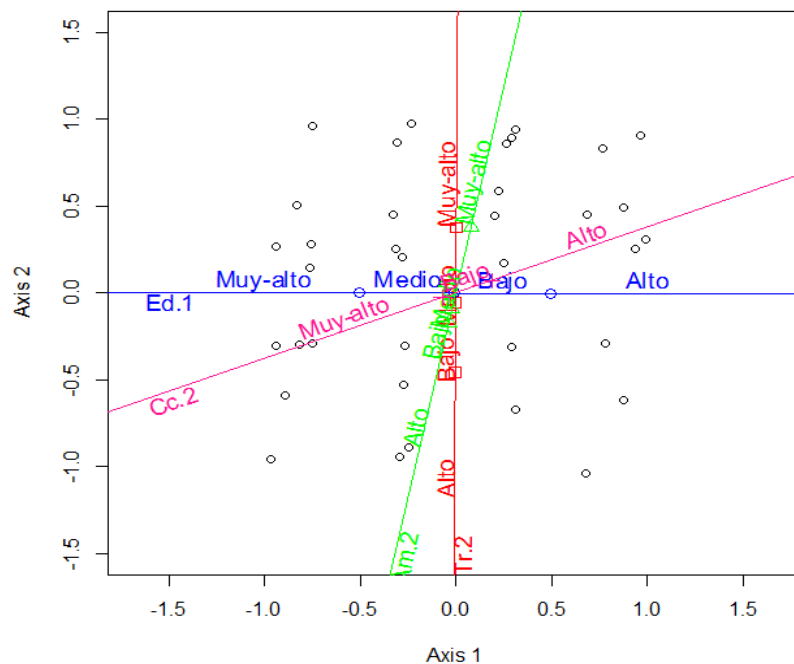


Figura 2-14 Regiones de predicción para una variable ordinal

2.12 Versión Inferencial del Biplot

Como hemos visto a lo largo de esta revisión bibliográfica, los métodos biplot constituyen una importante herramienta de representación, para explicar la variación conjunta de variables e individuos en un espacio bidimensional. La interpretación del biplot se hace a partir de los marcadores filas y marcadores columna, que definen *parámetros de estimación* representados como *puntos* y *vectores* en el gráfico. De esta manera, los parámetros en el biplot se estiman como valores puntuales, pero no aportan información acerca de la *precisión* de los estimadores. Es por ello que nace la idea de una versión inferencial de los métodos biplot, avalada por Nieto-Librero (2015) que se basa en el método *Bootstrap* (Efron, 1979). Para realizar el análisis Biplot de forma inferencial Nieto-Librero & Galindo-Villardón (2015), implementaron una interfaz gráfica de usuario en entorno R, que detallamos en el siguiente capítulo.

El *Bootstrap* sugiere que, a partir de una muestra original de la población, se generen sucesivamente nuevas muestras aleatorias y se lleve a cabo la inferencia con los resultados de las muestras obtenidas. En el proceso de remuestreo se van seleccionando las submuestras mediante muestreo con reposición; este procedimiento hace suponer que la muestra considerada en sí misma contiene la información básica de la población. Bootstrap es un método de remuestreo utilizado con frecuencia para aproximar el error estándar o la varianza de un estimador, y construir así intervalos de confianza o realizar contrastes de hipótesis sobre parámetros de interés. En lugar de fórmulas o modelos matemáticos abstractos, el bootstrap simplemente requiere un

ordenador capaz de simular un proceso de muestreo aleatorio de los datos. Afortunadamente, la potencia de los ordenadores actuales facilita considerablemente la aplicabilidad de este costoso método.

La expresión *bootstrap* puede referirse a *bootstrap* paramétrico o *bootstrap* no paramétrico. El *bootstrap* no paramétrico no es rígido respecto al cumplimiento de supuestos teóricos, en tal sentido es menos *restrictivo* que las técnicas convencionales. En su aplicación, en vez de asumir a priori una determinada distribución teórica, se utiliza la muestra original y se generan un gran número de sub-muestras que sirven de base para estimar la distribución muestral de los datos. De esta manera, pueden analizarse datos provenientes de distribuciones desconocidas o incluso abordarse situaciones, frente a las cuales no hay una solución analítica conocida (Efron & Tibshirani, 1994).

2.12.1 El Principio “Plug-In”

El “Principio “Plug-In” es un método simple de estimación de parámetros a partir de la muestra. La notación $\theta = \theta(F)$ donde F es la función de distribución de probabilidad, es un indicador de que el valor θ del parámetro es obtenido aplicando algún procedimiento numérico $\theta(\cdot)$ a la distribución F .

El “estimador plug-in” del parámetro $\theta = \theta(F)$ estaría definido por:

$$\hat{\theta} = \theta(\hat{F})$$

Esto quiere decir que se estima la función $\theta = \theta(F)$ de la distribución de probabilidad F , a partir de la misma función de la distribución empírica \hat{F} , $\hat{\theta} = \theta(\hat{F})$.

Bajo este razonamiento, un estimador plug-in de la esperanza $\theta = E_F(x)$ es:

$$\hat{\theta} = E_{\hat{F}}(x) = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}$$

El Jackknife y el Bootstrap son dos métodos utilizados para estimar o aproximar la distribución muestral de un estadístico y sus características.

2.12.2 La Muestra Bootstrap

Dada una muestra aleatoria

$x = x_1, x_2, \dots, x_n$ que proviene de una población desconocida.

$x^* = x_1^* + x_2^* + \dots + x_n^*$ es una muestra aleatoria bootstrap, con reemplazo

Si definimos θ un parámetro de interés a estimar,

entonces $\hat{\theta} = s(x)$ es el estimador del parámetro θ .

De la misma forma, entonces podemos definir un estimador para un parámetro de x^*

$$\hat{\theta}^* = s(x^*)$$

En la teoría estadística, uno de los errores estadísticos de gran relevancia, es el error estándar o desviación estándar de la distribución muestral.

La estimación bootstrap de este error estadístico se obtiene a partir de los siguientes pasos:

Paso 1: Seleccionar B muestras bootstrap independientes $x_1^*, x_2^*, \dots, x_B^*$ escogidas con reemplazo.

Paso 2: Calcular el estimador ($\hat{\theta}$) bootstrap del error estándar $\hat{\theta}_b^* = s_b(x^*)$ donde $b = 1, 2, \dots, B$

Paso 3: Estimar el error estándar de $\hat{\theta}$ con la desviación estándar de la muestra de las réplicas bootstrap, esto es,

$$\begin{aligned}\widehat{se}_B &= \widehat{SE}[s(X)] \\ &= \left(\frac{1}{B-1} \sum_{b=1}^B (s_b(x^*) - s(x^*))^2 \right)^{1/2}\end{aligned}$$

donde $s(x^*) = \frac{\sum_{b=1}^B s_b(x^*)}{B}$

También es posible obtener estimaciones del sesgo del estadístico de interés. Recordemos que el sesgo de un estimador es la diferencia entre el valor estimado y el valor verdadero.

De igual manera, el sesgo de una muestra bootstrap, se puede obtener mediante la diferencia entre la media del estimador y el valor verdadero del parámetro, para lo cual debemos obtener una muestra aleatoria bootstrap con reemplazo.

Supongamos que $\hat{\theta}$ es la estimación del estadístico sobre los datos de la muestra, la estimación bootstrap del sesgo sería:

$$\widehat{sesgo}_b(\hat{\theta}) = \hat{\theta}^* - \hat{\theta} \quad \text{donde } \hat{\theta}^* = \frac{\sum_{b=1}^B s_b(x^*)}{B}$$

2.12.3 Estimador Jackknife

Jackknife es un método de estimación no paramétrico que se usa para estimar medidas de variabilidad como el sesgo, error estándar y la varianza. Jackknife es también llamado “leave-one-out” y fue introducido por [Quenouille \(1950\)](#) y mejorado después por [Tukey \(1958\)](#).

Este método recalcula el valor del estadístico de interés en cada una de las muestras denominadas muestras Jackknife, de tamaño $(n - 1)$ obtenidas a

partir de la muestra conocida $x = (x_1, x_2, \dots, x_n)$, en la siguiente forma (Figura 2-15).

- La primera muestra Jackknife está formada por todas las observaciones de la muestra original, excepto la primera.
- La segunda muestra Jackknife está formada por todas las observaciones de la muestra original, excepto la segunda.
- La i –ésima muestra está formada por todas las observaciones de la muestra original, excepto la i –ésima muestra; así sucesivamente.

Bajo este procedimiento, Jackknife es un método que incorpora todas las muestras extraídas ($n - 1$), sin reposición, de la muestra original (Efron & Gong, 1983). Con la muestra original, estimamos $\hat{\theta} = s(x)$ siendo $s(x)$ la estadística de interés.

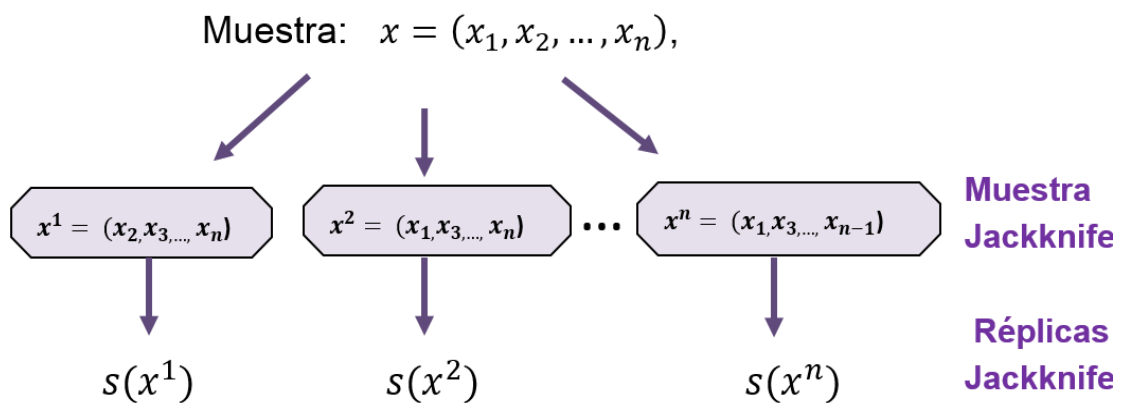


Figura 2-15 Metodología Jackknife

Considerando, $\hat{\theta}_{(.)} = \frac{1}{n} \sum_{i=1}^n \hat{\theta}_{(i)}$ y $\hat{\theta}_{(i)} = s(x_{(i)})$

El estimador Jackknife del sesgo se define como:

$$s\widehat{esg}o_{jackk} = (n - 1)(\widehat{\theta}_{(.)} - \widehat{\theta})$$

El estimador Jackknife del error estándar se define como:

$$\widehat{se}_{jackk} = \left[\frac{n-1}{n} \sum (\widehat{\theta}_{(i)} - \widehat{\theta})^2 \right]^{1/2}$$

2.12.4 Intervalos de confianza Bootstrap

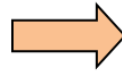
Tradicionalmente, el estadístico aplicado en la construcción de intervalos de confianza impone determinados supuestos sobre la variable aleatoria analizada; obtiene una estimación del estadístico de interés y consigue una estimación del error estándar de la distribución muestral del estadístico. Fijado un nivel de confianza $1 - \alpha$, el intervalo de confianza (IC) se obtiene mediante la expresión de la forma: $\widehat{\theta} \pm Z_{\alpha/2} \cdot \widehat{\sigma}_{\widehat{\theta}}$ donde $\widehat{\sigma}_{\widehat{\theta}}$ es la estimación del error estándar de la distribución muestral del estadístico y $Z_{\alpha/2}$ se corresponde con los percentiles asociados al nivel de confianza establecido.

El mayor aporte de la estrategia bootstrap, en el marco de la construcción de intervalos de confianza, consiste en la posibilidad de disponer de un procedimiento que solventa aquellas dificultades que surgen si se desconoce la distribución muestral de los estadísticos. Para la construcción de intervalos de confianza, de muestras grandes ($n \geq 20$), vamos a suponer que se tiene un estimador $\widehat{\theta}$ (ver *Figura 2-16*).

Estimador $\hat{\theta}$

$$\hat{\theta} \sim N(\theta, \text{Var}(\hat{\theta}))$$

de media desconocida
y varianza conocida.



$$Z = \frac{(\hat{\theta} - \theta)}{\text{se}} \sim N(0, 1)$$



Con lo cual podemos definir
los intervalos de confianza como:

$$[\hat{\theta} - z_{1-\alpha}\text{se}(\hat{\theta}), \hat{\theta} + z_{1-\alpha}\text{se}(\hat{\theta})]$$

Figura 2-16 Intervalos de Confianza Bootstrap para muestras grandes

Estos intervalos tienen una probabilidad $1 - \alpha$ de contener el verdadero valor de θ .

Para obtener los intervalos de confianza de muestras pequeñas (Figura 2-17), utilizamos la aproximación t de Student conocida:

$$\hat{\theta} \sim N(\theta, \text{Var}(\hat{\theta}))$$



$$Z = \frac{(\hat{\theta} - \theta)}{\text{se}} \sim t_{n-1}$$



obteniéndose los intervalos
de confianza en la forma

$$[\hat{\theta} - t_{(1-\alpha, n-1)}\text{se}(\hat{\theta}), \hat{\theta} + t_{(1-\alpha, n-1)}\text{se}(\hat{\theta})]$$

Figura 2-17 Intervalos de confianza Bootstrap para muestras pequeñas

Capítulo 3 **SOFTWARE** **SOBRE BIPLLOT**

3.1 Introducción

Si bien es cierto que los programas estadísticos comerciales genéricos son capaces de realizar análisis biplot, el coste total como herramienta analítica suele ser bastante elevado. Además, en algunos casos el uso de cada una de sus facultades queda limitada, debido a que son paquetes que se ofrecen separadamente. Sin embargo, el lenguaje R, a pesar de que inicialmente pueda tener una curva de aprendizaje exigente, es un lenguaje libre, abierto y gratuito, con una potencia analítica y una continua expansión en aplicaciones de todo tipo, que podemos adaptar a nuestras necesidades.

El lenguaje de programación R se ha convertido prácticamente en un referente internacional en el ámbito estadístico. Su amplia variedad de librerías orientadas al trabajo estadístico ha permitido que tanto investigadores como académicos y una gran comunidad de usuarios, adopten esta herramienta para cubrir sus necesidades de ajuste y modelado estadístico para pequeños conjuntos de datos, o para gestionar grandes volúmenes de datos que requieran métodos estadísticos de alto rendimiento.

Por otro lado, junto a la amplia gama de librerías, se ha desarrollado el entorno gráfico conocido como Interfaz Gráfica de Usuario (GUI), que permite trabajar con R, cuya principal virtud es que, al estar hecha a base de menús, facilita el manejo del entorno.

Evidentemente el lenguaje R posee una variedad de herramientas para el análisis de datos multivariantes. Como este trabajo está orientado al estudio de los métodos Biplot haremos una revisión de las librerías que dentro de sus funciones abordan las técnicas del biplot para el caso de datos de *dos dimensiones* (vías). Además de mostrar de forma resumida sus capacidades e ilustrar el uso de las funciones disponibles; en cada caso se ilustra con un ejemplo y finalmente se hace una comparativa de los resultados obtenidos, de forma general. El objetivo final es aplicar correctamente los diferentes algoritmos a datos reales e interpretar sus resultados.

3.2 Librerías en R para la construcción del Biplot

La búsqueda de ese objetivo se inicia con una revisión de los paquetes en entorno R (R-TEAM, 2014) que tienen implementada la descomposición y/o representación Biplot. En la Tabla 2 se ha recogido información de cada uno de ellos con su respectivo nombre y una breve descripción.

Tabla 2: Paquetes o librerías que realizan a descomposición Biplot

Librería-Referencia	Fecha de creación y de modificación	Descripción
Biplot {stats} (R-TEAM, 2014)	25-09-2013	Realiza el Análisis de Componentes Principales y devuelve el biplot de Gabriel
Calibrate (Graffelman, 2012)	21-01-2006 10-09-2013	Permite calibrar los ejes en el biplot y en el diagrama de dispersión, dibujando marcas a lo largo del vector.
bPCA (Faria & Demetrio, 2012)	17-08-2008 11-06-2018	Implementa la construcción de biplots en dos (2D) y tres (3D) dimensiones. Utiliza

		4 métodos de factorización: HJ, JK, GH y sqrt.
multbiplotR (Vicente-Villardón, 2017)	13-01-2015	Es la traducción a R del paquete “MultBiplot” desarrollado previamente en Matlab. Ejecuta diferentes tipos de biplot
NominalLogistic Biplot (Hernández & Vicente-Villardón, 2013a)	01-05-2014	Construye el biplot para datos nominales, en cuyo caso las variables deben ser politómicas; es decir, deben tener más de dos categorías.
OrdinalLogisticBiplot (Hernández & Vicente-Villardón, 2013b)	30-10-2013 16-01-2015	Construye el biplot en base a la existencia de un orden en los valores de las categorías de las variables en estudio.

A continuación, se explica más detalladamente el contenido y funcionalidad de cada librería, resaltando principalmente los argumentos que requiere para poner en práctica el análisis biplot.

3.2.1 Paquete Stats

Este paquete está instalado con el módulo base de R; es un método para la función genérica biplot. Produce un biplot a partir de resultados previos obtenidos de un ACP sobre la matriz de datos. El argumento que usa es:

```
biplot(x, choices=1:2, cex)
```

donde,

x Un objeto de la clase “princomp” o “prcomp”.

La función princomp() utiliza la descomposición espectral.

La función prcomp() usa la descomposición en valores singulares.

El formato simplificado de estas dos funciones es:

```
prcomp(x, scale = FALSE)
```

```
princomp(x, cor = FALSE, scores = TRUE)
```

Argumentos para `prcomp()` :

- `x`: una matriz numérica o un “data frame”
- `scale`: un valor lógico indicando si las variables deben escalarse antes de que tenga lugar el análisis

Argumentos para `princomp()` :

- `x`: una matriz numérica o un “data frame”
- `cor`: un valor lógico. Si es “TRUE”, los datos serán centrados y escalados antes del análisis

choices Vector de longitud 2, que especifica las componentes a graficar.

cex Factor de expansión de caracteres utilizado para etiquetar los puntos fila. Las etiquetas pueden ser de diferentes tamaños para filas y columnas al proporcionar un vector de longitud dos.

3.2.2 Paquete “calibrate”

Paquete para dibujar escalas calibradas con marcas graduadas, sobre los vectores variables en el biplot de correlación. Generalmente los trabajos que utilizan el biplot no suelen mostrar ejes calibrados¹, porque hay pocos softwares disponibles para hacerlo. Probablemente el primer ejemplo de un biplot con ejes calibrados ha sido el de [Gabriel & Odoroff \(1990\)](#), que hace referencia a los vectores columna calibrados como ejes biplot. Posteriormente,

¹ Las “p” variables están representadas por ejes no ortogonales, conocidos como *ejes biplot*.

[Greenacre \(1993\)](#) aborda el tema de la calibración biplot en el contexto del análisis de correspondencia, y [Gower & Hand \(1996\)](#), con más detalle, en base al análisis de componentes principales.

El paquete “calibrate” presentado por [Graffelman \(2012\)](#) desarrolla un procedimiento de calibración que consiste en dibujar una escala (lineal) a lo largo de un *eje* en una gráfica, con etiquetas numéricas y marcas de graduación.

Los ejes biplot no se refieren a los ejes perpendiculares del sistema de coordenadas, sino a las variables del biplot que se está calibrando; por lo tanto, la calibración que se produce es un biplot, con “*variables como ejes calibrados*”. Los ejes pueden ser calibrados realizando el análisis de componentes principales a partir de la matriz de covarianzas o de correlaciones.

La rutina del paquete para calibrar cualquier eje o vector en un biplot es la siguiente:

```
calibrate (g, yc, ticklab,Fp[,1:2], ticklabc,tl=0.5,axislab,cex.axislab,where,  
labpos)
```

donde,

- g** Un objeto de la clase “princomp”
- yc** Cargas factoriales del vector a ser calibrado
- ticklab** Marcas de las etiquetas del vector calibrado (función seq (from, to, by))
- Fp** Componentes principales (scores)
- ticklabc** Vector de marcas centrado
- tl** Longitud de las marcas del vector calibrado, por defecto toma el valor 0.05

- axislab** Etiqueta para el vector calibrado
- cex.axislab** Factor de expansión de caracteres para etiquetas de marca de eje
- where** Posición de la etiqueta (1=principio, 2=medio, 3=final)
- labpos** Posición de la etiqueta para el eje calibrado (1=debajo, 2=izquierda, 3=encima, 4=derecha)

El procedimiento de calibración se basa en el hecho de que los marcadores fila y marcadores columna para una matriz de datos en un biplot se pueden interpretar como coeficientes de regresión.

Se destacan dos puntos importantes de este paquete: primero, sirve para calibrar ejes obtenidos por diversos métodos multivariantes (*ACP, análisis de correspondencia, de correlación canónica o de redundancia*), manteniendo las operaciones de centrado y estandarización; segundo, se puede producir cualquier calibración lineal debido al hecho de que se puede especificar un factor de calibración. El biplot calibrado facilita su lectura e interpretación.

3.2.3 Paquete “bpca”

Este paquete dibuja los biplot en dos y tres dimensiones basados en ACP; además, ofrece la opción de interactuar en el gráfico 3D, rotándolo y ampliándolo.

La función interna plot (`bpca (X, d, center, scale, method, iec, var.rb, var.rd, limit)`) realiza la representación Biplot, donde:

- X** Una matriz de datos u objeto *prcomp* (basado en una matriz de correlación)

- d** Un vector que proporciona el primer y último valor propio a ser considerado por la reducción biplot. Puede ser $d=1:3$ o para un biplot 3D. El valor predeterminado es $d=1:2$.
- center** Valor numérico que indica el tipo de centrado a realizar: 0=sin centrado, 1=centrado global, 2=centrado por columnas; 3=doble centrado
- scale** Valor lógico que indica si las variables deben escalarse antes de que se realice el análisis: FALSE: sin escala, TRUE=con escala
- method** Un vector de cadena de caracteres que indica el método de factorización: "hj": HJ (simétrico)
 "sqrt": SQRT (raíz cuadrada simétrica)
 "jk": JK (preserva la métrica de las filas)
 "gh": GH (preserva la de las columnas)
- iec** Valor lógico, si $iec=TRUE$ la matriz de valores propios, las coordenadas de los objetos y las variables serán invertidas
- var.rb** Un valor lógico, si $var.rb=TRUE$ los coeficientes de correlación para todas las variables serán calculados
- var.rd** Un valor lógico, si $var.rd=TRUE$ el diagnóstico de la representación de variables proyectadas por el biplot

La salida de la función **bpca** devuelve un objeto de la clase `bpca.2d` o `bpca.3d`.

En cualquier caso, se puede explorar el objeto "bp" creado por la función "bpca" y generar una lista de objetos, que pueden resumirse con la función "summary(bp)":

- call** X
- eigenvalues** Un vector de valores propios
- eigenvectors** Un vector de vectores propios
- Number** Un vector del número de valores propios considerados en la reducción
- Importance** La varianza explicada general y parcial

Coord	Una lista con las coordenadas de las dos componentes: individuos y variables
var.rb	Una matriz de los coeficientes de correlación para todas las variables
var.rd	Una matriz del diagnóstico de la débil proyección de las correlaciones de las variables por la reducción biplot.

3.2.4 Paquete “MultBiplotR”

Es la traducción en R del paquete previo MultBiplot, desarrollado en Matlab. Proporciona varias técnicas multivariantes desde una perspectiva biplot.

La función `HJ.Biplot (X, dimension=3, Scaling=4, sup.rows = NULL, sup.cols = NULL)` ejecuta un HJ Biplot directamente.

Los argumentos que utiliza son:

X	Matriz de datos
dimension	Dimensión de la solución
scaling	Transformación de los datos originales (ver detalle de las posibles transformaciones al final de este apartado)
sup.rows	Filas suplementarias o ilustrativas, si las hay
sup.cols	Columnas suplementarias o ilustrativas, si las hay

La siguiente función permite realizar los biplots clásicos:

`PCA.Biplot (X, alpha=1, dimension=3, Scaling=4, sup.rows=NULL, sup.cols=NULL)`

donde,

X	Matriz de datos
alpha	Un número entre 0 and 1. 0 para GH-Biplot, 1 para JK-Biplot and 0.5 para SQRT Biplot. Para el HJ-Biplot se puede utilizar 2 o cualquier otro valor que no esté en el intervalo [0,1].
dimensión	Dimensión de la solución

Scaling Transformación de los datos originales.

Las transformaciones disponibles en el programa son las siguientes:

1="Raw Data":	Cuando no se requiere transformación.
2="Substract the global mean":	Elimina un efecto común a todas las observaciones
3="Double centering":	Residuales de la interacción. Útil para modelos AMMI
4="Column centering":	Sustrae la media de las columnas
5="Standardize columns":	Sustrae las medias de las columnas y divide por su desviación estándar
6="Row centering":	Sustrae la media de las filas
7="Standardize rows":	Divide cada fila por su desviación estándar.
8="Divide by the column means and center":	La dispersión resultante es el coeficiente de variación
9="Normalized residuals from independence"	Se usa para tablas de contingencia

La transformación se puede proporcionar a la función utilizando la cadena de caracteres entre comillas o solo indicando con el número asociado.

3.2.5 Paquete "NominalLogisticBiplot (NBL)"

La función principal de este paquete es el cálculo y la representación de un biplot para un conjunto de variables categóricas de tipo nominal. La sintaxis es la siguiente:

```
Plot (x, planex, planey, QuitNotPredicted, ReestimateInFocusPlane, proofMode, AtLeastR2, xlimi, xlimu, ylimi, ylimu, linesVoronoi, ShowAxis, PlotVars, PlotInd, LabelVar, LabelInd, CexInd, CexVar, ColorInd, ColorVar, SmartLabels, PchInd, PchVar, LabelValuesVar, ShowResults)
```

Los parámetros relacionados con esta sintaxis son los siguientes:

x	un objeto de la clase nominal.logistic.biplot.
planex	Dimensión para el eje X
planey	Dimensión para el eje y
QuitNotPredicted	Consulta si deben representarse las categorías no predichas en el gráfico? Por defecto es "TRUE".
ReestimateInFocusPlane	Estima los parámetros de las variables usando las dimensiones de la gráfica o bien utilizando aquellos almacenados en el objeto que se pasa como primer parámetro de la función. Si es "FALSO", los valores de los parámetros para otras dimensiones se establecen en 0. El valor predeterminado es "FALSO".
proofMode	Establece si cada variable se debe dibujar en una ventana separada o todas conjuntamente. Si es "FALSO" se realiza un solo gráfico con una leyenda para identificar cada variable.
AtLeastR2	Establece el valor de corte para trazar una variable atendiendo a su valor de Nagelkerke R ² .
xlimi, xlimu	Valor mínimo y valor máximo en el eje x
ylimi, ylimu	Valor mínimo y valor máximo en el eje y
linesVoronoi	Indica si las teselaciones para cada variable se deben dibujar. El valor predeterminado es FALSO y solo se pintan los puntos de la categoría para una mejor lectura del gráfico.
ShowAxis	Si debe o no mostrarse el eje. "TRUE" o "FALSE"
PlotVars	Si deben o no trazarse las variables. "TRUE" o "FALSE"
PlotInd	Si deben o no mostrarse los individuos. "TRUE" o "FALSE"
LabelVar	Si deben o no mostrarse las etiquetas de las variables.
LabelInd	Si deben o no mostrarse las etiquetas de los individuos
CexInd	Tamaño de los puntos individuos
CexVar	Tamaño de los puntos de las categorías
ColorInd	Color de los puntos de los individuos

ColorVar	Color de las variables
SmartLabels	Si deben imprimirse las etiquetas de texto de acuerdo con su posición en el gráfico
PchInd	Símbolo para los individuos
PchVar	Símbolo para las variables
LabelValuesVar	Lista con las etiquetas de texto para todas las variables
ShowResults	Si se muestran o no los resultados del proceso de cálculo de las regiones de predicción

3.2.6 Paquete “OrdinalLogisticBiplot (OLB)”

El paquete se utiliza para analizar y representar una matriz de elementos politómicos ordenados. La función principal de este paquete es `OrdinalLogisticBiplot`, cuya sintaxis es la siguiente:

```
Plot(x, planex, planey, AtLeastR2, xlimi, xlimu, ylimi, ylimu, margin, ShowAxis,
     PlotVars, PlotInd, LabelVar, LabelInd, CexInd, CexVar, ColorInd,
     ColorVar, PchInd, PchVar, showIIC, iicxi, iicxu, legendPlot, PlotClus,
     Clusters, chulls, centers, colorCluster, ConfidentLevel, addToExistingPlot)
```

Los parámetros relacionados con esta sintaxis son los siguientes:

x	Un objeto de la clase <code>ordinal.logistic.biplot</code> .
planex, planey	Dimensión para el eje “x” y para el eje “y” respectivamente
AtLeastR2	Establece el valor de corte para trazar una variable atendiendo a su valor de Nagelkerke. Se grafica una variable si su R^2 es mayor que este valor
xlimi	Valor mínimo en el eje x
xlimu	valor máximo en el eje x
ylimi	Valor mínimo en el eje y
ylimu	Valor máximo en el eje y

margin	Establece el espacio entre las variables trazadas y el borde de la ventana
ShowAxis	Si debe o no mostrarse el eje. "TRUE" o "FALSE"
PlotVars	Si deben o no trazarse las variables. "TRUE" o "FALSE"
PlotInd	Si deben o no mostrarse los individuos. "TRUE" o "FALSE"
LabelVar	Si deben o no mostrarse las etiquetas de las variables.
LabelInd	Si deben o no mostrarse las etiquetas de los individuos
CexInd	Tamaño de los puntos individuos
CexVar	Tamaño de los puntos de las categorías
ColorInd	Color de los puntos de los individuos
ColorVar	Color de las variables
PchInd	Símbolo para los individuos
PchVar	Símbolo para las variables
showLIC	Parámetro booleano para decidir si el usuario desea ver las curvas de información del elemento para cada variable.
licxi, iixcu	Límite inferior y superior respectivamente del eje X al trazar las curvas de información de las variables
legendPlot	Parámetro booleano para mostrar la leyenda de la gráfica. El valor predeterminado es "FALSO".
PlotClus	Parámetro booleano para mostrar los cluster estudiados. El valor predeterminado es "FALSO"
Clusters	Cluster asociado a cada variable. Por defecto es "FALSO".
chulls	Parámetro booleano para especificar si las figuras de los "convex hulls" serán trazadas
centers	Parámetro booleano para trazar los centros de cada cluster. El valor predeterminado es NULL.
colorCluster	Color para cada agrupación. Puede ser una matriz con la información de color para cada grupo. Valor por defecto: NULL
ConfidentLevel	Valor entre 0 y 1 para evitar valores extremos en el gráfico. El valor predeterminado es NULL.

addToExistingPlot Parámetro booleano para decidir si las variables trazadas se agregarán a un gráfico existente o no.

3.2.7 Paquete “BiplotForCategoricalVariables”

Paquete para la representación Biplot de datos categóricos mixtos que combina en un solo algoritmo, variables nominales y ordinales. Para su ejecución depende de los paquetes anteriores (NominalLogisticBiplot y OrdinalLogisticBiplot).

Para instalarlo se siguen los siguientes comandos dentro de R:

```
install.packages(c("NominalLogisticBiplot", "OrdinalLogisticBiplot",  
"mirt", "stats4", "lattice", "gmodels" , "MASS" ))  
install.packages("http://biplot.usal.es/classicalbiplot/categorical-  
biplot/ categoricalbiplottar.gz", repos = NULL, type="source")  
library(BiplotForCategoricalVariables)
```

Se ejecuta escribiendo en la consola la función,

`BiplotForCategoricalVariables()`.

3.3 Comparaciones gráficas de las librerías

En esta sección se explora el alcance de las distintas librerías que realizan la descomposición HJ-Biplot, para lo cual se analiza un conjunto de datos que miden el *Índice para una Vida Mejor (Better Life Index)* en 38 países durante el año 2017.

El índice es elaborado y publicado por la Organización para la Cooperación y el Desarrollo Económico (OCDE) (<https://www.direcon.gob.cl/ocde/>) para sus 34 estados miembros –la cual reúne a la mayoría de las economías

desarrolladas del mundo– y cinco economías emergentes (Brasil, Indonesia, India, Rusia y Sudáfrica).

El Índice para una Vida Mejor se creó con el fin de visualizar y comparar algunos de los factores clave –como *vivienda, ingresos, empleo, comunidad, educación, medioambiente, compromiso cívico, salud, satisfacción con la vida, seguridad y balance vida/trabajo*– que contribuyen al bienestar en los países de la OCDE. Cada tema se basa en uno o más indicadores específicos que hacen un total de 24 indicadores.

En este trabajo se tomó en cuenta un total de 38 países que contaban con la información de los 24 indicadores. Estos países son: Austria (AUT), Australia (AUS), Bélgica (BEL), Canadá (CAN), Chile (CHL), República Checa (CZE), Dinamarca (DNK), Estonia (EST), Finlandia (FIN), Francia (FRA), Alemania (DEU), Grecia (GRC), Hungría (HUN), Islandia (ISL), Irlanda (IRL), Israel (ISR), Italia (ITA), Japón (JPN), Korea (KOR), Letonia (LVA), Luxemburgo (LUX), México (MEX), Países Bajos (NLD), Nueva Zelanda (NZL), Noruega (NOR), Polonia (POL), Portugal (PRT), República Eslovaca (SVK), Eslovenia (SVN), España (ESP), Suecia (SWE), Suiza (CHE), Turquía (TUR), Reino Unido (GBR), Estados Unidos de América (USA), Brasil (BRA), Rusia (RUS) y Sudáfrica (ZAF).

En la Tabla 3 se resumen los diferentes factores y las variables asociadas a cada uno, con sus respectivas siglas para una mejor visualización en los gráficos que se explicarán más adelante.

En la Figura 3-1 se hace una comparación de estos datos utilizando las diferentes librerías en R.

Tabla 3: Factores y variables relacionadas con el Índice para una Vida Mejor

Factores	Variables relacionadas
Vivienda	Viviendas con servicios básicos (VI.1)
	Gastos en vivienda (VI.2)
	Habitaciones por persona (VI.3)
Ingreso	Ingreso familiar disponible neto ajustado (IN.1)
	Patrimonio financiero familiar neto (IN.2)
Empleo	Seguridad en el empleo (TR.1)
	Tasa de empleo (TR.2)
	Tasa de desempleo a largo plazo (TR.3)
	Ingreso medio (TR.4)
Comunidad	Red de apoyo social (Ap.0)
Educación	Logro educativo (Ed.1)
	Competencia de los estudiantes (Ed.2)
	Años de educación (Ed.3)
Medioambiente	Contaminación del aire (Am.1)
	Calidad del agua (Am.2)
Compromiso cívico	Participación de los interesados en la elaboración de regulaciones (Cc.1)
	Participación electoral (Cc.2)
Salud	Esperanza de vida (Sa.1)
	Estado de salud autopercebida (Sa.2)
Satisfacción	Satisfacción global ante la vida (SGVida)
Seguridad	Sentirse seguro por la noche (Se.1)
	Tasa de homicidios (Se.2)
Balance vida-trabajo	Jornada laboral muy larga (Eq.1)
	Tiempo libre dedicado al ocio y al cuidado personal (Eq.2)

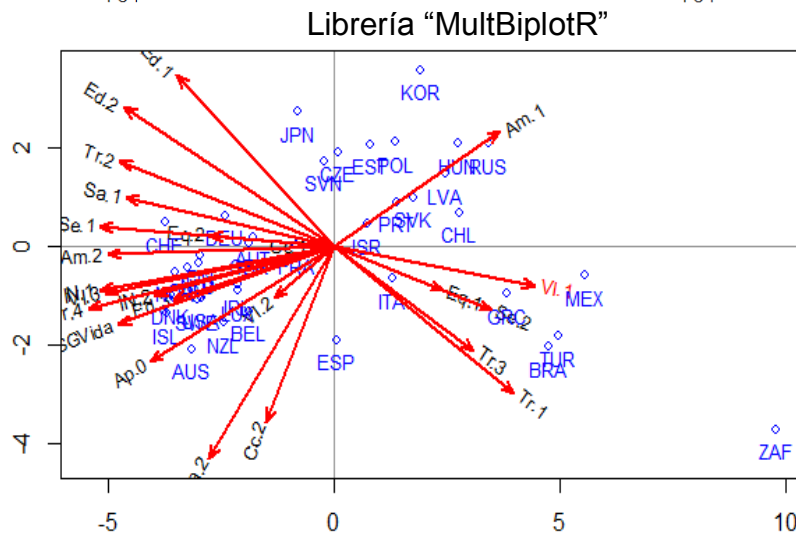
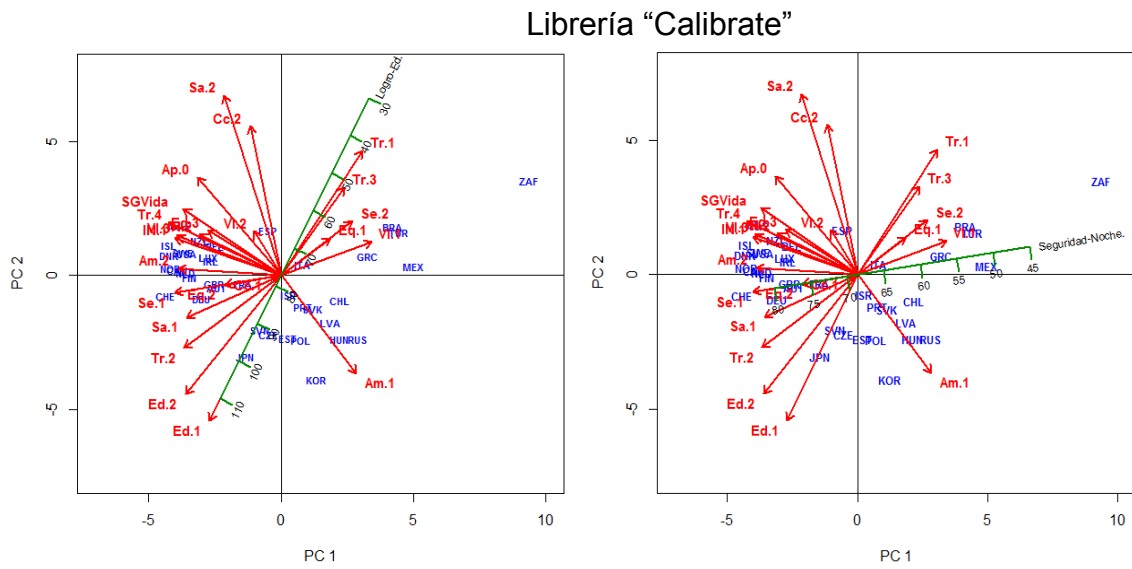
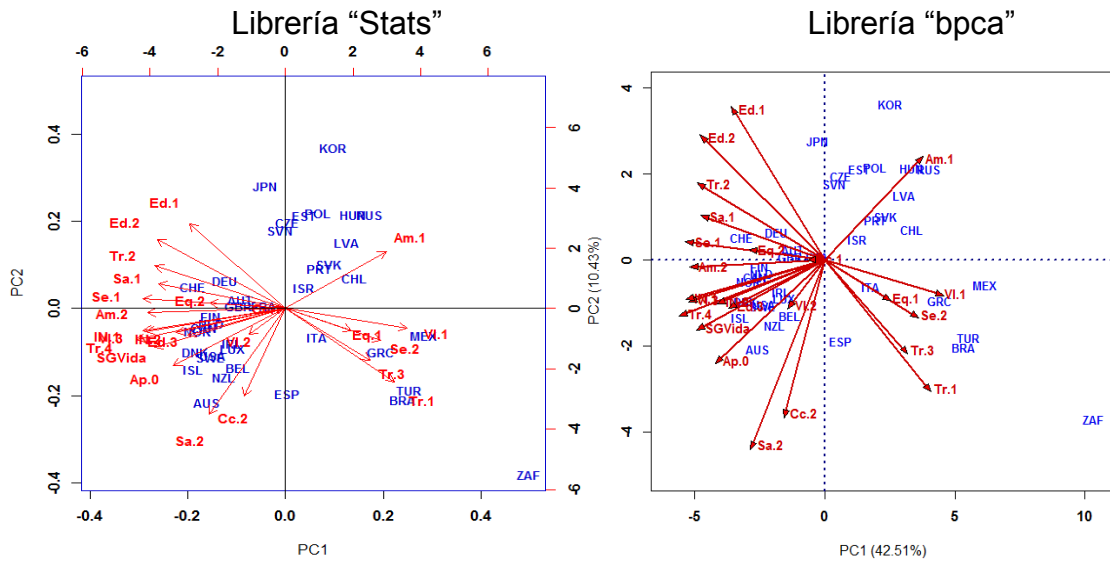


Figura 3-1 Comparación del HJ Biplot obtenida de diferentes librerías en R

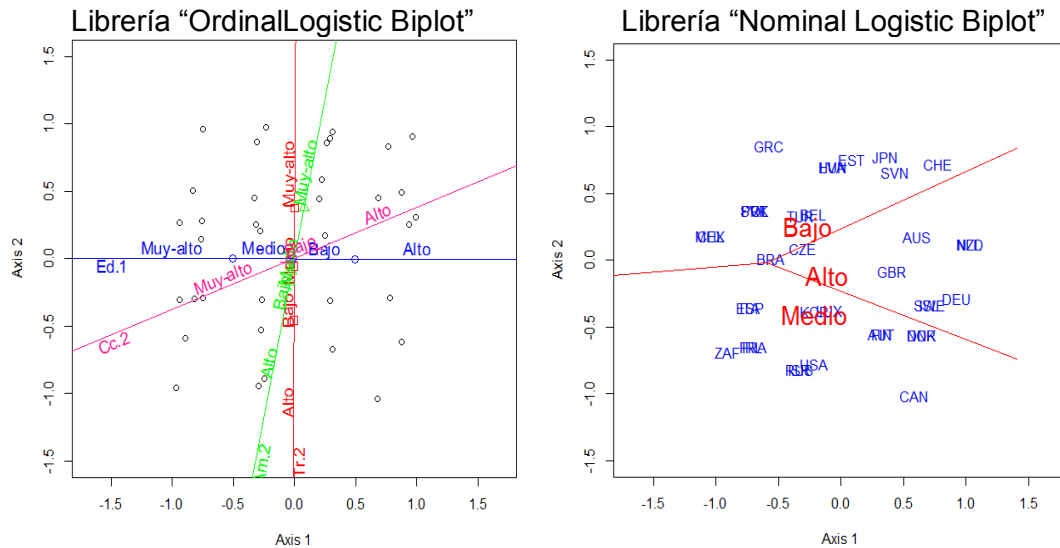


Figura 3-2 Extensiones del Biplot Logístico: Biplot Ordinal y Biplot Nominal

3.4 Interfaz Gráfica de Usuario (GUI)

Como sabemos, una interfaz gráfica permite a cualquier usuario, inclusive a aquellos con poco conocimiento, estimar modelos de forma rápida y sencilla.

Por lo tanto, una GUI es parte fundamental de cualquier aplicación y sirve como vía de aprendizaje que facilita el acceso a múltiples procedimientos estadísticos implementados en R, sin necesidad de conocer propiamente el lenguaje de programación.

Evidentemente, una interfaz de usuario permite un grado de flexibilidad, fluidez y productividad similar a la que tienen los expertos que utilizan lenguajes de programación, pero sin tener que aprenderlo.

En esta línea de trabajo, describiremos algunas interfaces gráficas que abordan técnicas estadísticas multivariantes y que, por tanto, tienen implementada la descomposición y/o representación Biplot (Tabla 4).

Tabla 4: Interfaces gráficas que realizan descomposición Biplot

GUI	Fecha de creación y modificación	Breve Descripción
BiplotGUI (La Grange et al., 2009)	13-08-2008 19-03-2013	Aborda varias técnicas multivariantes, desde una perspectiva biplot. Además, construye y calibra biplots.
multibiplotGUI (Nieto-Librero et al., 2012)	29-10-2012 19-06-2015	Construye e interactúa con biplot múltiples. Permite el manejo de gráficos en 2 y 3 dimensiones.
GGEbiplotGUI (Frutos & Galindo, 2013)	29-08-2011 17-02-2016	Interfaz diseñada para la construcción, interacción y manipulación de biplots GGE en R.
dynBiplotGUI (Egido, 2014)	04-11-2013 12-02-2017	Resuelve Biplot dinámicos y biplots clásicos. Soporta 4 lenguajes: español, inglés, francés y portugués
biplotbootGUI (Nieto-Librero & Galindo-Villardón, 2015)	22-06-2015	Realiza el análisis biplot de forma inferencial, combinándolo con los métodos Bootstrap.

En general, estas aplicaciones son bastante similares en cuanto a la forma para cargar los datos; sin embargo, en su ejecución, suelen estar bastante personalizadas.

3.4.1 Paquete “BiplotGUI”

Biplots (data) es la única función del paquete BiplotGUI, donde “data” contiene los datos a ser analizados. Esta función que se carga a través de la consola del software R da inicio a la interfaz. Ya dentro de la misma GUI, todos los contenidos disponibles se ejecutan a través de un menú de opciones.

Los parámetros que requiere el paquete y que van a la línea de comandos son:

Biplots (Data, groups, PointLabels, AxisLabels)

Data Una matriz o “data.frame” de datos numéricos. Las n muestras u observaciones son representadas como puntos en el biplot y las variables como ejes biplot calibrados.

groups Un vector o factor de longitud n que especifica la pertenencia al grupo de las muestras. Por defecto, todas las muestras se toman de un solo grupo. Las etiquetas de grupo se toman de este argumento.

PointLabels Un vector de longitud n que especifica las etiquetas de los puntos.

AxisLabels Un vector de longitud p que especifica las etiquetas de los ejes.

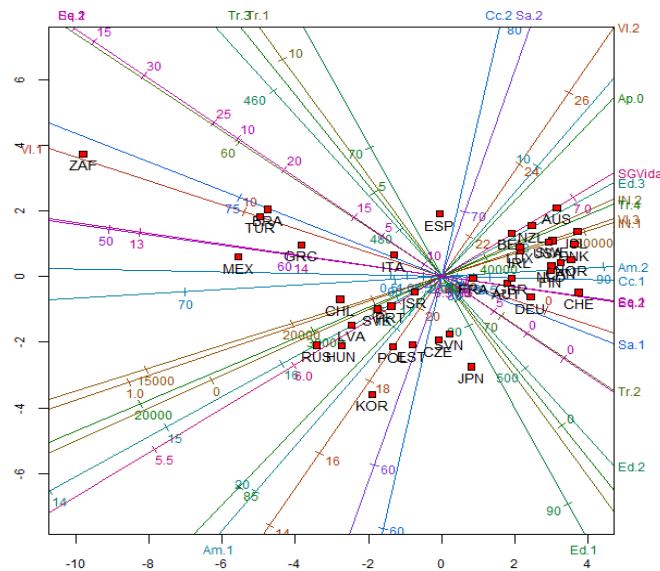


Figura 3-3 Biplot obtenido de la Interfaz Gráfica de Usuario "BiplotGUI"

3.4.2 Paquete “MultiBiplotGUI”

Este paquete permite representar e interactuar gráficos biplot en 2 y 3 dimensiones. Además, proporciona resultados inferenciales utilizando los métodos Bootstrap.

Para ejecutar el programa, solo escribimos en la consola de R: `multibiplot` (X, ni), siendo “X” un data frame en el que se han yuxtapuesto los diferentes conjuntos de datos que van a ser analizados; y “ni” un vector especificando el largo de cada conjunto de datos.

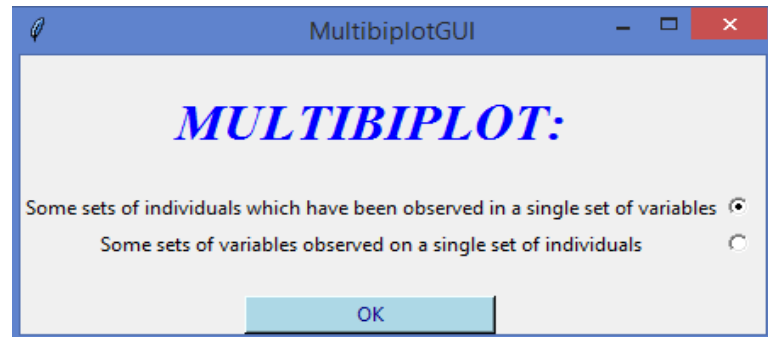


Figura 3-4 Ventana principal del paquete "multibiplotGUI"

El análisis de los datos en esta GUI, se pueden resumir en cinco pasos:

- (1) Estandarización de la matriz X, por columnas
- (2) Análisis individuales
- (3) Creación del biplot ponderado
- (4) Obtención de las medidas de calidad de representación
- (5) Cálculo de la bondad de ajuste.

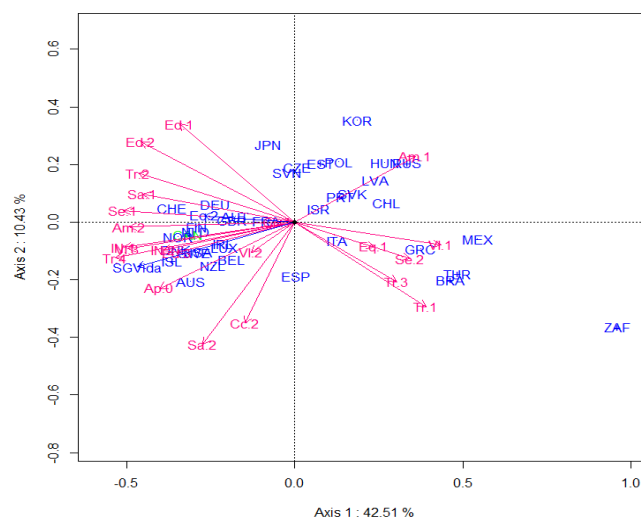


Figura 3-5 Biplot obtenido de la Interfaz Gráfica de Usuario "MultiBiplotGUI"

3.4.3 Paquete “GGEBiplotGUI”

El GGEBiplotGUI proporciona una herramienta de visualización para el análisis e interpretación de la interacción genotipo-ambiente, en la evaluación de cultivos. Para ejecutar y utilizar el programa, primero se debe cargar y sencillamente escribir en la consola lo siguiente: GGEBiplot (data).

La data a cargar puede ser un “data.frame” o una matriz de datos. Inmediatamente se abre una pantalla que da inicio a la interfaz (Figura 3-6). Esta pantalla inicial permite seleccionar el modelo de análisis: SVP (Singular Value Partitioning) para seleccionar el tipo de factorización que se quiere realizar, sea biplot clásico o HJ-Biplot; Centered By, para indicar el tipo de centrado a realizar; y, Scaled para escalar los datos o no.

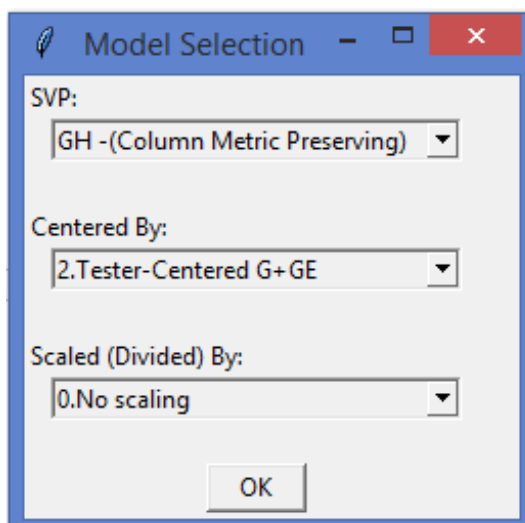


Figura 3-6 Ventana principal del paquete GGEBiplotGUI

Las factorizaciones de que dispone el programa son: GH, JK, SQRT y HJ-Biplot. Las funciones disponibles permiten:

- (a) Clasificar los cultivos según su desempeño en cualquier entorno dado
- (b) Clasificar los entornos según el rendimiento relativo de cualquier cultivo dado

- (c) Comparar el rendimiento de cualquier par de cultivos en diferentes ambientes
- (d) Evaluar los cultivos en base al rendimiento promedio como a la estabilidad
- (e) Agrupar los diferentes ambientes en base a los mejores cultivos,
- (f) Evaluar los entornos en base tanto a la capacidad discriminatoria como a la representatividad
- (g) Identificar el mejor cultivo en cada ambiente.

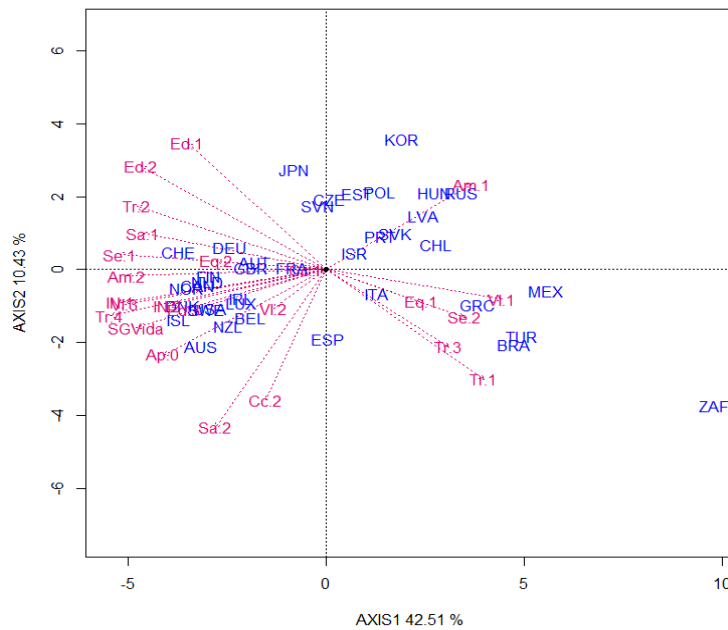


Figura 3-7 Biplot obtenido de la Interfaz Gráfica de Usuario "GGEbiplot"

3.4.4 Paquete "dynBiplotGUI"

Es un programa para el tratamiento de matrices de datos de dos y tres vías (cubo). La función dynBiplot ("es") inicializa la GUI en idioma español. La GUI se puede ejecutar en varios idiomas: "es"=Spanish, "en"=English, "pt"=Portuguese, "fr"=French. Este programa fue creado para dar soporte a la teoría del Biplot Dinámico.

La entrada de datos se puede realizar desde varias fuentes: Excel, SPSS, archivos de texto (txt, CSV) y del entorno R, e incluso se pueden cargar desde el portapapeles.

El programa se ejecuta a través de 4 fases: (1) Carga de datos, (2) Formato de datos, (3) Selección de filas y columnas, (4) Opciones para el análisis de los datos (Figura 3-8).

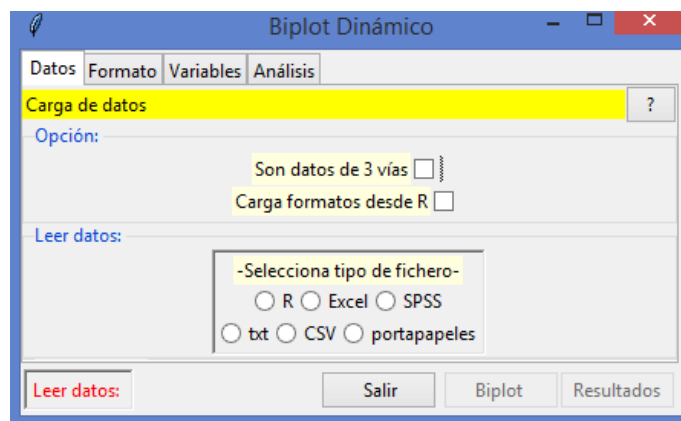


Figura 3-8 Ventana principal del paquete "dynBiplotGUI"

La fase de análisis dispone de cuatro zonas:

- (a) La zona de “Estandarización”, que tiene la opción de centrar o escalar los datos de forma independiente
- (b) La zona de “Análisis Biplot” que permite elaborar el HJ-Biplot, el GH-Biplot o el JK-Biplot
- (c) La zona “Ejes” que selecciona el número de ejes a calcular, y
- (d) La de “Opciones de Gráfico” que controla cómo se van a mostrar los gráficos Biplot.

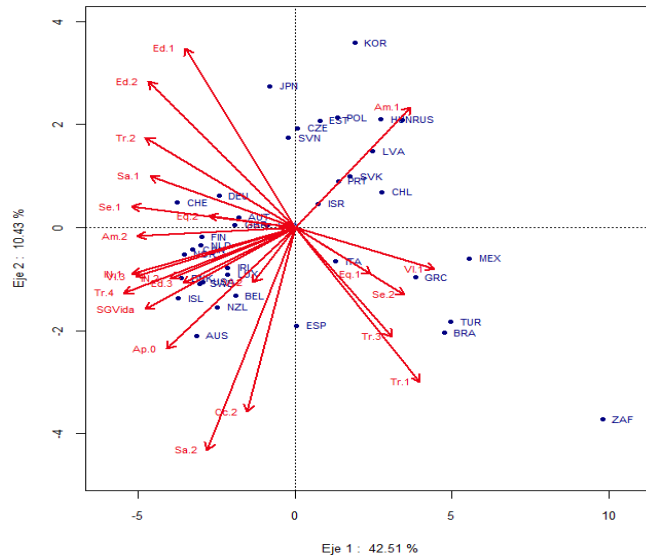


Figura 3-9 Biplot obtenido de la Interfaz Gráfica de Usuario "dynBiplotGUI"

De los paquetes mencionados, todos necesitan ser editados o mejorados, a excepción del “dynbiplot” que no necesita ser retocado; por lo tanto, una vez realizado se puede usar directamente en cualquier informe o presentación (ver Figura 3-9).

3.4.5 Paquete “biplotbootGUI”

El paquete biplotbootGUI se carga a través de la consola del software R, mediante el comando `library(biplotbootGUI)`. A continuación, se inicializa la interfaz mediante el comando `biplotboot(data)`, donde “data” contiene la base de datos para el análisis. Ejecutada la sentencia anterior se abre la ventana principal (Figura 3-10), la cual viene dividida en dos paneles.

Este programa se basa en la metodología Bootstrap para la presentación de sus resultados; esto quiere decir que, en vez de asumir una determinada distribución teórica, se utilizan la muestra original y se generan submuestras que sirven de base para estimar la distribución muestral.

En este sentido, en el panel izquierdo de la ventana principal del programa se debe indicar el número de submuestras que se van a seleccionar a partir de los datos de partida, además del nivel de confianza para calcular los intervalos de confianza. En esta parte de la pantalla también se tiene una opción que permite guardar los gráficos generados, en formato .eps o .pdf, tanto en blanco y negro como en color. En el panel derecho se presenta una lista con todos los resultados que proporciona el análisis Biplot.

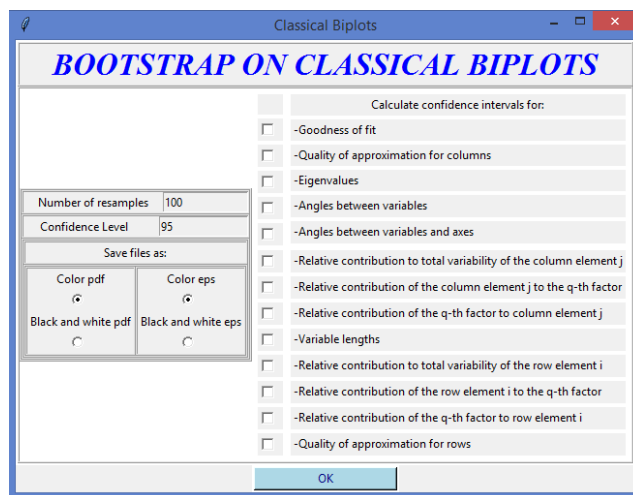


Figura 3-10 Ventana principal del paquete "biplotbootGUI"

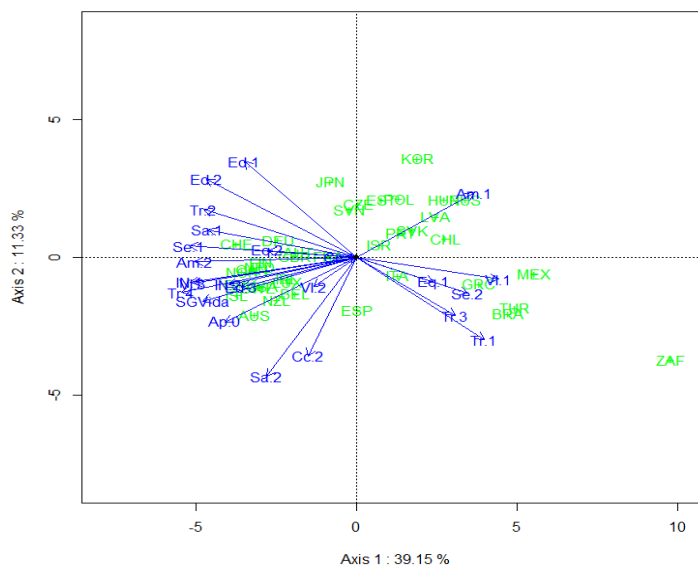


Figura 3-11 Biplot obtenido de la Interfaz Gráfica de Usuario "biplotbootGUI"

Tabla 5: Tabla comparativa de los resultados que devuelven las librerías y las GUI que realizan la descomposición Biplot

Librería	Biplot(s) Clásico(s)	HJ- Biplot	Autovalores (eigenvalues)	Proporción de varianza explicada	Coordenadas de los individuos	Coordenadas de las variables
Biplot {stats}	✓	-	-	-	-	-
bpca	✓	✓	✓	✓	✓	✓
Calibrate (*)	-	-	-	-	-	-
MultBiplotR	✓	✓	✓	✓	✓	✓
Interfaz gráfica de usuario (GUI)						
BiplotGUI (*)	-	-	✓	-	✓	✓
multibiplotGUI	✓	✓	✓	✓	✓	✓
GGEBiplotGUI	✓	✓	✓	✓	✓	✓
dynBiplotGUI	✓	✓	✓	✓	✓	✓
biplotbootGUI	✓	✓	✓	✓	✓	✓

(*) Calibra vectores de variables en el Biplot. Se basa en la idea de [Gower & Hand \(1996\)](#) en la cual las variables se representan como ejes calibrados

3.5 Otros Paquetes

Dentro del entorno estadístico de R existen también otros paquetes que si bien, no realizan la descomposición biplot, hacen referencia a este término (*biplot*), realizando en una misma gráfica la representación conjunta de coordenadas, obtenidas a través de otros métodos multivariantes (ver Tabla 6). A continuación, una descripción de cada uno de estos paquetes.

Tabla 6: Paquetes de R que utilizan el término Biplot, pero no realizan la DVS

Paquete	Fechas de creación y de actualización	Breve descripción
Vegan (Oksanen et al., 2013)	06-09-2001 04-02-2019	Muestra el biplot a partir de los resultados del Análisis de redundancia y la correlación canónica.
ade4 (Chessel et al. 2013; Chessel, Dufour, & Thioulouse, 2004; Dray et al. 2007)	10-12-2002 31-08-2018	Realiza el biplot a través del Análisis de Componentes Principales, Análisis de Redundancia, Análisis de Correspondencia y Análisis de coinercia.
Ade4TkGUI (Thioulouse & Dray, 2007, 2012)	29-09-2006 09-11-2015	Complementario al ade4, es una interfaz gráfica que ejecuta las principales funciones del programa ade4.
Ca (Greenacre & Nenadic, 2012; Nenadic & Greenacre, 2007)	28-07-2007 10-10-2018	Muestra el biplot desde los resultados del “Análisis de Correspondencia” (simple, múltiple, conjunto).
caGUI (Markos, 2012)	04-10-2009 29-10-2012	Es una interfaz gráfica diseñada para realizar las funciones del programa ca.

3.5.1 Paquete “Vegan”

Este paquete cuenta con una variedad de herramientas para la representación, ordenación y clasificación de individuos a partir del análisis multivariante de datos. El argumento central que utiliza en el análisis de datos es el “*análisis de redundancia*”.

Dentro de las técnicas de reducción de la dimensionalidad, es posible el Análisis de Componentes Principales (ACP), el Análisis de Coordenadas Principales (MDS), el Análisis Factorial de Correspondencia (AC) y el Multidimensional Scaling No Métrico (NMDS). En relación a los modelos de clasificación, permite realizar cluster jerárquicos y algoritmos partitivos (K-means y fuzzy C-means) y no partitivos.

La siguiente función devuelve el biplot:

`(x, choices, scaling, display, type, xlim, ylim, col, const, correlation)`, donde,

- | | |
|----------------|---|
| x | Objeto con extensión rda (análisis de redundancia) |
| choices | Ejes a mostrar (1 o 2) |
| scaling | Escalado para especies y puntuaciones de los sitios. Hay varias alternativas: las puntuaciones de las especies o de los sitios se escalan según los valores propios, o ambos se escalan simétricamente por la raíz cuadrada de los valores propios o se dejan sin escalar. El scaling se especifica entonces como “ <i>none</i> ”, “ <i>sites</i> ”, “ <i>species</i> ” o “ <i>symmetric</i> ”, que corresponden a los valores 0,1,2 y 3 respectivamente. |

display	Puntuaciones mostradas. Deben ser “species” para las especies y/o “sites” para las puntuaciones de los sitios.
type	Tipo de gráfico: puede ser de longitud 2 (por ejemplo, tipo = c ("texto", "puntos"), en cuyo caso el primer elemento describe cómo se manejan los puntajes de las especies y el segundo cómo se dibujan los puntajes del sitio
xlim, ylim	El límite de x y de y (min, max) de la gráfica
col	Colores usados para sitios y especies (en este orden), si solo se da un color se usa para ambos
const	Constante de escala general para scores.rda
correlation	Valor lógico (TRUE or FALSE), se puede utilizar para seleccionar puntuaciones similares a la correlación para ACP.

3.5.2 Paquete “ade4”

Este paquete contiene una amplia variedad de herramientas para el análisis de datos multivariantes, proporcionando varios métodos para el análisis de 1, 2, 3 y k-tablas. ade4 no realiza la descomposición biplot; sin embargo, implementa la función “**biplot**” sobre un objeto de clase *dudi* que contiene los resultados de un análisis multivariante.

El diagrama de dualidad (*dudi*), herramienta básica de ade4, consiste simplemente en una lista que contiene el triplete (X, Q, D) donde, X : es una matriz de datos $(n \times p)$, Q : es una matriz simétrica positiva $p \times p$ que calcula la

distancia entre los individuos, y D : es una matriz simétrica de dimensión $n \times n$ que calcula la distancia entre variables.

Cada método corresponde a un triplete en particular; así las diferentes funciones *dudi* devuelven resultados según corresponda a un análisis de componentes principales, análisis de coordenadas, análisis de correspondencia simple y múltiple, entre otros.

3.5.3 Paquete “Ade4TkGUI”

ade4TkGUI es una interfaz gráfica diseñada con el objetivo de ejecutar las principales funciones del programa ade4.

El núcleo del paquete es la función **ade4TkGUI ()**, que abre la ventana principal. En la ventana principal de la GUI, los botones se agrupan de acuerdo a su función: conjuntos de datos, análisis de una tabla, análisis de una tabla con grupos de filas, análisis de dos tablas, funciones gráficas y gráficos avanzados. Sólo los métodos de una y dos tablas están disponibles en ade4TkGUI. Las funciones de uso menos frecuente están disponibles a través de los menús de la barra de menús, que se encuentran en la parte superior de la ventana.

3.5.4 Paquete “ca”

Este paquete consta de dos partes, una para el análisis de correspondencia simple (AC) y otra para el análisis de correspondencia múltiple (ACM) y conjunta (ACJ).

La representación gráfica de los resultados se realiza con los llamados “*mapas simétricos*”. Existen dos opciones que especifica el tipo de mapa: la función

`plot.ca` que crea un mapa bidimensional del objeto creado a través del análisis de correspondencia simple; y, la función `plot.mjca` que lo hace para el análisis de correspondencia múltiple y conjunto.

Desde las funciones `plot.ca(), map=" "` y `plot.mjca(), map=" "` se listan a continuación los parámetros necesarios.

symbiplot: escala las filas y las columnas para que tengan variaciones iguales a los valores propios, lo que da un biplot simétrico, pero no conserva las métricas de fila ni de columna

simmetric: filas y columnas en coordenadas principales

rowprincipal mapa que preserva la métrica de las filas

colprincipal: mapa que preserva la métrica de las columnas

rowgab: mapa asimétrico con filas en coordenadas principales y columnas en coordenadas estándar multiplicadas por la masa del punto correspondiente

colgab: mapa asimétrico con columnas en coordenadas principales y filas en coordenadas estándar multiplicadas por la masa del punto correspondiente

rowgreen: mapa similar a rowgab, excepto que los puntos en coordenadas estándar se multiplican por la raíz cuadrada de las masas correspondientes.

colgreen: mapa similar a colgab, excepto que los puntos en coordenadas estándar se multiplican por la raíz cuadrada de las masas correspondientes.

3.5.5 Paquete “caGUI”

caGUI es una interfaz gráfica para el cálculo y visualización del análisis de correspondencia simple, múltiple y conjunto con las funciones del paquete ca. Para iniciar la interfaz se utiliza la función **caGUI ()**. La ventana de diálogo principal contiene dos pestañas, una para Análisis de Correspondencia Simple y otra para Análisis de Correspondencia Múltiple y Conjunto.

**Capítulo 4 SOLUCIONES
DISJUNTAS Y SPARSE
BIPLOT**

4.1 Sparse PCA y Soluciones Disjuntas

En la era del "Big Data", se trabaja con grandes cantidades de datos que surgen en diferentes disciplinas como las ciencias sociales, biología, ingeniería, astronomía, física y medicina, entre otras. Como ejemplo particular, en la biología molecular y la ciencia genómica, los datos de micro arreglos generan miles de características sobre tejidos y muestras de células, que pueden ser útiles para estudiar los diferentes tipos de cáncer o bien para diagnosticar enfermedades. La complejidad de estos datos exige el uso de técnicas capaces de simplificar la información original y proporcionar significado a los resultados obtenidos.

Las técnicas multivariantes para el análisis de datos se basan en el método de descomposición matricial que busca explorar grandes volúmenes de información, aprovechando su alta dimensionalidad. La representación más común es el Análisis de Componentes Principales ([Hotelling, 1933](#); [Pearson, 1901](#)) realizado a través de la Descomposición en Valores Singulares (DVS) de [Eckart & Young \(1936\)](#). Desde el punto de vista algebraico, los métodos biplot se basan en el mismo principio, la DVS de una matriz de datos.

En el biplot los datos se proyectan ortogonalmente sobre ejes de máxima variabilidad en un espacio de dimensión reducida. Así, cada componente principal se expresa como una combinación lineal de las variables originales y establecen su contribución a cada PC. Los coeficientes de las combinaciones, denominados cargas y habitualmente distintos de cero, generan el principal inconveniente del biplot: *su interpretación*.

Se han propuesto varios métodos para modificar el análisis de componentes principales, con el fin de mejorar la interpretación de sus resultados, que van desde las técnicas de rotación hasta la imposición de restricciones sobre las cargas factoriales del ACP.

Inicialmente, [Hausman \(1982\)](#) propuso restringir el valor que se le puede asignar a las cargas de las componentes principales a un conjunto de números enteros $\{-1,0,1\}$ con el interés de construir componentes simplificados. Posteriormente, [Vines \(2000\)](#) acoge la idea de [Hausman \(1982\)](#) y sugiere el uso de números enteros arbitrarios. Por su parte, [McCabe \(1984\)](#) plantea un método diferente que consiste en seleccionar un subconjunto de variables, identificadas bajo el concepto de variables principales, a partir de un criterio de optimización, sin necesidad de pasar por el ACP. Para resolver el problema, [Cadima & Jolliffe \(1995\)](#) presentan un método denominado “umbral simple”, que consiste en convertir las cargas factoriales con valores absolutos menores que cierto umbral, en cargas nulas. Tradicionalmente, para simplificar la estructura de las CPs y facilitar su interpretación se han utilizado las técnicas de rotación ([Jolliffe,1995](#)). Sin embargo, la reducción de la dimensionalidad de los datos no siempre es suficiente para facilitar la interpretación de las CPs. Una forma *ad hoc* consiste en utilizar las técnicas de regularización que, aunque requieren un parámetro de restricción para inducir vectores de proyección con cargas modificadas (nulas o cercanas a cero) y controlar el peso de las cargas, mejoran notablemente la interpretación de los resultados. Así, [Tibshirani \(1996\)](#) introdujo el método *LASSO (Least Absolute Shrinkage and Selection Operator)*, en el que combinó un modelo de regresión con un

procedimiento de contracción de algunos parámetros hacia cero imponiendo una penalización sobre los coeficientes de regresión. Años después, [Jolliffe & Uddin \(2000\)](#) presentan una solución que modifica el enfoque tradicional de CPs en dos etapas (ACP + rotación). Plantean la técnica de componentes simplificados ScoT (Simplified component Technique), en la cual las CPs originales seguidas de la rotación VARIMAX se combinan en un solo paso para inducir cargas con poca densidad, manteniendo la disminución de la proporción de varianza explicada. Como las cargas obtenidas mediante ScoT consiguen valores pequeños distintos de cero, pero no nulos; [Jolliffe, Trendafilov, & Uddin \(2003\)](#), proponen el algoritmo SCoTLASS (Simplified Component Technique subject to LASSO), imponiendo una restricción de forma tal que algunas cargas se hacen completamente nulas, pero sacrificando la varianza. En el mismo sentido, [Zou et al. \(2006\)](#) proponen un algoritmo de regresión penalizado usando la técnica “*Elastic Net*” (llamado *Sparse PCA*) ([Zou & Hastie, 2005](#)) que resolvieron eficientemente utilizando la regresión de ángulo mínimo ([Efron, Hastie, Johnstone, & Tibshirani, 2004](#)). Sujeto a restricciones de cardinalidad (número de cargas cero, por componente), [Moghaddam et al. \(2006\)](#) construyen un algoritmo para componentes *Sparse*. Seguidamente, [d’Aspremont et al. \(2007\)](#) explican la restricción de cardinalidad en base a programación semidefinida.

Aprovechando algunas de las ideas anteriores, [Shen & Huang \(2008\)](#) conectan el ACP con la DVS de los datos y obtienen CPs *Sparse* mediante penalizaciones de regularización (sPCA-rSVD). [Witten, Tibshirani, & Hastie \(2009\)](#) unifican el enfoque de aproximación matricial de bajo rango de [Shen &](#)

Huang (2008) con el criterio de máxima varianza de Jolliffe et al. (2003) y el método sparse PCA de (Zou et al., 2006), para dar una solución general al problema de sparse PCA. En el mismo contexto, Farcomeni (2009) sugirió maximizar la varianza explicada penalizada por la cardinalidad de las CPs Sparse. Además, Vichi & Saporta (2009) presentan una modificación al PCA, que permite identificar componentes de máxima varianza y garantiza que cada variable contribuya solo a uno de los ejes factoriales. Qi, Luo, & Zhao (2013) proponen una forma relativamente simple para abordar el problema. Mediante una extensión del ACP clásico, construyen componentes sparse reemplazando la norma l_2 en problemas de valores propios tradicionales con una nueva norma que es la combinación de las normas l_1 y l_2 .

Por otro lado, Mahoney & Drineas (2009) formulan una alternativa a la DVS a través de la descomposición de la matriz CUR, expresada en un pequeño número de filas y/o columnas en una aproximación de bajo rango de la matriz original. Su objetivo es procurar una mejor interpretación de la información, seleccionando las variables más importantes de la matriz de datos. Es una técnica diferente, en el sentido de que no está orientada a la obtención de ejes factoriales, lo cual le otorga de alguna manera, ventajas frente a métodos tradicionales. La CUR descompone una matriz A_{ij} ($m \times n$) como el *producto* de tres matrices C , U y R donde $C_{m \times r}$ y $R_{r \times n}$ están formadas por un subconjunto de columnas (Figura 4-1) y filas (Figura 4-2) de la matriz original, respectivamente. En tanto, $U_{r \times r}$ es una matriz de bajo rango construida cuidadosamente a partir de las columnas de C y las filas de R .

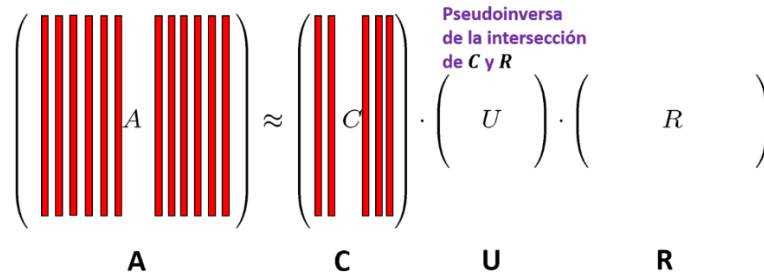


Figura 4-1 CUR: Selección de columnas (C)

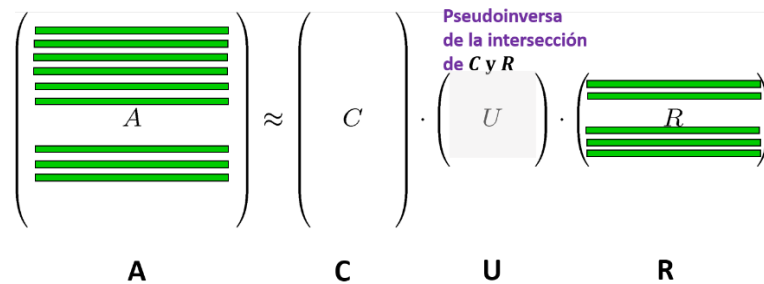


Figura 4-2 CUR: Selección de Filas (R)

Sea A una matriz de m filas y n columnas. $A \in \mathbf{R}^{m \times n}$.

El primer paso en la descomposición de la matriz CUR es la DVS de la matriz A . La DVS devuelve los vectores singulares izquierdo y derecho para poder calcular las puntuaciones de más influencia tanto de filas como de columnas, esto es: $A = U \Sigma V^T$ donde, $U = [u^1 \ u^2 \ \dots \ u^m]$ y $V = [v^1 \ v^2 \ \dots \ v^n]$ son matrices ortogonales y Σ es una matriz diagonal.

Sea u^ξ y v^ξ el ξ -ésimo vector singular izquierdo y derecho de A , la j -ésima columna de A puede ser aproximada mediante los k valores y vectores singulares como:

$$A^j = \sum_{\xi=1}^k (\sigma_\xi u^\xi) v_j^\xi$$

donde v_j^ξ es la j -ésima coordenada del ξ -ésimo vector singular derecho.

Tanto filas como columnas son seleccionadas según la probabilidad que determina el “*puntaje de importancia*” para cada columna (o fila) en aproximación de dimensión reducida.

Las aproximaciones de la descomposición CUR se pueden obtener a través de diferentes criterios. [Stewart \(1999\)](#) desarrolló el método *cuasi-Gram-Schmidt* para obtener una descomposición de la matriz CUR y construir una aproximación de bajo rango al proyectar en los espacios abarcados por columnas y filas. Un límite de error aportado por Stewart para la calidad de la aproximación es:

$$\|A - CUR\|_F^2 \leq \varepsilon_C^2 + \varepsilon_R^2.$$

[Goreinov, Tyrtshnikov, & Zamarashkin \(1997\)](#) y [Goreinov & Tyrtshnikov \(2001\)](#) se interesaron en aplicaciones como la dispersión, en las que las matrices de coeficientes grandes se pudiesen aproximar mediante matrices de bajo rango y desarrollaron una descomposición de la matriz CUR a la que llamaron “*pseudoskeleton*”. Demostraron que si la matriz A_{ij} es aproximada por una matriz de rango k dentro de una precisión ε , es decir, si existe una matriz E tal que $\text{rango}(A - E) \leq k$ y $\|E\|_2 \leq \varepsilon$. Entonces existe una opción de columnas (C) y filas (R), y una matriz U de baja dimensión, construida a partir de los elementos de R y C tal que $A \approx CUR$ en el sentido de que:

$$\|A - CUR\|_2 \leq \varepsilon (1 + 2\sqrt{km} + 2\sqrt{kn})$$

[Frieze, Kannan, & Vempala \(2004\)](#), muestrean aleatoriamente columnas de A según una distribución de probabilidad que depende de las normas euclidianas

de esas columnas. Si el número de columnas elegidas es polinomial en k y $1/\epsilon$ (para algunos parámetros de error), entonces las garantías de error de la forma

$$\|A - P_C A\|_F \leq \|A - A_K\|_F + \epsilon \|A\|_F$$

se pueden obtener con alta probabilidad. $P_C A$ denota la proyección de A en el subespacio abarcado por las columnas de A .

Posteriormente, [Drineas, Kannan, & Mahoney \(2006\)](#) construyeron una descomposición de la matriz CUR al elegir columnas y filas simultáneamente. A partir de la matriz A , construyen aleatoriamente una matriz C de columnas, una matriz R de filas y una matriz U tal que

$$\|A - CUR\|_F \leq \|A - A_K\|_F + \epsilon \|A\|_F \text{ con alta probabilidad.}$$

[Mahoney & Drineas \(2009\)](#) definen este factor de importancia para cada columna, denotado como π_j como se muestra a continuación (Drineas et al., 2006):

$$\pi_j \approx \frac{1}{k} \sum_{p=1}^k (v_j^p)^2 \text{ para } j= 1, 2, \dots, n$$

donde v_j^p es la j -ésima componente del p -ésimo vector singular derecho de A
 k es el número aleatorio de *columnas* a utilizar

De manera similar, los puntajes de importancia para cada *fila* se calculan a partir de los mejores k vectores singulares izquierdos de A .

$$\pi_i \approx \frac{1}{k} \sum_{p=1}^k (u_i^p)^2 \text{ para } i= 1, 2, \dots, m$$

La matriz \mathbf{U}_{rxr} se construye a partir de las matrices C y R como sigue:

- (1) Sea $W_{r \times r}$ la matriz que resulta de la intersección de las columnas y las filas elegidas de C y R , respectivamente.
- (2) Obtenida W , se calcula la DVS de W ; es decir $W = X\Sigma Y^T$
- (3) Calculamos la pseudo inversa de Moore-Penrose (Σ^+) de la matriz diagonal Σ . Es decir, reemplazamos todos los elementos distintos de cero en Σ por sus respectivos inversos.
- (4) Obtenemos $U = Y(\Sigma^+)^2 X^T$.

Finalmente tenemos la descomposición de la matriz $M \approx CUR$. (matriz CUR). El algoritmo de descomposición de la matriz CUR se ha aplicado con éxito a muchos dominios, incluyendo oncología, en el análisis en tumores (Bodor, Csabai, Mahoney, & Solymosi, 2012); en el análisis de imágenes médicas (Mahoney, Maggioni, & Drineas, 2008); en Química y Biología, en la identificación de iones y en el estudio de muestras biológicas complejas (Yang, Rübél, Prabhat, Mahoney, & Bowen, 2015); en Bioinformática (Mahoney & Drineas, 2009); en el análisis de datos de texto (Drineas, Mahoney, & Muthukrishnan, 2008).

A partir de la descomposición CUR, es posible realizar el HJ-Biplot, utilizando solo las variables de mayor relevancia en la matriz de datos original. El proceso inicia calculando los *leverages scores* para cada variable y seleccionando las variables con los scores más altos. Una vez reducida la dimensionalidad de los datos se aplica el HJ-Biplot sobre el nuevo conjunto de datos. De esta manera se eliminan las variables que tienen poca influencia y se mantienen las variables que aportan mayor variabilidad.

Bajo este enfoque, se resume el procedimiento a través de los siguientes pasos:

- (1) Calcular los *leverages scores* para cada variable.
- (2) Determinar el número “*k*” de variables deseadas.
- (3) Crear el nuevo conjunto de datos utilizando las “*k*” variables de mayor puntaje.
- (4) Realizar un HJ Biplot a partir del nuevo conjunto de datos.

Siguiendo estos pasos, obtenemos la representación CUR + HJ Biplot (ver Figura 4-3) para la matriz de datos definida en la tabla 3. A partir de este enfoque se puede observar que de 24 variables que tiene la matriz original, el algoritmo devuelve una representación que contiene sólo 15 variables, aquellas con las puntuaciones más altas; de esta manera se ha eliminado información menos relevante.

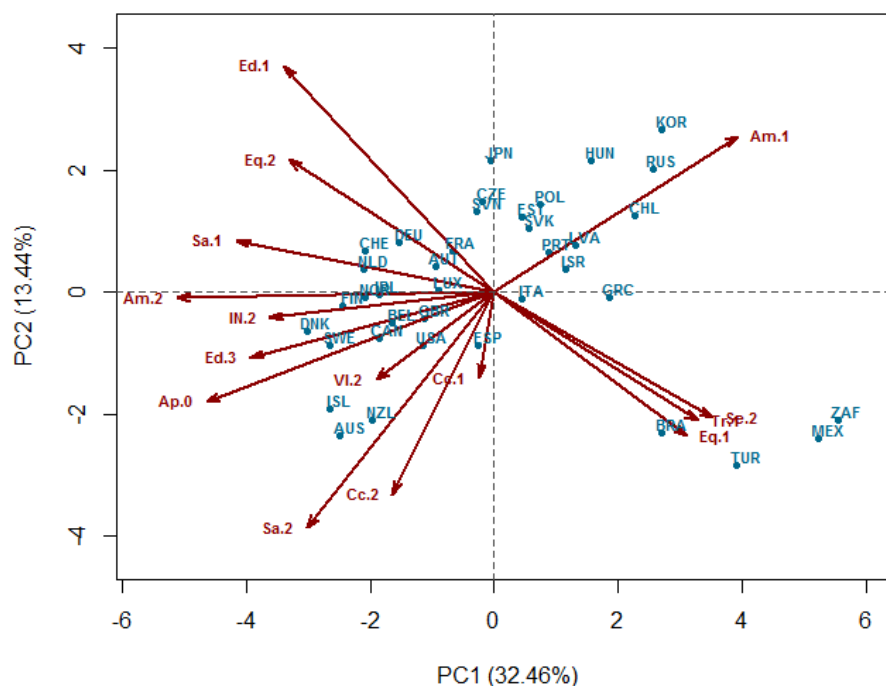


Figura 4-3 CUR + HJ Biplot. Datos sobre el Índice para una vida mejor

Un algoritmo para la representación HJ-Biplot es propuesto por Nieto-Librero (2015) y Nieto-Librero et al. (2017). Definido como *Disjoint Biplot* (DBiplot), es un método que construye ejes factoriales disjuntos garantizando que cada variable de la matriz de datos original contribuya solamente a una componente principal (ver Figura 4-4). El algoritmo parte de una clasificación aleatoria de las variables en las CP's, y mediante un procedimiento iterativo busca la clasificación óptima que conduzca a la maximización de la variabilidad explicada. La representación gráfica de objetos y variables en este nuevo espacio de dimensión reducida se realiza a través del HJ-Biplot. Para ello, una función denominada *CDBiplot* se aloja dentro de la interfaz gráfica **biplotbootGUI** (Nieto-Librero & Galindo-Villardón, 2015). La interfaz posee tres funciones principales. La función *CDBiplot* ejecuta la interfaz gráfica para construir el *Disjoint Biplot* (DBiplot); la representación de clusters, mediante la función *Clustering Biplot* (CBiplot), y el *Clustering Disjoint Biplot* permite la representación conjunta tanto de individuos como de variables.

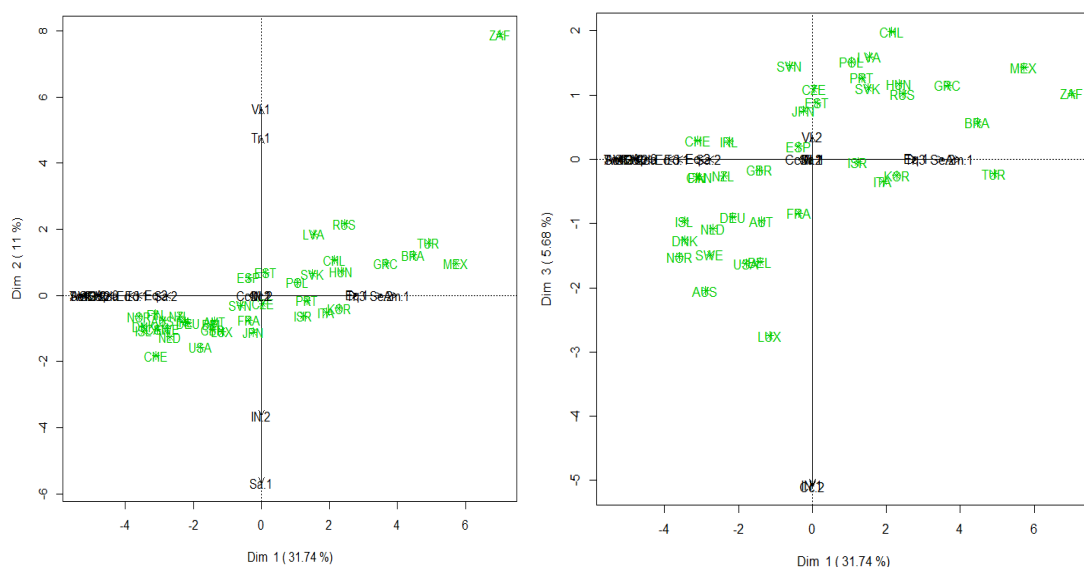


Figura 4-4 Representación del Disjoint Biplot: izquierda (plano 1-2) y derecha (plano 1-3)

El Sparse PCA también puede ser abordado a través del enfoque bayesiano. De esta manera, la forma estándar del ACP es reemplazada por su versión probabilística (Trendafilov, 2014) siguiendo la misma estrategia que los métodos mencionados anteriormente.

Trendafilov (2014) realiza una profunda revisión de los principales enfoques para la interpretación del PCA. Similarmente, Zhang, Xu, Yang, Li, & Zhang (2015) dan una visión general de los algoritmos *sparse* desde el punto de vista de la teoría de la optimización matemática.

4.2 Sparse Biplot

En el contexto del Biplot, no hemos encontrado ninguna evidencia que formule algoritmos alternativos para penalizar o contraer las cargas de las componentes principales, con el fin de mejorar la interpretación de la información que aportan los datos de alta dimensionalidad. En esta línea de trabajo, este documento propone nuevas alternativas de representación Biplot que consisten en adaptar restricciones para contraer y/o producir cargas nulas en las componentes, en base a las teorías de regularización *Ridge*, *LASSO* y *Elastic Net*. En cada caso se demuestra el funcionamiento de los algoritmos mediante la creación en lenguaje R del paquete “SparseBiplots”, diseñado exclusivamente para dar soporte a la nueva metodología planteada. La implementación del paquete se realiza con los datos del *Índice para una Vida Mejor* utilizados en el capítulo anterior. Además, se utiliza una muestra en la que analizamos indicadores sociales de grandes compañías que informan sobre *Responsabilidad Social Corporativa (RSC)* de conformidad con el

modelo del Global Reporting Initiative (Cubilla-Montilla, Nieto-Librero, Galindo-Villardón, Vicente Galindo, & Garcia-Sanchez, 2019).

El Biplot *Sparse* aborda el problema de encontrar una combinación lineal de las variables, determinado por un vector de cargas *sparse* que maximiza la variabilidad de los datos o minimiza el error de construcción. Este enfoque mejora notablemente la capacidad de interpretar los ejes (*Sparse*) obtenidos.

El Biplot se reformula como un modelo de regresión lineal simple, bajo el concepto de minimización del “*error de reconstrucción (E)*”.

$$E = \left[\|X - \hat{X}\|^2 \right] = \text{Tr}(E[(X - \hat{X})(X - \hat{X})^T])$$

Esto quiere decir que se busca minimizar la diferencia entre la matriz original y los datos que se obtendrían proyectando en el espacio original las p nuevas variables.

La formulación del Biplot como un problema de regresión, impone restricciones en las cargas factoriales para producir “*ejes modificados*”. Obviamente, la incorporación de una restricción adicional proporciona unas dimensiones que, en general, no explican toda la varianza que se explica en las dimensiones originales. No obstante, las técnicas de regularización², proporciona cierta estabilidad en el proceso y mejora la capacidad de generalización del modelo. La idea principal de esta tesis es estudiar diferentes métodos de regularización con el fin de implementarlos en el Biplot y generar modelos más interpretables.

² El parámetro de regularización λ controla la fuerza o importancia que le damos a la regularización en el proceso de optimización.

De los diferentes métodos de regularización, nos vamos a centrar en tres de ellos: Ridge (Hoerl & Kennard, 1970), LASSO, (Tibshirani, 1996) y Elastic Net (Zou et al., 2006). Estos métodos han sido ampliamente cubiertos en la literatura cuando se aplican a modelos clásicos de regresión lineal y componentes principales, pero no han sido estudiados en el contexto de las técnicas Biplot. Para cada método de regularización se presenta en detalle el procedimiento para su implementación en el HJ Biplot, teniendo en cuenta las particularidades o características propias de cada uno.

4.3 Ridge HJ Biplot

La regresión *Ridge* fue propuesta originalmente por Hoerl & Kennard (1970) como un método alternativo frente al problema de colinealidad, en un modelo lineal general estimado por mínimos cuadrados. Se formula añadiendo una restricción al modelo lineal, definida en términos de la norma l_2 de $\beta = \beta_1, \dots, \beta_p$

$$l_2 = (\|\beta\|_2)^2 = \sum_{j=1}^p \beta_j^2$$

La regresión *Ridge* impone una penalización al tamaño de los coeficientes de regresión, con lo cual estos coeficientes disminuyen su valor. Los coeficientes estimados por Ridge, $\hat{\beta}^{ridge}$, son los valores que minimizan la suma de los residuos al cuadrado penalizada:

$$\hat{\beta}^{ridge} = \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j X_{ij} \right)^2 + \lambda \sum_{j=1}^p \beta_j^2 = \|Y - \beta X\|^2 + \lambda \sum_{j=1}^p \beta_j^2$$

donde $\lambda \geq 0$

Una forma equivalente de escribirlo es:

$$\hat{\beta}^{ridge} = \arg \min \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j X_{ij} \right)^2 \text{ sujeto a } \sum_{j=1}^p \beta_j^2 \leq \lambda$$

siendo λ el parámetro de regularización que controla la reducción de β_j . Cuanto mayor sea λ , mayor contracción de los coeficientes.

Usualmente las variables son estandarizadas previamente, para evitar que la penalización varíe frente a cambios de escala.

Sabemos que $\hat{\beta}^{MCO} = (X^T X)^{-1} X^T Y$, es la estimación por mínimos cuadrados de β . En un principio se ha planteado la inestabilidad de $\hat{\beta}^{MCO}$ ya que $(X^T X)^{-1}$ no se puede calcular cuando $p \gg n$.

La regresión *Ridge* plantea una solución a este problema con unos parámetros β contraídos. Esto nos lleva al estimador *Ridge*:

$$\hat{\beta}^{ridge} = (X^T X + \lambda I)^{-1} X^T Y \quad \text{donde } I \text{ es la matriz identidad } p \times p.$$

Como se observa, la solución *Ridge* sigue siendo una función lineal de Y . Lo que se hace es añadir una cantidad, λ ($\lambda \geq 0$) a la diagonal de $X^T X$ antes de invertir la matriz, para alcanzar una matriz no singular. A partir de esta restricción, los nuevos β se obtienen mediante la siguiente expresión:

$$\hat{\beta}^{ridge} = \frac{\hat{\beta}^{MCO}}{1 + \lambda}$$

Ahora bien,

- Si $\lambda \rightarrow 0$ estamos en el caso de mínimos cuadrados ordinarios.
- Si $\lambda \rightarrow \infty$, $\hat{\beta}^{ridge} \rightarrow 0$ y estamos ante un estimador sesgado de β .

Así, mediante la penalización *Ridge* si bien contraemos las estimaciones de los coeficientes a cero; introducimos sesgo, pero reducimos la varianza de las estimaciones. Al contraer todos los coeficientes hacia cero no se consigue la nulidad de ninguno de éstos; por tanto, la regresión *Ridge* no selecciona variables, permaneciendo todas en el modelo.

Para entender el funcionamiento de la regresión *Ridge* en el Biplot, utilizamos la DVS de la matriz de datos X .

Sea X una matriz de datos $n \times p$ con rango $K \leq \min(n, p)$. La descomposición en valores singulares de los datos se puede escribir de la forma:

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

donde $U^T U = I_n$, $V^T V = I_p$, $d_1 \geq d_2 \geq \dots \geq d_k > 0$

Sea $X \rightarrow \hat{X} = UDV^T$ donde UD contiene las coordenadas para filas y DV contiene las coordenadas para las columnas. Expresamos:

$$A = UD = XV$$

y

$$B = VD$$

Entonces, A contiene las componentes principales (de la matriz XX^T), y las columnas de V las correspondientes cargas de las PC's.

Denotemos \mathbf{u}_k la columna k de U , \mathbf{v}_k la columna k de V , y d_k el k -ésimo elemento diagonal de la matriz diagonal D . Es conocido (ver por ejemplo [\(Eckart & Young, 1936\)](#) que para cualquier $r \leq K$,

$$\sum_{k=1}^r d_k \mathbf{u}_k \mathbf{v}_k^T = \arg \min \|X - \hat{X}\|^2$$

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:

$$\hat{X} = AD^{-1}B^T + E$$

Así, la estimación de cargas penalizadas en el HJ Biplot implementando *Ridge* está dada por:

$$V_{ridge} = \arg \min \| X - AD^{-1}B^T \|^2 + \lambda \| V \|^2$$

donde, λ es el valor de la regularización *Ridge*

A partir de la regularización *Ridge*, las nuevas cargas de las dimensiones factoriales en el Biplot, se puede obtener mediante la siguiente expresión:

$$V_{ridge} = \frac{V}{1 + \lambda}$$

El método de regularización *Ridge* impone restricción a los elementos de V para obtener el Biplot. Así, el SPARSE HJ Biplot a partir de la regularización *Ridge* puede ser resumido mediante el siguiente algoritmo:

Algoritmo 1: Algoritmo SPARSE HJ Biplot usando la regularización *Ridge*

1. Se tiene una matriz de datos $n \times p$.
 2. Se transforman los datos (centrar o estandarizar).
 3. Se realiza la descomposición en DVS de la matriz original de datos.
 4. Se toma el vector de cargas modificadas V_{ridge} obtenido por el método de regularización *Ridge*.
 5. Se procede a calcular los marcadores filas y marcadores columnas que llevan a la representación Biplot.
 6. Se grafica el *Ridge* HJ Biplot obtenido mediante los pasos anteriores.
-

Se ha implementado una aplicación para obtener el *Ridge* HJ Biplot en el lenguaje R, que da soporte práctico al algoritmo diseñado; y que a la vez viene a llenar un punto abierto para obtener componentes sparse en el contexto del Biplot. Esta aplicación se encuentra alojada en el paquete SparseBiplots.

4.4. LASSO HJ Biplot

LASSO es una técnica regularizada que asegura *sparsity*³ en el modelo de regresión, sujeta a una adecuada elección del parámetro de contracción. Motivado por encontrar una técnica que permitiera no solo reducir coeficientes a cero, sino también seleccionar variables, Tibshirani (1996) propuso LASSO “*Least Absolute Shrinkage and Selection Operator*”, una herramienta que favorece la interpretación de las estimaciones.

La regresión LASSO se formula añadiendo una restricción a los coeficientes del modelo lineal, definida en términos de la norma l_1 de $\beta = \beta_1, \dots, \beta_p$:

$$l_1 = \|\beta\|_1 = \sum_{j=1}^p |\beta_j|$$

Esta restricción es igual al valor absoluto de la magnitud de los coeficientes. El objetivo fundamental de la penalización LASSO es inducir *sparsity* en el modelo de regresión lineal y por tanto conseguir el mayor número de coeficientes con valor cero. LASSO fuerza a cero aquellos coeficientes de regresión de los predictores que están menos correlacionados con la variable respuesta; es decir, los valores de coeficientes pequeños son contraídos a cero de forma inmediata.

Los coeficientes estimados por LASSO, $\hat{\beta}^{LASSO}$, son los valores que minimizan

³ Sparsity (esparsidad): condición de la penalización que se refiere a la selección automática de variables estableciendo que coeficientes suficientemente pequeños serán nulos.

$$\hat{\beta}^{LASSO} = \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j X_{ij} \right)^2 + \lambda \sum_{j=1}^p |\beta_j| = \|Y - \beta X\|^2 + \lambda \sum_{j=1}^p |\beta_j|$$

donde $\lambda \geq 0$ es el parámetro de regularización de LASSO.

Una forma equivalente de escribirlo es:

$$\hat{\beta}^{LASSO} = \arg \min \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j X_{ij} \right)^2 \text{ sujeto a } \sum_{j=1}^p |\beta_j| \leq \lambda$$

La penalización *LASSO* produce un modelo *sparse* pero el número de variables seleccionadas no puede exceder al número de observaciones. Si $p \gg n$, *LASSO* selecciona como máximo n variables. Por otro lado, si hay un grupo de variables entre las cuales las correlaciones por parejas son muy altas, entonces *LASSO* tiende a seleccionar sólo una variable del grupo, sin importar cuál de ellas selecciona. El auge en los últimos años en la investigación y aplicación de técnicas tipo *LASSO* se debe principalmente a la existencia de problemas donde $p \gg n$ y al desarrollo paralelo de algoritmos eficientes ([Tibshirani, 2011](#)).

La regresión lineal *LASSO*, minimiza la suma de los residuos al cuadrado sujeta a una restricción de desigualdad en los parámetros. Esta restricción produce un estimador no lineal en la variable respuesta; en consecuencia, la función objetivo no es diferenciable y la solución *LASSO* se convierte en un problema de programación cuadrática, que debe encontrar solución a través de un algoritmo de optimización. Con el fin de profundizar en algunas de estas soluciones, se han estudiado teórica y experimentalmente distintos algoritmos.

Algunos de estos son: el “*nonnegative garrote*” (Breiman, 1995), el “*Smoothly Clipped Absolute Deviation*” (SCAD) (Fan & Li, 2001), el método “*Coordinate Descent*” (Friedman, Hastie, & Tibshirani, 2010), el “*Hard-thresholding operator* y *Soft-Thresholding operator*” (Donoho & Johnstone, 1994).

A continuación, se detallan estos dos últimos operadores, el “*Hard-thresholding operator* y *Soft-Thresholding operator*”.

El *soft-thresholding operator* (Operador de umbralización suave) se define como:

$$\hat{\beta}^{soft} = S(\hat{\beta}^{OLS}, \lambda) = \begin{cases} \hat{\beta}^{OLS} - \lambda & \text{si } \hat{\beta}^{OLS} > \lambda \\ 0 & \text{si } -\lambda \leq \hat{\beta}^{OLS} \leq \lambda \\ \hat{\beta}^{OLS} + \lambda & \text{si } \hat{\beta}^{OLS} < -\lambda \end{cases}$$

Esta solución presenta dos características:

- (1) Los coeficientes de regresión *LASSO* se contraen a cero, a medida que el valor del parámetro λ aumenta.
- (2) La solución *LASSO* es una función lineal a trozos, respecto al valor de λ . Esto se ilustra en la Figura 4-5.

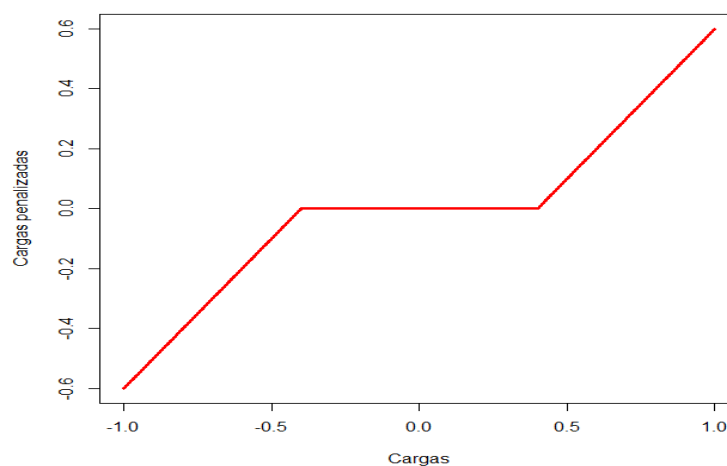


Figura 4-5 Ilustración de la regla del operador soft-thresholding

El *hard-thresholding operator* (operador de umbralización fuerte) se define como

$$\hat{\beta}^{hard} = H(\hat{\beta}^{OLS}, \lambda) = \begin{cases} \hat{\beta}^{OLS} & \text{si } |\hat{\beta}^{OLS}| > \lambda \\ 0 & \text{si } |\hat{\beta}^{OLS}| \leq \lambda \end{cases}$$

Este operador trunca un coeficiente a cero, si el valor absoluto de éste es más pequeño que un valor predeterminado. De esta manera, tanto los valores de los coeficientes positivos como los negativos se "reducen" hacia cero. A diferencia del *soft thresholding*, el *hard thresholding* no es continuo.

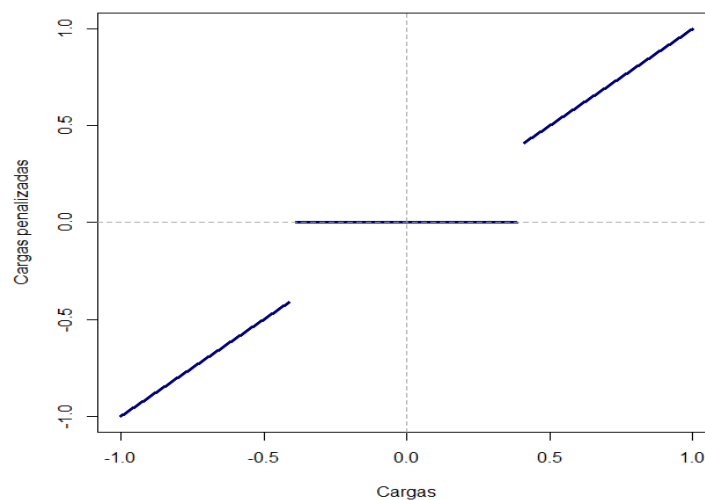


Figura 4-6 Ilustración de la regla del operador *hard-thresholding*

En el contexto del Biplot, el método de penalización *LASSO* combina un modelo de regresión con un procedimiento de contracción de algunos parámetros hacia cero y selección de variables, imponiendo la penalización sobre las cargas de las componentes principales.

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:

$$\hat{X} = AD^{-1}B^T + E$$

Así, la estimación de cargas penalizadas en el HJ Biplot implementando LASSO se plantea como:

$$V_{lasso} = \arg \min \|X - AD^{-1}B^T\|^2 + \lambda \sum_{j=1}^p \|V\|_1$$

donde, λ es el parámetro de penalización.

A medida que aumenta el valor de λ se van anulando más cargas. Primero se anulan aquellas que aportan menos información; es decir, aquellas cargas que tienen un valor más pequeño.

A continuación, resumimos los operadores que se utilizan para definir el Biplot mediante la penalización LASSO.

El *soft-thresholding operator* se define como:

$$V^{soft} = S(V, \lambda) = \begin{cases} V - \lambda & \text{si } V > \lambda \\ 0 & \text{si } -\lambda \leq V \leq \lambda \\ V + \lambda & \text{si } V < -\lambda \end{cases}$$

Las cargas se contraen a cero, a medida que el valor del parámetro λ aumenta.

Cuando las columnas de X son ortogonales, la solución al problema de LASSO es:

$$V^{soft} = \text{sign}(V)(|V| - \lambda)_+$$

El *hard-thresholding operator* se define como

$$V^{hard} = H(V, \lambda) = \begin{cases} V & \text{si } |V| > \lambda \\ 0 & \text{si } |V| \leq \lambda \end{cases}$$

Implementando cualquiera de estos dos operadores, obtenemos la solución LASSO en el HJ Biplot.

La representación *LASSO* HJ Biplot se resume en el siguiente algoritmo:

Algoritmo 2: Algoritmo SPARSE HJ Biplot usando la regularización *LASSO*

1. Se tiene una matriz de datos $n \times p$.
 2. Se transforman los datos (centrar o estandarizar).
 3. Se realiza la descomposición en DVS de la matriz original de datos.
 4. Se penalizan los vectores de cargas de acuerdo al método elegido en la penalización *LASSO*, para obtener las cargas modificadas. En este trabajo se han utilizado dos operadores de umbralización (soft-thresholding y hard-thresholding).
 5. Se toma el vector de cargas modificadas V_{lasso} obtenido.
 6. Se procede a calcular los marcadores filas y marcadores columnas que llevan a la representación Biplot.
 7. Se grafica el *LASSO* HJ Biplot obtenido mediante los pasos anteriores.
-

También se ha implementado una aplicación en el lenguaje R, para dar soporte práctico al algoritmo y obtener el *LASSO* HJ Biplot. Esta se encuentra alojada en el paquete SparseBiplots.

A manera de síntesis, el tipo de restricción entre *Ridge* y *LASSO* deriva importantes conclusiones que presentamos en la siguiente tabla:

Tabla 7: Comparación teórica entre la restricción Ridge y LASSO en el Biplot

Regularización <i>Ridge</i>	Regularización <i>LASSO</i>
Usa la norma l_2	Usa la norma l_1
Es un método que NO selecciona variables.	Es un método de selección de variables.
No penaliza las cargas.	Penaliza las cargas.
Contrae todas las cargas hacia cero, pero sin llegar a hacerlas nulos.	Produce estimaciones nulas para algunas cargas y no nulas para otras.

En los últimos años se han presentado algunas generalizaciones y extensiones de las técnicas presentadas anteriormente. Éstas buscan retener las ventajas de *LASSO* como método de estimación y selección de variables, y al mismo tiempo mantener la principal propiedad de la regularización *Ridge* (encoger los vectores de cargas).

4.5. *Elastic Net* HJ Biplot

En el análisis de regresión, [Zou & Hastie \(2005\)](#) proponen el método *Elastic Net*, que combina los métodos de regularización *Ridge* y *LASSO*. Por tanto, es un método de regularización que penaliza el tamaño de los coeficientes de regresión en base a ambas normas l_1 y l_2 .

Los coeficientes estimados por medio del método *Elastic Net*, $\hat{\beta}^{elasticnet}$, son los valores que minimizan:

$$\hat{\beta}^{elasticnet} = \arg \min_{\beta} \|y_i - X\beta\|^2 + \lambda_2 \sum_{j=1}^p \beta_j^2 + \lambda_1 \sum_{j=1}^p |\beta_j|$$

donde $\lambda_1 > 0$ y $\lambda_2 > 0$ son ambos parámetros de complejidad.

Por un lado, el término

$$\lambda_1 \sum_{j=1}^p |\beta_j|$$

apunta a soluciones *sparse*.

Por otro lado, el término

$$\lambda_2 \sum_{j=1}^p \beta_j^2$$

sugiere que predictores altamente correlacionados obtengan coeficientes estimados similares.

De forma equivalente, se puede escribir:

$$\hat{\beta}^{elasticnet} = \arg \min \|y_i - X\beta\|^2$$

sujeto a

$$\sum_{j=1}^p \beta_j^2 \leq \lambda \text{ y } \sum_{j=1}^p |\beta_j| \leq \lambda$$

A partir de esta restricción, [Zou & Hastie \(2005\)](#) proponen calcular los nuevos β mediante la siguiente expresión:

$$V^{soft} = \text{sign}(V) \frac{(|V - \lambda_1|)_+}{1 + \lambda_2}$$

donde se combina las normas l_1 y l_2 utilizando el operador soft-thresholding.

El método de regularización Elastic Net también puede ser implementado en el Biplot, combinando LASSO y Elastic Net para derivar cargas modificadas. 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:

$$\hat{X} = AD^{-1}B^T + E$$

A partir del método de regularización *Elastic Net* se derivan cargas modificadas para el Biplot, de la forma:

$$V_{elasticnet} = \arg \min \|X - AD^{-1}B^T\|^2 + \lambda_2 \sum_{j=1}^p V_j^2 + \lambda_1 \sum_{j=1}^p |V_j|$$

Considerando los primeros k ejes factoriales, se definen las matrices:

$$A_{p \times k} = [\alpha_1, \alpha_2, \dots, \alpha_k] \text{ y } B_{p \times k} = [\beta_1, \beta_2, \dots, \beta_k].$$

Para cualquier $\lambda_2 > 0$, sea:

$$(\hat{A}, \hat{B}) = \underset{A, B}{\operatorname{argmin}} \sum_{i=1}^n \|x_i - AB^T x_i\|^2 + \lambda_2 \sum_{j=1}^k \|\beta_j\|^2 + \lambda_{1,j} \sum_{j=1}^k \|\beta_j\|_1$$

sujeto a $A^T A = I_{K \times K}$

donde $\lambda_{1,j}$ es el parámetro de penalización *LASSO* para inducir esparsidad

λ_2 es el parámetro de regularización para contraer las cargas

Este problema se puede resolver alternando la optimización sobre A y B usando el algoritmo LARS-EN (Zou & Hastie, 2005).

Para A fija: **B** se obtiene resolviendo el siguiente problema:

$$\begin{aligned} \hat{\beta}_j &= \underset{\beta_j}{\operatorname{argmin}} \|X\alpha_j - X\beta_j\|^2 + \lambda_2 \|\beta_j\|^2 + \lambda_{1,j} \|\beta_j\|_1 \\ &= (\alpha_j - \beta_j)^T X^T X (\alpha_j - \beta_j) + \lambda_2 \|\beta_j\|^2 + \lambda_{1,j} \|\beta_j\|_1 \end{aligned}$$

donde cada $\hat{\beta}_j$ es un estimador Elastic Net.

Para B fija: **A** se ignora la parte de penalización y se minimizando

$$\underset{A}{\operatorname{argmin}} \sum_{i=1}^n \|x_i - AB^T x_i\|^2 = \|X - XBA^T\|^2$$

sujeto a $A^T A = I_{K \times K}$

Este es un problema de Procrustes, y la solución es proporcionada por la *DVS*,

$(X^T X)B = UDV^T$ y se toma $\hat{A} = UV^T$.

Recientemente Erichson, Zheng, & Manohar (2018) plantean otra solución al problema de optimización, utilizando el método de proyección variable.

Para ajustar los parámetros λ_2 y λ_1 no existen métodos claros y establecidos.

Se sugiere probar varias combinaciones y elegir la que proporcione equilibrio entre la varianza explicada y la esparsidad, dando preferencia a la varianza.

Los pasos para la implementación del método de regularización Elastic Net en el HJ Biplot se detallan en el siguiente algoritmo.

Algoritmo 3: Algoritmo SPARSE HJ Biplot usando la regularización Elastic Net

1. Se tiene una matriz de datos $n \times p$.
2. Se fija un valor de tolerancia (1×10^{-5}).
3. Se transforman los datos (centrar o estandarizar).
4. Se realiza la descomposición en DVS de la matriz original de datos.
5. Se toma A como las cargas de las primeras k componentes $V[1:k]$.
6. Se calcula β_j mediante:

$$\beta_j = (\alpha_j - \beta_j)^T X^T X (\alpha_j - \beta_j) + \lambda_2 \|\beta_j\|^2 + \lambda_{1,j} \|\beta_j\|_1$$

7. Se actualiza A mediante la DVS de $X^T X \beta$:

$$X^T X \beta = U D V^T \rightarrow A = U V^T$$

8. Se actualiza la diferencia entre A y B

$$dif_{AB} = \frac{1}{p} \sum_{i=1}^p \frac{1}{|\beta_i|^2 |\alpha_i|^2} \sum_{j=1}^m \beta_{ij} - \alpha_{ij}$$

9. Se repiten los pasos 4,5 y 6 hasta que $dif_{AB} < tolerancia$
 10. Se normalizan las columnas $\hat{V}_j^{EN} = \frac{\beta_j}{\|\beta_j\|}, j = 1, \dots, k$
 11. Se procede a calcular los marcadores filas y marcadores columnas que llevan a la representación Biplot.
 12. Se grafica el *Elastic Net* HJ Biplot obtenido mediante los pasos anteriores.
-

En este caso, también se ha diseñado una aplicación en el lenguaje R, para obtener la representación *Elastic Net* HJ Biplot. Esta aplicación forma parte del paquete denominado "Sparsebiplots".

En el siguiente esquema se presentan los pasos que describen la aplicación de los métodos de regularización sobre el HJ Biplot, que llevan a la obtención de ejes modificados.

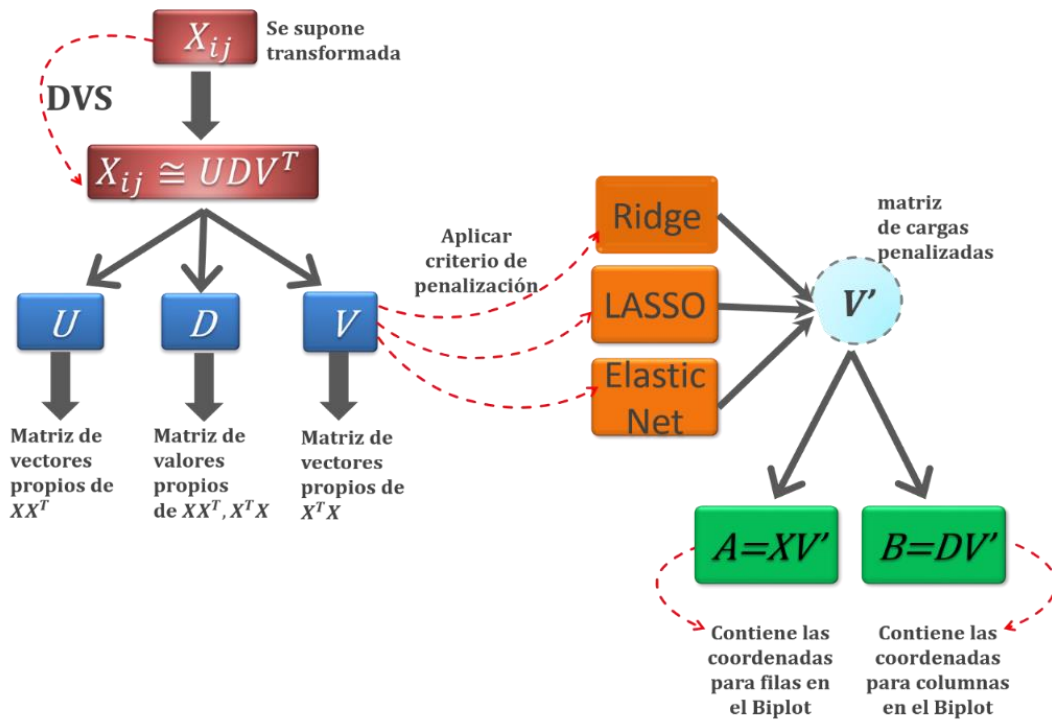


Figura 4-7. Esquema del Sparse HJ Biplot

4.6 Varianza Total Explicada por las componentes sparse

En el Biplot se obtienen *cargas ortogonales* y componentes *no correlacionadas* a partir de la transformación de las variables originales. Ambas condiciones se cumplen ya que dada la matriz de covarianza $\hat{\Sigma} = X^T X$ entonces $V^T V = I$ y $V^T \hat{\Sigma} V$ es una matriz diagonal. Tomando \hat{Z} como las CP's Sparse estimadas, la varianza total explicada se determina por $\text{tr}(\hat{Z}^T \hat{Z})$.

Las CP's Sparse son capaces de producir cargas ortogonales, pero sus componentes están correlacionadas (Jolliffe et al., 2003). Bajo estas condiciones, no es apropiado calcular la varianza explicada como se hace en el Biplot ordinario porque se estaría sobreestimando la varianza verdadera; puesto que la dimensión absorbe la información propia de ella, más la

compartida con el resto de los ejes. En el análisis se quiere que cada componente sea independiente de las anteriores; por lo tanto, de existir dependencia lineal entre componentes, ésta debe ser eliminada.

Dentro de las alternativas para obtener la varianza total ajustada y dar solución a este problema en componentes Sparse, [Hui Zou et al. \(2006\)](#) sugieren usar vectores de proyección para remover la dependencia lineal. Estos autores denotan $\hat{Z}_{j \cdot 1, \dots, j-1}$ el residual después de ajustar \hat{Z} por $\hat{Z}_{1, \dots, j-1}$, esto es:

$$\hat{Z}_{j \cdot 1, \dots, j-1} = \hat{Z}_j - H_{1, \dots, j-1} \hat{Z}_j$$

donde $H_{1, \dots, j-1}$ es la matriz de proyección sobre $\{\hat{Z}_i\}_1^{j-1}$.

Entonces la varianza ajustada de \hat{Z}_j es $\|\hat{Z}_{j \cdot 1, \dots, j-1}\|^2$ y la varianza total explicada queda definida como $\sum_{j=1}^k \|\hat{Z}_{j \cdot 1, \dots, j-1}\|^2$. Cuando las CP's Sparse estimadas \hat{Z} estén no correlacionadas, dicha fórmula coincide con $\text{tr}(\hat{Z}^T \hat{Z})$.

La descomposición QR , donde Q es una matriz ortonormal y R una matriz triangular superior, es otra forma más sencilla de estimar la varianza ajustada.

Tomando $\hat{Z} = QR$ se tiene que $\|\hat{Z}_{j \cdot 1, \dots, j-1}\|^2 = R_{jj}^2$.

Luego, la varianza total explicada es

$$\sum_{j=1}^k R_{jj}^2$$

4.7 Paquete en R para Sparse Biplot: “SparseBiplots”

En el interés de proporcionar una herramienta que implemente los algoritmos descritos en las secciones anteriores, se ha creado un paquete en el lenguaje R, disponible en <https://cran.r-project.org/web/packages/SparseBiplots/index.html>. Este paquete, denominado SparseBiplots, permite utilizar el algoritmo que desea con sus propios datos y generar un gráfico con la representación de interés; además, permite obtener resultados numéricos como las coordenadas de los individuos y de las variables, los valores propios, las cargas factoriales y la varianza explicada por cada componente.

Package: SparseBiplots

Type: Package

Title: HJ Biplot using diferents ways of penalization

Version: 1.0.0

Authors: Mitzi Cubilla Montilla <mitzi@usal.es>,

Ana Belén Nieto Librero <ananieto@usal.es>,

Purificación Galindo Villardón <pgalindo@usal.es>

Maintainer: Mitzi Cubilla Montilla <mitzi@usal.es>

Depends: R (>= 2.10), sparsepca

Description: The SparseBiplots package contains a set of functions that allow to represent multivariate on a subspace of low dimension, in such a way that most of the variability of the information is captured. This representation is carried out through the HJ Biplot methodology. A first method performs Galindo's HJ-Biplot (1986). Then, the package implements three new techniques and constructs in each case the HJ Biplot, adapting restrictions to contract and / or produce zero charges in the main components, based on the regularization theories. It implements three methods of regularization: *Ridge*, *LASSO* and *Elastic Net*. The functions are as follows: *Ridge_HJBiplot*, *LASSO_HJBiplot* and *ElasticNet_HJBiplot*.

Encoding: UTF-8

LazyData: true

Imports: sparsepca, Rdpack

RoxygenNote: 6.1.1

RdMacros: Rdpack

Built: R 3.4.2: 2019-06-21 13:10:00 UTC

HJ Biplot

HJ Biplot	HJ Biplot based on Principal Components Analysis
-----------	--

Description

This function performs the representation of HJ Biplot (Galindo, 1986).

Usage

```
HJBiplot (X, transform_data = 'scale', ind_name=FALSE, vec_name = TRUE)
```

Details

Algorithm used to construct the HJ Biplot. The Biplot is obtained as result of the configuration of markers for individuals and markers for variables in a reference system defined by the factorial axes resulting from the Decomposition in Singular Values (DVS).

Value

HJBiplot returns a list containing the following components:

loadings	array_like; the loadings of the principal components.
coord_ind	array_like; matrix with the coordinates of individuals.
coord_var	array_like; matrix with the coordinates of variables.
eigenvalues	array_like; vector with the eigenvalues.
explvar	array_like; an vector containing the proportion of variance explained by the first 1, 2,..,k principal components obtained.

Author(s)

Mitzi Cubilla Montilla, Ana Belén Nieto Librero y Purificación Galindo Villardón

References

- Gabriel, K. R. (1971). The Biplot graphic display of matrices with applications to principal components analysis. *Biometrika*, 58(3), 453-467.
- Galindo, M. P. (1986). Una alternativa de representacion simultanea: HJ-Biplot. *Questiio*, 10(1), 13-23.

Examples

```
data(mtcars)
HJBiplot(mtcars, transform_data = 'scale', ind_name = TRUE)
```

Ridge_HJBiplot

<i>Ridge</i> HJ Biplot	<i>Ridge</i> HJ Biplot based on Principal Components Analysis
------------------------	---

Description

This function performs the representation of the SPARSE HJ Biplot applying the *Ridge* regularization, on the original data matrix, implementing the norm L2.

Usage

```
Ridge_HJBiplot (X, lambda, transform_data = 'scale', ind_name=FALSE,
vec_name = TRUE)
```

Arguments

X	array_like; A data frame which provides the data to be analyzed. All the variables must be numeric.
lambda	float; Tuning parameter for the <i>Ridge</i> penalty
ind_name	bool; If it is TRUE it prints the name for each row of X. If it is FALSE (default) does not print the names.
vec_name	bool; If it is TRUE (default) it prints the name for each column of X. If it FALSE does not print the names.
Transform_data	character; A value indicating whether the columns of X (variables) should be centered or scaled. Options are: "center" that removes the columns means and "scale" that removes

the columns means and divide by its standard deviation.
For default it is "scale".

Details

Algorithm used to contract the loads of the main components towards zero, but without achieving the nullity of any. If the penalty parameter is less than or equal to $1e-4$ the result is like Galindo's HJ Biplot (1986).

Value

Ridge_HJBiplot returns a list containing the following components:

loadings	array_like; penalized loadings, the loadings of the sparse principal components.
coord_ind	array_like; matrix with the coordinates of individuals.
coord_var	array_like; matrix with the coordinates of variables.
eigenvalues	array_like; vector with the eigenvalues penalized.
explvar	array_like; an vector containing the proportion of variance explained by the first 1, 2,..,k sparse principal components obtained.

Author(s)

Mitzi Cubilla Montilla, Ana Belén Nieto Librero y Purificación Galindo Villardón

References

- Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55-67.
- Galindo, M. P. (1986). Una alternativa de representación simultánea: HJ-Biplot. *Questiio*, 10(1), 13-23.
- Zou, H., Hastie, T. and Tibshirani, R. (2004) Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2), 265-286

Examples

```
data(mtcars)
```

```
Ridge_HJBiplot (mtcars, 0.2, transform_data = 'scale', ind_name = TRUE)
```

LASSO_HJBiplot

LASSO HJ Biplot LASSO HJ Biplot based on Principal Components Analysis

Description

This function performs the representation of the SPARSE HJ Biplot applying the *LASSO* regularization, on the original data matrix, implementing the norm L1.

Usage

```
LASSO_HJBiplot(X, lambda, transform_data = 'scale', operator = 'Hard-Thresholding', ind_name=FALSE, vec_name = TRUE)
```

Arguments

X	array_like; A data frame which provides the data to be analyzed. All the variables must be numeric.
lambda	float; Tuning parameter for the <i>LASSO</i> penalty
operator	character; The operator used to solve the norm L1.
ind_name	bool; If it is TRUE it prints the name for each row of X. If it is FALSE (default) does not print the names.
vec_name	bool; If it is TRUE (default) it prints the name for each column of X. If it is FALSE does not print the names.
Transform_data	character; A value indicating whether the columns of X (variables) should be centered or scaled. Options are: "center" that removes the columns means and "scale" that removes the columns means and divide by its standard deviation. For default is "scale".

Details

Algorithm that performs a procedure of contraction and selection of variables. *LASSO* imposes a penalty that causes the charges of some components to be reduced to zero. By producing zero loadings for some components and not zero

for others, the *LASSO* technique performs selection of variables. As the value of the penalty approaches one, the loadings approach zero.

Value

LASSO_HJBiplot returns a list containing the following components:

loadings	array_like; penalized loadings, the loadings of the sparse principal components.
n_ceros	array_like; number of loadings equal to zero in each component.
coord_ind	array_like; matrix with the coordinates of individuals.
coord_var	array_like; matrix with the coordinates of variables.
eigenvalues	array_like; vector with the eigenvalues penalized.
explvar	array_like; an vector containing the proportion of variance explained by the first 1, 2,..,k sparse principal components obtained.

Author(s)

Mitzi Cubilla Montilla, Ana Belén Nieto Librero y Purificación Galindo Villardón

References

- Galindo, M. P. (1986). Una alternativa de representacion simultanea: HJ-Biplot. *Questiío*, 10(1), 13-23.
- Tibshirani, R. (1996). Regression shrinkage and selection via the *LASSO*. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267-288.
- Tibshirani, R. (2011). Regression shrinkage and selection via the *LASSO*: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3), 273-282.

Examples

```
data(mtcars)
LASSO_HJBiplot(mtcars, 0.2, transform_data = 'scale', operator = 'Hard-Thresholding',ind_name = TRUE)
```

ElasticNet_HJBiplot

Elastic Net HJ Biplot *Elastic Net* HJ Biplot based on Principal Components Analysis

Description

This function is a generalization of the *Ridge* regularization method and the *LASSO* penalty. Realizes the representation of the SPARSE HJ Biplot through a combination of *LASSO* and *Ridge*, on the data matrix. This means that with this function you can eliminate weak variables completely as with the *LASSO* regularization or contract them to zero as in *Ridge*.

Usage

```
ElasticNet HJBiplot (X, lambda = 1e-04, alpha = 1e-04, transform_data = 'scale',  
ind_name=FALSE, vec_name = TRUE)
```

Arguments

X	array_like; A data frame with the information to be analyzed
lambda	float; Tuning parameter of the <i>LASSO</i> penalty. Higher values lead to sparser components.
alpha	float; Tuning parameter of the <i>Ridge</i> shrinkage
transform_data	character; A value indicating whether the columns of X (variables) should be centered or scaled. Options are: "center" or "scale". For default is "scale".
ind_name	bool; Logical value, if it is TRUE it prints the name for each row of X. If it is FALSE (default) does not print the names.
vec_name	bool; Logical value, if it is TRUE (default) it prints the name for each column of X. If it is FALSE does not print the names.

Details

Algorithm used to perform automatic selection of variables and continuous contraction simultaneously. With this method, the model obtained is simpler and more interpretable.

It is a particularly useful method when the number of variables is much greater than the number of observations.

Value

ElasticNet_HJBiplot returns a list containing the following components:

loadings	array_like; penalized loadings, the loadings of the sparse principal components.
n_ceros	array_like; number of loadings equal to zero in each component.
coord_ind	array_like; matrix with the coordinates of individuals.
coord_var	array_like; matrix with the coordinates of variables.
eigenvalues	array_like; vector with the eigenvalues penalized.
explvar	array_like; an vector containing the proportion of variance explained by the first 1, 2,..,k sparse principal components obtained.

Author(s)

Mitzi Cubilla Montilla, Ana Belén Nieto Librero y Purificación Galindo Villardón

References

- Galindo, M. P. (1986). Una alternativa de representacion simultanea: HJ-Biplot. *Questiío*, 10(1), 13-23.
- Erichson, N. B., Zheng, P., Manohar, K., Brunton, S. L., Kutz, J. N., & Aravkin, A. Y. (2018). Sparse principal component analysis via variable projection. arXiv preprint arXiv:1804.00341.
- Zou, H., & Hastie, T. (2005). Regularization and variable selection via the *Elastic Net*. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2), 301-320.

Examples

```
data(mtcars)
ElasticNet_HJBiplot (mtcars, 0.1, 0.01, transform_data = 'scale',
ind_name=TRUE)
```

Arguments

X	array_like; A data frame which provides the data to be analyzed. All the variables must be numeric.
ind_name	bool; If it is TRUE it prints the name for each row of X. If it is FALSE (default) does not print the names.
vec_name	bool; If it is TRUE (default) it prints the name for each column of X. If it FALSE does not print the names.
Transform_data	character; A value indicating whether the columns of X (variables) should be centered or scaled. Options are: "center" that removes the columns means and "scale" that removes the columns means and divide by its standard deviation. For default is "scale".

4.8 Aplicaciones Prácticas

A continuación, se presentan los resultados de la aplicación práctica de los argumentos teóricos expuestos previamente. Para el desarrollo de este apartado utilizamos dos conjuntos de datos, uno en el que se tienen más individuos que variables ($n > p$), y otro en que hay más variables que individuos ($p > n$); con el interés de explorar las herramientas formuladas y examinar la capacidad de utilizarlas en diferentes contextos.

Primero se analiza la información que se presentó en la sección 3.2 (Tabla 3) correspondientes al conjunto de datos que miden el *Índice para una Vida Mejor (Better Life Index)* en los países pertenecientes a la OCDE. Recordando que esta información recoge observaciones de 38 países sobre 24 variables.

Utilizamos otro conjunto de datos que contiene las 500 mayores compañías a nivel mundial según Fortune Global 500 (<http://fortune.com/global500>). La muestra corresponde a 201 compañías de 29 países que divulgaron el informe sobre *Responsabilidad Social Corporativa* en el año 2015, de conformidad con el Global Reporting Initiative (GRI). Concretamente, se evaluaron en total 25 *indicadores sociales* clasificados en 4 categorías: 9 relativos a “Derechos Humanos” (HR), 6 sobre “Prácticas Laborales” (LA), 5 sobre “Responsabilidad sobre Productos” (RP), y 5 relacionados a “Sociedad” (SO). Estos indicadores se analizan en conjunto a las dimensiones culturales de Hofstede (2011): colectivismo (collectivism), feminismo (femenity), distancia al poder (LOW PDI), tolerancia a la incertidumbre (U_Tolerance) y visión a largo plazo (Long_Term). El programa utilizado en ambos casos es R-TEAM (2014).

Con respecto al primer conjunto de datos, en la Figura 4-9 se presenta el Biplot obtenido mediante el método de regularización *Ridge*, que visualmente es muy similar al HJ-Biplot (Figura 4-8) puesto que este algoritmo solamente contrae las cargas hacia cero, manteniendo la proporción de varianza explicada igual que en el HJ Biplot.

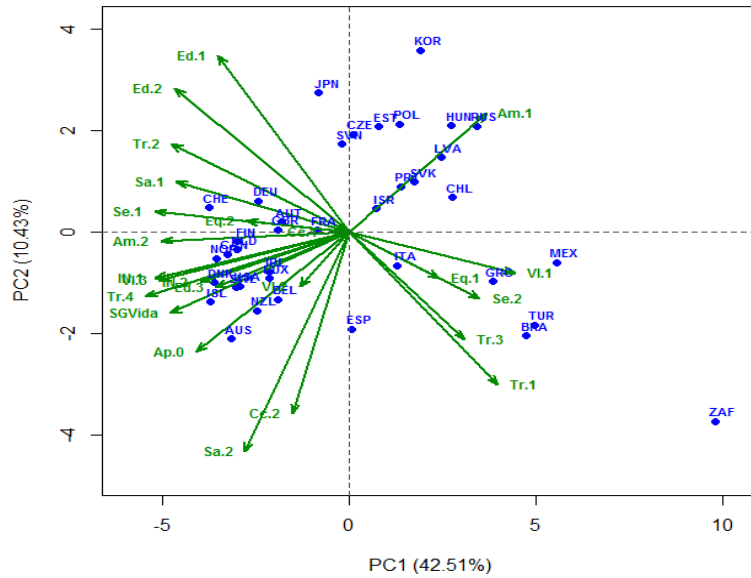


Figura 4-8 HJ Biplot. Datos sobre el índice para una vida mejor

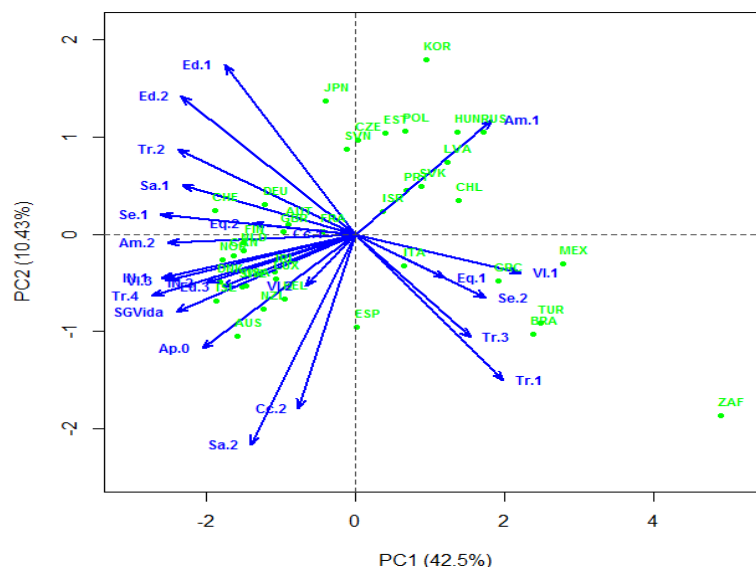


Figura 4-9 Ridge HJ Biplot. Datos sobre el índice para una vida mejor

Analizando estos resultados gráficos, en primer lugar, se destaca que casi todas las variables que miden el índice para una vida mejor se encuentran asociadas a la primera componente del plano factorial. Los países nórdicos y algunos países europeos son los que perciben mejores condiciones de vida y mayor bienestar.

También se observa que la satisfacción global ante la vida (SGVida) está relacionada con una gran parte de los índices para una vida mejor en los países miembros de la OCDE. La variable más altamente asociada al índice es el ingreso medio. Por otro lado, este nivel de satisfacción revela una relación inversa con aspectos asociados al empleo, como seguridad en el empleo (TR.1), jornada laboral muy larga (Eq.1), tasa de desempleo a largo plazo (TR.3); así como también con la tasa de homicidios (Se.2). En cuanto a la tasa de homicidios, México, Brasil y Grecia encabezan las posiciones.

Para favorecer la interpretación de los datos y evaluar qué factores parecen ser más relevantes en el estudio del índice para una vida mejor, se realiza el análisis mediante las técnicas de regularización *LASSO* y *Elastic Net*. Las variables más relevantes mediante estas técnicas quedan mejor identificadas en la Tabla 8.

En la Figura 4-10 se representa el biplot mediante el método de regularización *LASSO* (*soft thresholding operator*) donde se puede apreciar que algunas cargas factoriales se contraen a cero, quedando representadas sobre los ejes factoriales.

Las variables, logro educativo (Ed.1) y competencia de los estudiantes (Ed.2) son las que aportan más información, ya que sus vectores son los más largos.

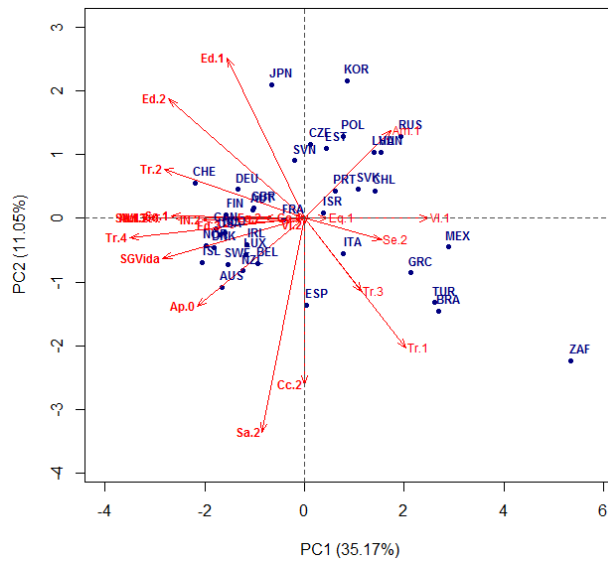


Figura 4-10 LASSO HJ Biplot. Datos sobre el Índice para una vida mejor

En la Figura 4-11 se presenta el Biplot obtenido mediante el método de penalización *Elastic Net*, en donde queda de manifiesto que algunas cargas se anulan totalmente, tanto en la primera componente, como en la segunda, quedando representados sobre los ejes factoriales.

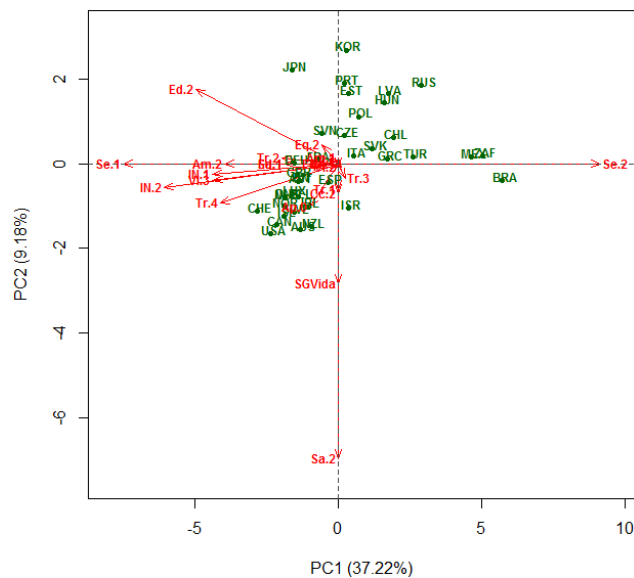


Figura 4-11 Elastic Net HJ Biplot. Datos sobre el Índice para una vida mejor

Tabla 8: Cargas de las primeras tres componentes principales obtenidas mediante HJ Biplot, Disjoint Biplot y los métodos de penalización Ridge, LASSO y Elastic Net (datos del Índice para una vida mejor)

VAR	<u>HJ-Biplot</u>			<u>Disjoint HJ Biplot</u>			<u>Ridge HJ Biplot</u>			<u>LASSO_{soft} HJ Biplot</u>			<u>Elastic Net HJ Biplot</u>		
	CP2	CP3		CP1	CP2	CP3	CP1	CP2	CP3	CP1	CP2	CP3	CP1	CP2	CP3
VI.1	0.228	-0.084	0.086	0.000	1	0.000	0.114	-0.042	0.043	0.1281	0.000	0.000	0.000	0.0068	0.0078
VI.2	-0.068	-0.110	0.044	0.000	0.000	1	-0.034	-0.055	0.022	0.000	-0.0105	0.000	0.000	-0.0075	0.000
VI.3	-0.262	-0.098	0.030	1	0.000	0.000	-0.131	-0.049	0.015	-0.1616	0.000	0.000	-0.2258	-0.0422	0.000
IN.1	-0.266	-0.093	-0.016	0.000	0.000	1	-0.133	-0.047	-0.008	-0.1671	0.000	0.000	-0.2260	-0.0256	0.000
IN.2	-0.206	-0.103	-0.040	0.000	1	0.000	-0.103	-0.051	-0.020	-0.1052	-0.0029	0.000	-0.3135	-0.0599	0.000
Tr.1	0.204	-0.312	0.286	0.000	1	0.000	0.102	-0.156	0.143	0.1047	-0.2114	0.1856	0.000	-0.0646	0.4303
Tr.2	-0.244	0.180	-0.182	1	0.000	0.000	-0.122	0.090	-0.091	-0.1450	0.0807	-0.0818	-0.1020	0.0129	-0.2650
Tr.3	0.160	-0.220	0.456	1	0.000	0.000	0.080	-0.110	0.228	0.0589	-0.1197	0.3561	0.0127	-0.0364	0.6025
Tr.4	-0.280	-0.132	0.018	1	0.000	0.000	-0.140	-0.066	0.009	-0.1805	-0.0319	0.000	-0.2123	-0.0977	0.000
Ap.0	-0.210	-0.244	0.121	1	0.000	0.000	-0.105	-0.122	0.061	-0.1109	-0.1442	0.0212	-0.0521	-0.1054	0.000
Ed.1	-0.180	0.362	0.083	1	0.000	0.000	-0.090	0.181	0.041	-0.803	0.2615	0.000	-0.0957	0.000	-0.0436
Ed.2	-0.240	0.294	0.156	1	0.000	0.000	-0.120	0.147	0.078	-0.1404	0.1949	0.0558	-0.2564	0.1832	0.000
Ed.3	-0.184	-0.112	0.060	1	0.000	0.000	-0.092	-0.056	0.030	-0.0837	-0.0122	0.000	0.000	0.000	0.000
Am.1	0.190	0.244	0.057	1	0.000	0.000	0.095	0.122	0.029	0.0891	0.1432	0.000	0.000	0.0134	0.000
Am.2	-0.258	-0.018	0.010	1	0.000	0.000	-0.129	-0.009	0.005	-0.1588	0.000	0.000	-0.2053	0.000	0.000

Cc.1	-0.026	0.001	-0.324	1	0.000	0.000	-0.013	0.000	-0.162	0.000	0.000	-0.2245	0.000	0.000	0.000
Cc.2	-0.078	-0.372	-0.069	0.000	0.000	1	-0.039	-0.186	-0.034	0.000	-0.2718	0.000	0.000	-0.0733	0.000
Sa.1	-0.238	0.104	-0.030	0.000	1	0.000	-0.119	0.052	-0.015	-0.1379	0.0037	0.000	-0.0954	0.000	0.000
Sa.2	-0.144	-0.448	-0.070	1	0.000	0.000	-0.072	-0.224	-0.035	-0.0436	-0.3490	0.000	0.000	-0.7255	0.0435
Se.1	-0.268	0.043	0.113	1	0.000	0.000	-0.134	0.021	0.057	-0.1678	0.000	0.0134	-0.3868	0.000	0.000
Se.2	0.180	-0.132	-0.301	1	0.000	0.000	0.090	-0.066	-0.150	0.0795	-0.0352	-0.2009	0.4721	0.000	-0.1244
Eq.1	0.124	-0.093	-0.375	1	0.000	0.000	0.062	-0.047	-0.187	0.0241	0.000	-0.2749	-0.0160	0.000	-0.0156
Eq.2	-0.142	0.023	0.454	0.000	0.000	0.000	-0.071	0.012	0.227	-0.0417	0.000	0.3532	-0.0288	0.0473	0.0635
SGVida	-0.246	-0.166	-0.216	0.000	0.000	0.000	-0.122	-0.083	-0.101	-0.1464	-0.0652	-0.1163	0.000	-0.2955	-0.3171
sparsity	0	0	0	9	20	21	0	0	0	3	8	13	9	8	14
% var. expl.	42.51	10.43	9.34	31.74	11.00	5.68	42.51	10.43	9.34	35.17	11.05	11.63	37.22	9.18	13.08
% total		62.28			48.42			62.28			57.85			59.48	

La varianza explicada en las tres primeras CP's es menor en cada uno de los algoritmos SPARSE Biplot que en el HJ Biplot. De esta forma las cargas que no son importantes para la componente principal son nulas y la interpretación de los resultados es mucho más sencilla al saber exactamente que variables aportan información a cada componente en el Biplot. Comparando los resultados de las tres (3) componentes principales obtenidas a través del HJ Biplot, el *Disjoint Biplot* y las tres técnicas *propuestas SPARSE Biplot (Ridge, LASSO y Elastic Net)*, se observa que en el caso de la restricción *Ridge*, las cargas se reducen a la mitad con respecto al HJ-Biplot usual. En el *Disjoint Biplot* hay que recordar que cada variable contribuye solo a una componente. En cambio, el efecto de las restricciones *LASSO y Elastic Net* en las componentes es más evidente, sobre todo gráficamente, de acuerdo a l criterio de penalización implementado.

A continuación, se ejecutan los algoritmos sobre el conjunto de datos relativos a indicadores sociales reportados según el Global Reporting Initiative (GRI), en combinación con las dimensiones culturales de los países ([Hofstede, 2011](#)). En este caso, utilizamos una muestra de 201 empresas pertenecientes a 29 países de diferentes regiones del mundo. El análisis se ha realizado solamente con los indicadores GRI relativos a la dimensión social *que son omitidos* con mayor frecuencia en las memorias de Responsabilidad Social Corporativa. El análisis del HJ Biplot está publicado en [Cubilla-Montilla et al. \(2019\)](#) y va a ser comparado y analizado con el Disjoint Biplot y los resultados del HJ Biplot obtenidos mediante las regularizaciones: Ridge, LASSO y Elastic Net. A través de estos métodos evaluamos la influencia que los valores culturales tienen sobre la divulgación de los indicadores. De esta manera podemos visualizar las relaciones entre indicadores y caracterizar a las compañías de acuerdo al desarrollo cultural de su país de origen y su compromiso con la divulgación de información.

Desde el punto de vista empírico, los análisis realizados nos permiten afirmar que el 48% de las empresas de mayor tamaño a nivel mundial, líderes en sostenibilidad y la utilización de las guías GRI en la elaboración de sus memorias de sostenibilidad, no reportan el 52% de los indicadores sociales que se recogen en la guía G4. En especial, existe una omisión en la revelación de los indicadores relativos a Derechos Humanos, situándose el porcentaje de opacidad en el 75%. Los primeros dos ejes explican aproximadamente el 55% de la variabilidad de los datos; permitiendo utilizar el plano factorial 1-2 para representar la información en el HJ-Biplot (Figura 4-12).

El primer valor propio 20.43 es significativamente más alto que el segundo valor propio (2.53), lo que expresa que la primera dimensión representa la mayor parte de la información. Analizando los resultados a través del HJ-Biplot (Figura 4-12), se observa que el primer cuadrante está formado por las compañías con mayor divulgación de la sostenibilidad en Prácticas laborales y trabajo digno, así como también en Indicadores de Sociedad; estas empresas se encuentran localizadas en países dominados por una cultura “colectivista”. En el cuarto cuadrante, se localizan compañías cuyos reportes se caracterizan principalmente por la divulgación de indicadores relacionados a Derechos humanos, lo cual está asociado de manera positiva con culturas de “baja distancia al poder” y “tolerancia a la incertidumbre”. El segundo y tercer cuadrante, reúne un conjunto de empresas que, a la luz de los datos, son las menos sostenibles. Refleja que es un grupo poco activo en el reporte de indicadores sociales; sin embargo, son empresas dominadas por países que posee un sistema cultural “feminista” y con “visión a largo plazo”. En concreto, los resultados evidencian que las empresas ubicadas en países con sistemas culturales asociados a mayores niveles de colectivismo, tolerancia a la incertidumbre y menor distancia al poder, tienden a divulgar indicadores sociales debido a que los grupos de interés presentan una mayor preocupación por el bienestar social común. Por el contrario, las presiones normativas, asociadas a dimensiones culturales de feminidad y orientación a largo plazo carecen de impacto en la divulgación de indicadores sociales. Evidentemente, las empresas que revelan mayor volumen de información no reportan todos los indicadores necesarios para conocer su impacto y poder evaluar riesgos.

En la Figura 4-13 se presenta el resultado obtenido mediante *Ridge HJ Biplot*. La interpretación se da en los mismos términos que lo expresado anteriormente para el HJ Biplot, ya que la distribución de las variables se mantiene, mientras que las coordenadas de los individuos y de las variables se contraen, lo que se puede reconocer claramente en los ejes factoriales.

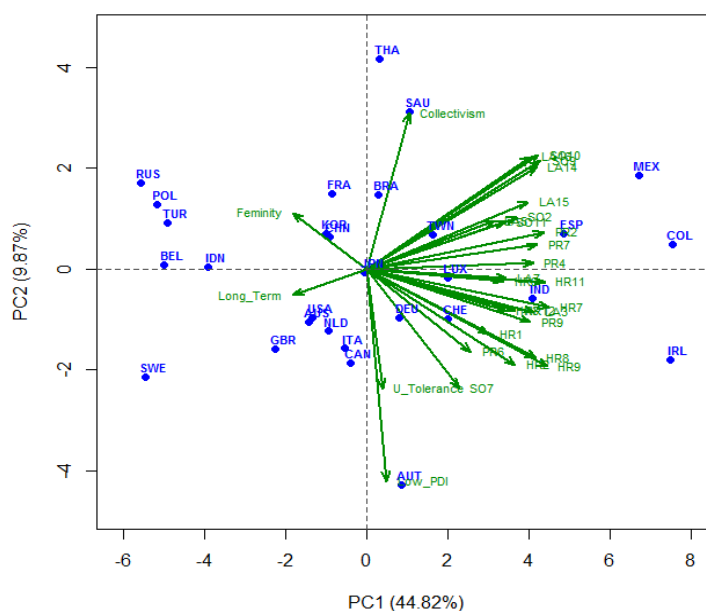


Figura 4-12 HJ Biplot. Datos sobre indicadores sociales y dimensiones culturales

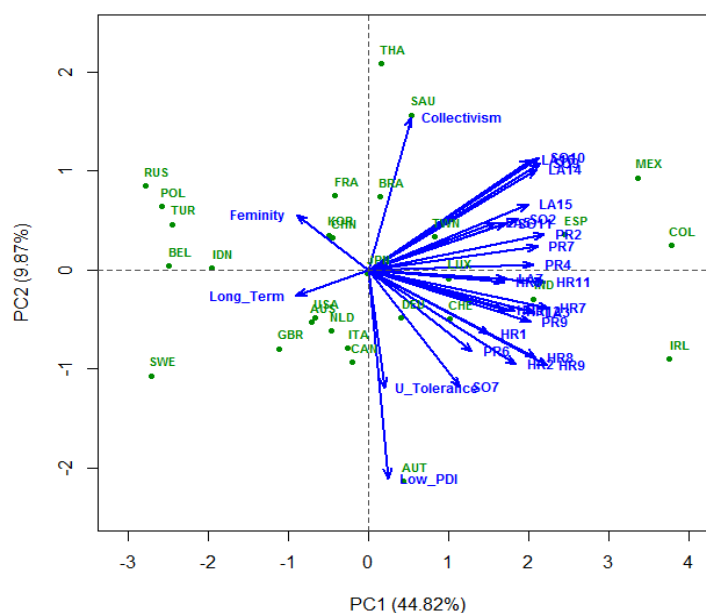


Figura 4-13 Ridge HJ Biplot. Datos sobre indicadores sociales y dimensiones culturales

En la siguiente gráfica Figura 4-15 se representa los datos mediante el Disjoint Biplot. Los ejes extraídos son disjuntos, con lo cual cada variable contribuye a la conformación de un solo eje y esto, por supuesto, también influye en la disminución de la variabilidad explicada por cada eje.

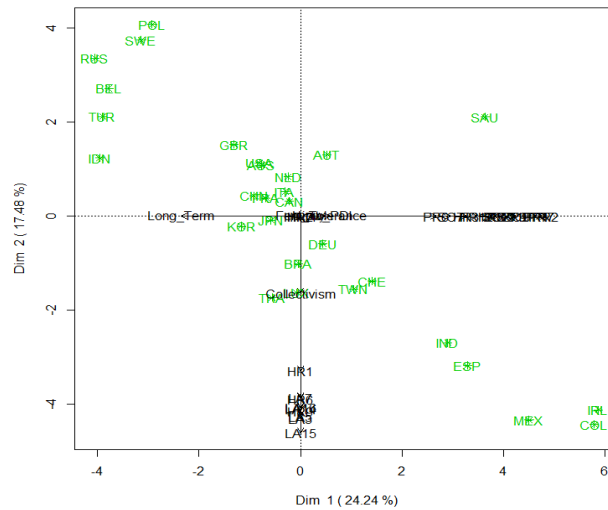


Figura 4-15 Disjoint Biplot. Datos sobre indicadores sociales y dimensiones culturales

El gráfico que representa el *Elastic Net* HJ Biplot se muestra en la Figura 4-16. Se observa que reteniendo los dos primeros ejes se alcanza un porcentaje de varianza explicada de 43.62%, la interpretación de los resultados es similar al Disjoint Biplot al conocer exactamente que variables aportan información a cada eje (ver Tabla 9).

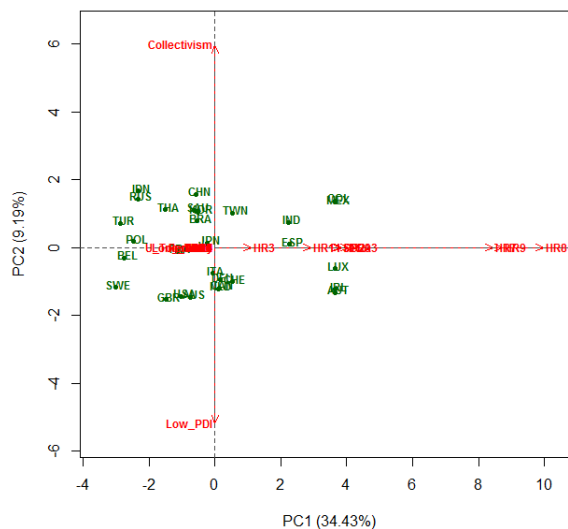


Figura 4-16 Elastic Net HJ Biplot. Datos sobre indicadores sociales y dimensiones culturales

Tabla 9. Cargas de las tres primeras componentes principales obtenidas mediante el HJ Biplot, Disjoint Biplot y los métodos de regularización Ridge, LASSO y Elastic Net (datos sobre indicadores sociales y cultura)

VAR	HJ-Biplot			Disjoint HJ Biplot			Ridge HJ Biplot			LASSO _{HARD} HJ Biplot			Elastic Net HJ Biplot		
	CP1	CP2	CP3	CP1	CP2	CP3	CP1	CP2	CP3	CP1	CP2	CP3	CP1	CP2	CP3
LA3	0.2130	-0.0940	-0.2320	0.000	1	0.000	0.1065	-0.0470	-0.1160	0.2130	0.000	-0.2319	0.2196	0.000	0.000
LA5	0.1619	0.1043	-0.2642	0.000	1	0.000	0.0810	0.0521	-0.1321	0.000	0.000	-0.2642	0.000	0.000	0.000
LA7	0.1776	-0.0196	-0.1540	0.000	1	0.000	0.0888	-0.0098	-0.0770	0.000	0.000	0.000	0.000	0.000	0.000
LA14	0.2164	0.2215	0.1988	0.000	1	0.000	0.1082	0.1108	0.0994	0.2164	0.2215	0.000	0.000	0.000	0.1174
LA15	0.2052	0.1451	-0.1367	0.000	1	0.000	0.1026	0.0726	-0.0683	0.2052	0.000	0.000	0.000	0.000	0.000
LA16	0.2081	0.2446	0.0438	0.000	1	0.000	0.1040	0.1223	0.0219	0.2080	0.2445	0.000	0.000	0.000	0.000
HR1	0.1553	-0.1422	0.0939	0.000	1	0.000	0.0776	-0.0711	0.0470	0.000	0.000	0.000	0.000	0.000	0.000
HR2	0.1893	-0.2094	-0.2573	0.000	0.000	1	0.0946	-0.1047	-0.1287	0.000	-0.2094	-0.2573	0.1995	0.000	0.000
HR3	0.1759	-0.0905	-0.0867	1	0.000	0.000	0.0880	-0.0453	-0.0433	0.000	0.000	0.000	0.0581	0.000	0.000
HR6	0.1743	-0.0267	-0.1363	0.000	1	0.000	0.0871	-0.0133	-0.0682	0.000	0.000	0.000	0.000	0.000	0.000
HR7	0.2315	-0.0843	0.0175	0.000	0.000	1	0.1157	-0.0421	0.0087	0.2315	0.000	0.000	0.4363	0.000	0.000
HR8	0.2150	-0.1944	-0.2026	1	0.000	0.000	0.1075	-0.0972	-0.1013	0.2150	0.000	-0.2026	0.5147	0.000	0.000
HR9	0.2297	-0.2130	-0.1725	0.000	1	0.000	0.1149	-0.1065	-0.0863	0.2297	-0.2130	0.000	0.4507	0.000	0.000
HR11	0.2267	-0.0281	-0.2343	0.000	0.000	1	0.1134	-0.0140	-0.1171	0.2267	0.000	-0.2342	0.1510	0.000	0.000
HR12	0.1885	-0.0909	-0.0451	1	0.000	0.000	0.0943	-0.0454	-0.0225	0.000	0.000	0.000	0.000	0.000	0.000

SO2	0.1914	0.1121	-0.0521	1	0.000	0.000	0.0957	0.0560	-0.0260	0.000	0.000	0.000	0.1961	0.000	0.000
SO7	0.1178	-0.2601	0.3398	1	0.000	0.000	0.0589	-0.1301	0.1699	0.000	-0.2601	0.3398	0.000	0.000	0.0413
SO9	0.2213	0.2356	0.0288	1	0.000	0.000	0.1107	0.1178	0.0144	0.2213	0.2355	0.000	0.000	0.000	0.000
SO10	0.2190	0.2475	0.0327	1	0.000	0.000	0.1095	0.1238	0.0163	0.2189	0.2475	0.000	0.000	0.000	0.000
SO11	0.1763	0.1017	0.1377	1	0.000	0.000	0.0882	0.0509	0.0688	0.000	0.000	0.000	0.000	0.000	0.000
PR2	0.2255	0.0795	0.2771	1	0.000	0.000	0.1128	0.0398	0.1385	0.2255	0.000	0.2771	0.000	0.000	0.5262
PR4	0.2116	0.0123	0.2991	1	0.000	0.000	0.1058	0.0062	0.1496	0.2116	0.000	0.2991	0.000	0.000	0.5516
PR6	0.1330	-0.1807	0.2089	1	0.000	0.000	0.0665	-0.0903	0.1044	0.000	0.000	0.2089	0.000	0.000	0.000
PR7	0.2172	0.0521	0.3092	1	0.000	0.000	0.1086	0.0261	0.1546	0.2172	0.000	0.3092	0.000	0.000	0.5832
PR9	0.2082	-0.1160	0.0643	1	0.000	0.000	0.1041	-0.0580	0.0321	0.2082	0.000	0.000	0.2092	0.000	0.000
LOW_PDI	0.0256	-0.4628	0.0002	1	0.000	0.000	0.0128	-0.2314	0.0001	0.000	-0.4628	0.000	0.000	-0.569	0.000
Collectivism	0.0542	0.3386	-0.2018	0.000	1	0.000	0.0271	0.1693	-0.1009	0.000	0.3386	-0.2018	0.000	0.6563	0.000
Feminity	-0.0934	0.1217	0.0401	0.000	0.000	1	-0.0467	0.0608	0.0201	0.000	0.000	0.000	0.000	0.000	0.000
U_Tolerance	0.0203	-0.2612	0.1854	1	0.000	0.000	0.0102	-0.1306	0.0927	0.000	-0.2612	0.000	0.000	0.000	0.000
Long_Term	-0.0938	-0.0574	-0.1949	1	0.000	0.000	-0.0469	-0.0287	-0.0975	0.000	0.000	0.000	0.000	0.000	0.000
sparsity	0	0	0	14	20	26	0	0	0	16	20	19	21	28	25
% var. expl.	44.82	9.87	8.39	24.24	17.48	8.45	44.82	9.87	8.39	38.61	11.28	9.30	34.43	9.19	14.12
% total		63.08			50.17			63.08			59.19			57.74	

La variabilidad explicada por cada eje en el algoritmo *Ridge* HJ Biplot es igual que la que explica el HJ Biplot ya que este algoritmo solamente contrae las cargas, pero como se ha comentado a lo largo del trabajo, no las hace nulas.

Por otro lado, se aprecia (Tabla 9) que la variabilidad explicada por cada eje en el Disjoint Biplot es menor que la explicada en el HJ Biplot, ya que el objetivo principal del DBiplot es encontrar la mejor clasificación de los individuos, más que maximizar la variabilidad.

Finalmente, los resultados de las metodologías de regularización (*Ridge*, *LASSO* y *Elastic Net*) utilizadas para obtener las distintas versiones del HJ Biplot se centran en quedarse con las variables más importantes para cada componente, haciendo nulas las cargas menos importantes. De esta manera, la proporción de varianza explicada se concentra en las primeras CP's con mayor o menor porcentaje (según el valor de penalización utilizado).

En síntesis, el hecho de que las grandes compañías afirman que reportan de acuerdo con la guía, pero no revelan todos los indicadores establecidos en ella, puede ser una práctica para camuflar problemas reales de sostenibilidad. En las grandes empresas, la ambigüedad entre los reportes de sostenibilidad y la declaración real y oportuna de los indicadores establecidos en la guía puede interpretarse como una forma de ocultar información y encubrir elementos importantes de divulgación, en un intento para hacer creer que la compañía está comprometida con las expectativas sociales, de cara a fortalecer su imagen y lograr ventajas competitivas.

CONCLUSIONES

1. Los resultados de la revisión bibliográfica realizada ponen de manifiesto que los métodos Biplot han logrado un creciente desarrollo desde su formulación en 1971, en el que Ruben Gabriel introduce el GH y el JK Biplot; el MANOVA Biplot de una vía en 1972, el MANOVA Biplot para tablas de 2 vías, en 2004; HJ Biplot en 1986, Biplot Canónico en 1992, Modelos AMMI GGE Biplot en 2001, Biplot para datos composicionales, en 2002, MANOVA Biplot para tablas de 3 vías en 2004, Biplot Logístico externo en 2008, el HJ Biplot inferencial, en 2015, el Biplot Logístico Nominal en 2016, Biplot Logístico Ordinal en 2016, y el HJ Biplot Composicional en 2016.
2. Frente a la dinámica de desarrollo actual de los métodos Biplot no hemos encontrado referencia alguna que implemente técnicas de regularización en el Biplot, por lo que el principal aporte de este trabajo consiste en la propuesta de nuevas versiones Biplot (SPARSE HJ Biplot), en base a los métodos de regularización *Ridge* (en base a la norma l_2), *LASSO* (en base a la norma l_1) y *Elastic Net* (combinando ambas normas).
3. La aplicación de una penalización en términos de la norma l_2 proporciona como resultado cargas pequeñas, pero no nulas; mientras que la aplicación de una penalización en términos de la norma l_1 tiende a dar como resultado que muchas cargas factoriales se reduzcan exactamente a cero y algunas otras cargas reciban poca o ninguna contracción. La penalización resultante de combinar las normas l_1 y l_2 tiende a dar un resultado intermedio, con menos cargas establecidas en cero que un ajuste en un valor de la norma l_1 , y una mayor contracción de los otros coeficientes.
4. Consideramos que el uso de las técnicas de regularización sparse en el Biplot, proporcionan soluciones eficientes a problemas planteados por la alta dimensionalidad de los datos.

5. Se ha realizado una comparación entre los resultados obtenidos mediante el HJ-Biplot, el Disjoint Biplot y los diferentes métodos del SPARSE HJ Biplot, aplicados a datos sobre grandes compañías a nivel mundial que reportan sus indicadores sociales mediante el Global Reporting Initiative. Así, se ha probado que al pasar del HJ Biplot a los resultados del Disjoint y de los Sparse HJ Biplot hemos reducido el número de variables, descartando las que menos aportan y seleccionando aquellas más importantes. Cabe mencionar que las variables culturales “feminidad” y “visión a largo plazo” no ejercen influencia en la divulgación de los indicadores sociales.
6. Hemos propuesto una versión combinada de los métodos Biplot y la descomposición CUR, la cual supone un gran avance en el análisis de grandes masas de datos, ya que nos permite seleccionar las variables más relevantes antes de aplicar los métodos BIPLLOT. Esto contribuye a reducir la dimensionalidad del problema pasando a hiperespacios de menor dimensión, lo cual supone una mayor calidad de representación en las representaciones euclídeas resultantes de los análisis BIPLLOT.
7. Para dar soporte a los nuevos métodos del SPARSE HJ Biplot que hemos planteado, se ha implementado una librería en el lenguaje de programación R, facilitando su aplicación a cualquier conjunto de datos mediante funciones específicas.
8. El software desarrollado en R, llamado SparseBiplots, permite aplicar todas las innovaciones teóricas de esta tesis doctoral.

Are cultural values sufficient to improve stakeholder engagement human and labour rights issues?

Mitzi Cubilla-Montilla^{1,2}  | Ana-Belén Nieto-Librero¹ | Ma Purificación Galindo-Villardón¹ | Ma Purificación Vicente Galindo¹ | Isabel-María García-Sánchez³ 

¹Department of Statistics, Campus Miguel de Unamuno, University of Salamanca, Salamanca, Spain

²University of Panama, Department of Statistics, Ringgold Standard Institution, Panama City, Panama

³Instituto Multidisciplinar de Empresa-IME, Campus Miguel de Unamuno, University of Salamanca, Salamanca, Spain

Correspondence

Mitzi Cubilla-Montilla, Departamento de Estadística, Universidad de Salamanca, Campus Miguel de Unamuno, Salamanca, Spain.
Email: mitzi@usal.es; mitzi.cubilla@up.ac.pa

Abstract

The complexity of the business world and current business models has motivated an increasing number of companies to disclose corporate information through sustainability reports. This reporting and stakeholders engagement may bring shared value to business and society in general although working towards sustainable development goals. This work adopts a new analytical approach by determining the global reporting initiative indicators related to labour practices and decent work, human rights, society, and product responsibility that are reported less frequently by companies. The final objective is to predict the influence that society's cultural values will play as a normative institutional pressure in their evolution. The results obtained for a sample comprising the 201 largest international companies that report in accordance with the recommendations of the G4 Guide in 2015 indicate that more than 50% of these large companies do not report specific mechanisms implemented to avoid violations of human rights and labour rights, or information on incidents related to production and commercial relations. Regulatory pressures associated with cultural values have limited effectiveness as drivers of greater corporate transparency in this area, as they are able to predict a favourable evolution for only 40% of companies that currently do not report.

KEYWORDS

biplot, corporate social responsibility, GRI, social indicators, stakeholder engagement, sustainable development

1 | INTRODUCTION

The corporate social responsibility (CSR) reports have as their main objective to inform stakeholders about the level of sustainability of business performance in order to enable stakeholders to make better decisions. The impact derived from these informative practices is extremely beneficial for companies, especially in terms of improving their image, reputation, and access to financing with better conditions (e.g., Cuadrado-Ballesteros, García-Sánchez, & Martínez Ferrero, 2016; Dhaliwal, Li, Tsang, & Yang, 2014; García-Sánchez & Noguera-Gámez, 2017b, 2017c; Martínez-Ferrero, Ruiz-Cano, & García-Sánchez, 2016).

However, the preparation of this information is expensive, and substantial funds must be allocated (Li, Gong, Zhang, & Koh, 2018). Such preparation may also entail costs for the owners derived from the sensitivity of the information issued (García-Sánchez & Martínez-Ferrero, 2017). Thus, researchers have focused on analysing the internal and external factors that lead organisations to disclose a greater volume of CSR information as well as its quality level (García-Sánchez, Martínez-Ferrero, & García-Benau, 2018; Martínez-Ferrero, García-Sánchez, & Ruiz-Barbadillo, 2018).

In relation to the internal factors, there is high agreement about the explanatory capacity of profitability (Allouche & Laroche, 2005), company size (McWilliams & Siegel, 2000), level of risk (Arora &

Dharwadkar, 2011), activity sector (Ndemanga & Koffi, 2009; Reverte, 2009; Ullmann, 1985), and certain characteristics of the board of directors (Frias-Aceituno, Rodríguez-Ariza, & García-Sánchez, 2013b; Fuente, García-Sánchez, & Lozano, 2017; García-Sánchez & Martínez-Ferrero, 2017; Prado-Lorenzo & García-Sánchez, 2010; Prado-Lorenzo, García-Sánchez, & Gallego-Álvarez, 2009a; Rodríguez-Ariza, Aceituno, & Rubio, 2014) in relation to corporate transparency in CSR.

In terms of external factors, institutional theory suggests that there are three main forces that drive organisational actions: normative, coercive, and mimetic. However, the number of studies that have evaluated the impact of the institutional environment is smaller, despite the strong effect that these factors have on business decisions (Khanna, Palepu, & Srinivasan, 2004; Riahi-Belkaoui & AlNajjar, 2006). Empirically, a considerable number of studies focus their efforts on analysing the informative practices of corporate information in certain countries (Deegan & Gordon, 1996; Díez, García, & Gago, 2012; R. Gray, Javad, Power, & Sinclair, 2001; Trotman & Bradley, 1981). Nevertheless, according to Scott (2008), it is the other dimensions of the institutional environment that cause companies to adopt different strategies of legitimisation before society, due to discrepancies between the coercive and normative pressures they support in their country of origin (Baughn, Bodie, & McIntosh, 2007; Buhr & Freedman, 2001; Van der Laan Smith, Adhikari, & Tondkar, 2005; Xiao, Gao, Heravi, & Cheung, 2005) and the existing mimetic forces at the sectoral level (Amor-Esteban, Galindo-Villardón, & García-Sánchez, 2018).

Thus, several researchers have evidenced the effect that several institutional pressures exert on CSR disclosure strategies (Frias-Aceituno, Rodríguez-Ariza, & García-Sánchez, 2013a; García-Sánchez, Cuadrado-Ballesteros, & Frías-Aceituno, 2016; García-Sánchez, Prado-Lorenzo, & Frías-Aceituno, 2013; García-Sánchez, Rodríguez-Ariza, & Frías-Aceituno, 2013). More concretely, the previous authors determined that legal and cultural systems are the main coercive and normative forces in explaining CSR disclosure and performance. Cultural values—the feminine and collective national cultural dimensions of society—are important boosters of the report of more comparable and useful CSR information.

However, all these paper focus on global measures of CSR or environmental information disclosure without analysing more deeply which information is revealed (or not revealed) by firms. In this sense, it is necessary to consider that external factors are doubtless when internal factors are not drivers of CSR disclosure regarding several issues that companies may be disinterested in disclosing (García-Sánchez & Noguera-Gámez, 2017a).

In line with this thinking, this work aims to determine the role played by the regulatory pressures of the country of origin of companies as drivers of information on the social impacts of business performance, which companies do not report or report less frequently than other information. This objective can be subdivided into two parts: first, the paper aims to demonstrate those indicators of the global reporting initiative (GRI) related to the social dimension of business performance that is more often omitted in the CSR reports issued by companies; second, it seeks to predict which regulatory forces are

determinant in the favourable evolution of indicators and to quantify their capacity or explanatory power.

For this, we use a sample of 201 companies belonging to 29 countries from different regions of the world. The sample only includes large listed companies because they are the most active in terms of sustainability and corporate transparency (García-Sánchez et al., 2016), besides being a model for other companies and reporting more frequently in accordance with the most demanding criteria of the GRI-G4 guide (García-Sánchez & Martínez-Ferrero, 2017).

Consideration of the GRI indicators is a consequence of the relevance that these guides have worldwide (Naeem & Welford, 2009) and the role they play in the issuance of useful and internationally comparable information (García-Sánchez & Martínez-Ferrero, 2018; Manetti, 2011; Martínez-Ferrero, Suárez-Fernández, & García-Sánchez, 2018). In relation to information on social issues, the reporting of information focuses on analysis of the following dimensions: labour practices and decent work, human rights, society, and product responsibility. Previous empirical evidence has been oriented mainly towards the study of information on environmental issues, giving less importance to the social dimension of CSR, despite the important impact it has in certain labour-intensive sectors and the use of other, not natural, resources. In addition, the social dimension is subject to greater discretion on the part of the companies, although it is less regulated, being in numerous occasions an entirely voluntary business decision.

Analysis of normative pressure is carried out based on the cultural dimensions proposed by Hofstede (1983): collectivism, femininity, tolerance of uncertainty, power distance, and long-term orientation. According to Campbell (2007), these forces can aid or reinforce the responsible behaviour of organisations in response to the demands of society. This opinion is shared by many other authors, including Doh and Guay (2006), Galaskiewicz and Burt (1991), Matten and Moon (2008), and Scott (2008).

The two-way multivariate methodology used comprises a combined method that integrates the principal coordinates analysis with logistic regression (LR) to build an external logistic biplot. This technique allows simultaneous representation of the companies, the GRI indicators, and the cultural forces, allowing us to predict the evolution that the indicators should undergo, taking into account the normative pressure that the companies support.

The results obtained show that there are differences between countries in relation to the dissemination of information, justified by cultural values related to power distance, collectivism, femininity, tolerance of uncertainty, and long-term vision. The large multinationals outsource their production to third countries in which labour legislation is non-existent or extremely lax in order to lower their production costs, and 50% of these companies omit relevant information in this regard. The demands coming from collectivist societies, tolerant of uncertainty, and with a low power distance, are a driving force for achieving greater corporate transparency. However, only 40% of companies that do not report show a greater tendency to disclose social information about CSR driven by these regulatory pressures.

In addition to this introduction, the document is structured in four sections. In the following, the theoretical framework of analysis is established, addressing a brief description of CSR reports and the relevance of the institutional characteristics considered in the analysis. Following this, the methodology includes a description of the data and the methods of analysis used. The penultimate section contains the results obtained. The document concludes with a discussion and a summary of our main findings.

1.1 | Cultural values and CSR reports: The theoretical framework

1.2 | CSR reports and their evolution

CSR reports are intended to inform stakeholders about the economic, environmental, and social performance of an organisation in a given period of time, establishing transparent and reliable communication with the stakeholders (Orozco, Acevedo, & Acevedo, 2014). For this purpose, the reports expand upon the content of traditional financial statements (Williams & Pei, 1999), offering information on aspects that go beyond the corporate economic result (Gray, Kouhy, & Lavers, 1995), providing data on labour practices, relationships with suppliers and customers, environmental management policies and systems, community activities, charitable contributions, and the effect of the company's products on consumer health and safety (Williams, 1999).

The significant growth in the preparation of CSR reports during the last two decades (Gray et al., 1995) has led academics to analyse and examine the quantity, content, and quality of these reports as well as the factors that influence these characteristics (Clarkson, Li, Richardson, & Vasvari, 2008; da Silva Monteiro & Aibar-Guzmán, 2010; Hackston & Milne, 1996; Haniffa & Cooke, 2005; Hassan & Ibrahim, 2012; Jairo, 2013; Mio, 2010).

Thus, in general, the scientific community has determined that the guidelines proposed by the GRI are the main reference in the preparation of CSR reports at an international level (Dilling, 2010; Hess, 2008; López, García, & Rodríguez, 2007; Perez-Batres, Doh, Miller, & Pisani, 2012; Prado-Lorenzo, García-Sánchez, & Gallego-Álvarez, 2009b; Rasche, 2009; Tsang, Welford, & Brown, 2009; White, 2006) and that, in general terms, profitability, size, and activity sector are key factors in the decision to begin disclosing information on CSR and in improving its content, usefulness, and comparability (e.g., Albertini, 2014; Gibson & O'Donovan, 2007; Higgins, Milne, & Van Gramberg, 2015; Islam & McPhail, 2011; Kolk & Pinkse, 2010; Lauwo, 2018; Romolini, Fissi, & Gori, 2014; Russo-Spena, Tregua, & De Chiara, 2018; Secchi, 2006).

However, when researchers gather the opinions of stakeholders, they observe that the credibility and usefulness of the CSR reports are insufficient for them to interact with the companies in order to address social and environmental issues and improve CSR practices in business (O'Dwyer, Unerman, & Hession, 2005). In addition, other authors, through content analysis, have evidenced important

deficiencies in CSR disclosure practices, highlighting the margins of improvement regarding compliance with international disclosure standards and effective engagement with stakeholders, for example, Skouloudis, Evangelinos, and Kourmousis (2010) for Greece; Mio (2010) for Italy; Moseñe, Burritt, Sanagustín, Moneva, and Tingey-Holyoak (2013) for Spain; Asif, Searcy, dos Santos, and Kensah (2013) for The Netherlands; Ahmad and Mohamad (2014) for Malaysia; Yadava and Sinha (2016) for India; and Rodrigue (2014) for Canada.

In addition, Laine (2010) shows that the companies that pioneered the dissemination of CSR reports broadcast polyphonic information; at present, companies use a fairly similar rhetoric, being able to engage in the same discourse for unsustainable behaviours. O'Donovan (2002), Pellegrino and Lodhia (2012), Hahn and Lülfes (2014), and Morrison, Wilmshurst, and Shimeld (2016), amongst others, observe that companies use CSR reports strategically, offering an analysis and argumentation of a point of view that allows them to obtain additional benefits in the promotion of dialogue with stakeholders. Despite the conciseness, comparability, and understandability of the information issued, it is possible to question its utility (Boiral, 2013) and the extent to which sustainability reports should be considered as a simulation used to camouflage real sustainability problems and project an idealised vision of the performance of mining and energy companies.

However, research has analysed the environmental information reported (or not reported), in general, in specific industries (Adler, Mansi, Pandey, & Stringer, 2017; Boiral, 2016; Kleinman, Kuei, & Lee, 2017; Leong, Hazelton, Taplin, Timms, & Laurence, 2014; Talbot & Boiral, 2015a, 2015b) or the level of CSR information standardisation (Belal & Owen, 2015; Lock & Seele, 2016; Michelon, Pilonato, & Ricceri, 2015).

In this paper, we improve upon the previous evidence in two ways. We analyse the social indicators (labour practices and decent work, human rights, society, product responsibility) that companies include or not in CSR reports, along with the role that normative institutional pressures may play in the improvement of social information disclosure. In this regard, some studies have found substantial differences in the quality of CSR reports between countries, which indicates that the institutional environment is a fundamental determinant of CSR disclosure practices (Baughn et al., 2007; Buhr & Freedman, 2001; Fekrat, Inclan, & Petroni, 1996; Freedman & Stagliano, 1992; Gamble, Hsu, Jackson, & Tollerson, 1996; Meek, Roberts, & Gray, 1995; Van der Laan Smith et al., 2005; Williams, 1999; Williams & Pei, 1999; Xiao et al., 2005).

1.3 | Normative forces as CSR disclosure prediction

Institutional theory establishes that companies are economic units that operate within an environment formed by the "rules of the game" that establish the different institutions comprising that environment and affecting its operation (Campbell, 2007; Campbell, Hollingsworth, & Lindberg, 1991), causing organisations that operate in similar institutional environments to adopt homogeneous forms of behaviour

(Claessens & Fan, 2002; La Porta, Lopez-de-Silanes, Shleifer, & Vishny, 1998). As a result, organisations become isomorphic in order to achieve greater stability and survival, facilitating institutional legitimacy (DiMaggio & Powell, 1983).

Three types of force or pressure determine this organisational isomorphism. Normative pressures allow organisations to acquire legitimacy within the profession and within the society in which they operate. Coercive pressures entail compliance with rules arising from external pressures exerted by the government or regulatory agencies. Mimetic forces refer to the process of imitation that some organisations adopt to resemble others, usually those that are most successful (Perez-Batres, Miller, & Pisani, 2011).

For Campbell (2006), companies will have a better predisposition towards socially responsible behaviour, and therefore provide more information, when operating in strong coercive and normative institutional environments, and especially, according to Campbell (2007), Doh and Guay (2006), Galaskiewicz and Burt (1991), Matten and Moon (2008), and Scott (2008), in environments with greater regulatory pressure. According to Campbell (2007), normative pressures can help to reinforce the responsible behaviour of organisations regarding the demands of society. According to Scott (2008), cultural systems should be considered normative forces, because they introduce a prescriptive, evaluative, and obligatory dimension into social life.

The different characteristics or cultural values of society play an important role in business decisions and actions, especially in the field of CSR (Esteban, Villardón, & Sánchez, 2017) and, consequently, in the information disclosed and its usefulness in the process of rendering accounts to stakeholders (Prado Lorenzo, García Sánchez, & Blázquez Zaballos, 2013; Yusoff, Othman, & Yatim, 2014). On the one hand, Hofstede (2003) points out that culture, in this sense, shapes the cognitive schemes of the individual or organisation, programming patterns of behaviour related to the cultural context. On the other hand, Tsakumis (2007) argues that national cultural dimensions explain the similarities and general differences in the collective mental programming of the human mind that distinguish one society from another.

According to Hofstede (1991), culture is always a collective phenomenon that is learned, not inherited and that influences human behaviour. It transcends in a similar way both the values of the individual and the values that govern the behaviour of companies (Vitell, Paolillo, & Thomas, 2003) through the effect that cultural values have on decision-making processes in the business world (Kim & Kim, 2010; Singhapakdi, Vitell, & Leelakulthani, 1994; Su, 2006). Consequently, culture influences the CSR practices that implement and the dissemination that they issue in this regard (Richerson & Boyd, 2005).

In this sense, various investigations show the impact that the cultural system has on different types of business report (Adams & Kuasirikun, 2000; Buhr & Freedman, 2001; García-Sánchez et al., 2016; Hope, 2003; Kim & Kim, 2010; Langlois & Schlegelmilch, 1990; Neu, Warsame, & Pedwell, 1998; Orij, 2010; Salter & Niswander, 1995; Tsakumis, 2007; Welford, 2004).

Different models have tried to explain the cultural differences between countries, but the model that has received greater recognition in the literature on corporate disclosure and CSR is the

five-dimension model proposed by Hofstede (2001). These dimensions are large or small power distance, individualism versus collectivism, masculinity versus femininity, tolerance versus aversion to uncertainty, and long- versus short-term orientation. These dimensions have been used by numerous investigators (e.g., Aceituno, da Conceição Marques, & Ariza, 2013; Christie, Kwon, Stoeberl, & Baumhart, 2003; Maignan, 2001; Orij, 2010; Ringov & Zollo, 2007; Van der Laan Smith et al., 2005; Vitell et al., 2003; G. Williams & Zinkin, 2008).

The power distance dimension describes a perception of the social hierarchy in terms of equality or inequality. A large power distance reflects that the levels of power are vertically stratified, establishing different hierarchies of power. In this sense, in countries with less power distance, companies will feel greater pressure to develop CSR practices and disclose them, whereas companies located in countries with greater power distance will be more focused on satisfying the interests of shareholders ahead of other interest groups (García-Sánchez et al., 2016).

The individualism/collectivism dimension reflects the tendency of society to reinforce individual or collective well-being. Collectivist societies are formed by individuals who think more of themselves as members of a group; their identity is therefore based on the social system to which they belong (de Mooij & Hofstede, 2010). Individuals with higher collectivism values have stronger links with society (Hofstede & Hofstede, 2005) and will therefore lead companies to show greater concern for sustainability, issuing more information about the impact of their actions on the social and environmental levels.

The dimension of masculinity/femininity refers to the role played by gender within society. Masculine cultures are usually assertive, strong, competitive, and oriented towards material success, whereas in feminine cultures, cooperation, modesty, consensus, and concern for quality of life prevail. Stakeholders in companies with a feminine orientation tend to be more open in the dissemination of information on CSR because they are more supportive societies (Gray, 1988) and demonstrate greater sensitivity to other perspectives of business performance related to the needs of society and the opportunities to satisfy them (García-Sánchez et al., 2016).

The dimension of aversion to versus tolerance of uncertainty determines the level at which people prefer structured situations over unstructured situations. Societies with a lower level of aversion to uncertainty are more open to change, have fewer rules, and have more flexible guidelines. In this sense, stakeholders in societies with a lower level of aversion to uncertainty have higher expectations of CSR practices than those in countries with a greater aversion to uncertainty. In the former institutional environments, CSR practices are motivated by legislation, which forces companies to behave in more sustainable ways (García-Sánchez et al., 2016). Countries with high values of aversion to uncertainty have a preference for secrecy and issue less information about CSR (Gray, 1988).

The long-/short-term orientation dimension describes the time horizon of a society. Cultures with short-term orientation value traditional methods spend a considerable amount of time developing relationships. A company with a long-term vision requires greater

CSR practices and corporate information, providing information about the company's behaviour and its future impact (García-Sánchez et al., 2016).

Prado-Lorenzo, García-Sánchez, and Blázquez-Zaballos (2013), Yusoff et al. (2014), and García-Sánchez et al. (2016) observe that companies whose country of origin is characterised by higher cultural values in the dimensions of collectivism and femininity are more sensitive to the disclosure of information on CSR, followed by the dimensions of greater long-term orientation, and, subsequently, greater tolerance of uncertainty and a lower index of power distance. However, Gallén and Peraita (2017) show that companies in more feminine countries disclose more information about CSR, this information is not of a higher quality.

In this sense, in relation to the theoretical framework and the previous empirical evidence, we defend that companies located in countries in which higher cultural values associated with the dimensions of collectivism, femininity, tolerance of uncertainty, long-term vision, and less power distance predominate will generally disclose greater volumes of information on CSR and, in particular, indicators associated with the social dimensions related to labour practices and decent work, human rights, society, and product responsibility.

However, the impact that these dimensions have on corporate transparency will depend on the type of information to be disclosed. Thus, the cultural features associated with collectivism, femininity, and long-term vision will be the first determinants of levels of corporate transparency. However, the dimensions of tolerance of uncertainty and less power distance will be the true drivers of those indicators that have lower frequency in their disclosure because, for various reasons, companies do not report them.

2 | METHODOLOGY

2.1 | Population and sample

To successfully address the objectives of this work, we considered the 500 largest companies worldwide according to Fortune Global 500¹ for the year 2015. This population is selected because large companies have a higher social impact, which implies a certain obligation in sustainability practices and disclosure of information about them. In addition, the media and the general public demand more information from large companies and their turnover is sometimes higher than the GDP of some countries.

The sample corresponds to 201 companies from 29 countries that prepare and disseminate a CSR report in accordance with the GRI model, specifically version G4. This guide is a document used worldwide to provide a standard report on the content of sustainability reports and to give them credibility and transparency. The data analysed in relation to the indicators of the G4 version were obtained from the information published in sustainability reports on companies'

¹The Fortune Global 500, also known as Global 500, is an annual ranking of the top 500 corporations worldwide as measured by revenue and the list is compiled and published annually by Fortune magazine.

websites through a content analysis. Specifically, the presence or absence of each of the social indicators related to the categories of labour practices and decent work, human rights, society, and product responsibility was evaluated. As already indicated, the social dimension of CSR lacks the academic coverage of environmental indicators, even when it concerns information of a more discretionary nature, because it is less regulated.

2.2 | Variables: Social indicators and cultural values

Table 1 reflects the GRI indicators related to the social dimension proposed in the G4 guide. They correspond to 48 indicators, 16 relating to describing business actions in relation to labour practices and decent work, 12 to human rights, 11 to commitment to society, and nine to product responsibility.

To determine normative pressures, we use the national cultural features proposed by Hofstede (2001), which take values between 0 and 100. Specifically, the relative cultural dimensions are large or small power distance, individualism versus collectivism, masculinity versus femininity, tolerance versus aversion to uncertainty, and long- versus short-term orientation. Except for this last dimension, higher values imply companies less oriented towards CSR and corporate transparency, which is why we will use their inverse. Table 2 summarises Hofstede's cultural dimensions:

2.3 | Analysis technique

The information used in our analysis was organised in a binary data matrix $X (i \times j)$, in which rows (i) correspond to the 201 largest companies in the world and columns (j) correspond to the 48 variables or binary indicators referring to social aspects. All indicators are binary variables that take the value one when the characteristic is present (reported indicator) and zero when it is absent (undisclosed).

The ordering and nature of the data require the use of two-way techniques that allow graphic representations that facilitate visual analysis of the results with strong statistical support that guarantees both its adequate interpretation and prediction of its evolution. Specifically, the external logistic biplot and HJ-biplot have been used to analyse the information through the MultiBiplot program, a programming environment integrated in MATLAB and developed by Vicente-Villardón (2010), available on the website: <http://biplot.usal.es>. By implementing both techniques, we approximate a data matrix in a way that allows its description or modelling through geometric maps constructed as projections of point cloud rows and columns on subspaces of optimum adjustment.

2.3.1 | External logistic biplot

For the analysis, we use a methodology proposed by Vicente-Villardón, Galindo-Villardón, and Blázquez-Zaballos (2006) and extended by Demey, Vicente-Villardón, Galindo-Villardón, and Zambrano (2008) that combines principal coordinates analysis and

TABLE 1 Social indicators (GRI-G4)

Social dimension	Indicator	Indicator code
Labor practices and decent work	Employment	LA1, LA2, and LA3
	Labor/management relations	LA4
	Occupational health and safety	LA5, LA6, LA7, and LA8
	Training and education	LA9, LA10, and LA11
	Diversity and equal opportunity	LA12
	Equal remuneration for women and men	LA13
	Supplier assessment for labor practices	LA14 and LA15
	Labor practices grievance mechanisms	LA16
Human rights	Investment	HR1 and HR2
	Non-discrimination	HR3
	Freedom of association and collective bargaining	HR4
	Child labor	HR5
	Forced or compulsory labor	HR6
	Security practices	HR7
	Indigenous rights	HR8
	Assessment	HR9
	Supplier human rights assessment	HR10 and HR11
	Human rights grievance mechanisms	HR12
Society	Local communities	SO1 and SO2
	Anti-corruption	SO3, SO4, and SO5
	Public policy	SO6
	Anti-competitive behaviour	SO7
	Compliance	SO8
	Supplier human rights assessment	SO9 and SO10
	Grievance mechanisms for impacts on society	SO11
	Product responsibility	Customer health and safety
Product and service labelling		PR3, PR4, and PR5
Marketing communications		PR6 and PT7
Customer privacy		PR8
Compliance		PR9

TABLE 2 Normative institutional pressures: Cultural values

0 ←	Hofstede's cultural dimensions	→ 100
Low power distance	PDI	High power distance
Collectivism	INV	Individualism
Femininity	MAS	Masculinity
Low uncertainty avoidance	UAI	High uncertainty avoidance
Short-term orientation	LTO	Long-term orientation

Note. INV: individualism; LTO: long-term orientation; MAS: masculinity; PDI: power distance; UAI: uncertainty avoidance index.

logistic regression in the same algorithm to build the technique known as an external logistic biplot (see Figure 1). In this paper, the graphic representation allows visualizing the relationships between the companies and indicators. Let $\pi_{ij} = E(x_{ij})$, the probability that the social indicator of CSR j is present at company i , and x_{ij} the observed probability (0 or 1), where the value zero indicates that the attribute is absent and one that it is present. π_{ij} can be written as

$$\pi_{ij} = \frac{e^{b_{j0} + \sum_{s=1}^I b_{js} a_{is}}}{1 + e^{b_{j0} + \sum_{s=1}^I b_{js} a_{is}}}$$

where

a_{is} ($i = 1, \dots, I$) and b_{js} ($j = 1, \dots, J$) are the model parameters.

This model is equivalent to the generalised linear model that uses the logit function as a link function to avoid scale problems (Demey et al., 2008).

For more detailed on the geometric properties of the external logistic biplot and the rules for you interpretation, consult other researchers (see Gallego-Álvarez, Ortas, Vicente-Villardón, & Álvarez Etxeberria, 2017; Gallego-Álvarez & Vicente-Villardón, 2012; Torres-Salinas, Robinson-García, Jiménez-Contreras, Herrera, & López-Cózar, 2013; Vicente-Villardón et al., 2006). This technique has been used in others studies (e.g., García-Pérez, Muñoz-Torres, & Fernández-Izquierdo, 2018; Tejedor-Flores, Galindo-Villardón, & Vicente-Galindo, 2016; Vicente Galindo, Vaz, & de Noronha, 2015; de Noronha Vaz, Galindo, de Noronha Vaz, & Nijkamp, 2015; P. V. Galindo, de Noronha Vaz, & Nijkamp, 2011; Rodero, Sanz-Valero, & Galindo-Villardón, 2018).

2.3.2 | Predictive analysis in the logistic biplot

The regression coefficients are the vectors that show the direction that best predicts the probability of the presence of each index. For each variable, the ordering diagram is divided into two regions that predict the presence ($\pi_{ij} > 0.5$) or absence ($\pi_{ij} < 0.5$) of the attribute. A line that is perpendicular to the vector that represents the variable or indicator separates the two regions. For this work, the projection of a company in the direction of any vector (indicator) is interpreted as predicting the probability of the presence of that indicator in the company.

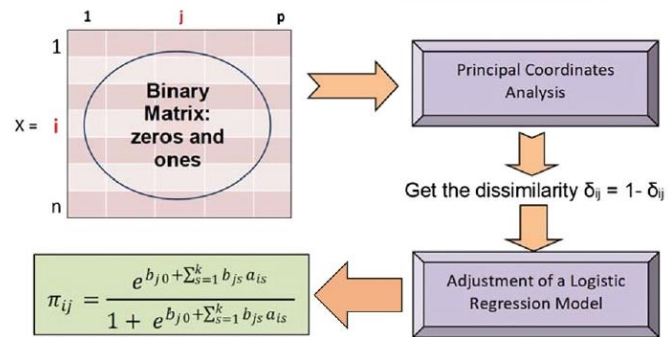


FIGURE 1 Algorithm for the application of the external logistic biplot [Colour figure can be viewed at wileyonlinelibrary.com]

2.3.3 | HJ-biplot technique

A biplot is a statistical approach for graphically depicting a data matrix $X (i \times j)$ obtained from analysing i individuals according to j numerical characteristics (Gabriel, 1971; Gabriel & Odoroff, 1986; Gower & Hand, 1995). Gabriel (1971) introduced biplot methods in the context of principal component analysis. The classical approach of biplot methods consists of two parts: first, a decomposition is performed in singular values of the data matrix; then the matrix is factored into row markers (individuals) and column markers (variables) (Eckart & Young, 1936; Golub & Reinsch, 1970).

In this study, we used the HJ-biplot statistical technique for multivariate data (Galindo, 1986), which has the capacity to represent both rows and columns simultaneously in a reduced dimension space. For this paper, the main goal of the HJ-biplot is to describe the relations between the rows (companies, which are displayed as points) and the columns (CSR social indicators and cultural system variables, displayed as vectors) according to the guidelines for the interpretation of the HJ-biplot:

- This, in turn, allows us to identify companies with similar behaviours—that is, we interpret the distance between points in relation to similarity: countries close to each other have similar profiles.
- In addition, the relationships between the social indicators of CSR, normative forces, and the relationships between them are described, meaning that acute angles between vectors are associated with a positive correlation between them, obtuse angles with a negative correlation, and right angles with uncorrelated indicators.
- To rank companies with respect to sustainability indicators and the normative forces, the orthogonal projections of the points (companies) on the vectors (indicators) are ordered in relation to each indicator and each normative.

This technique has been used in several contexts (for some examples, Demey et al., 2008; Esteban et al., 2017; Gallego-Álvarez, Galindo-Villardón, & Rodríguez-Rosa, 2015; Nieto-Librero, Sierra, Vicente-Galindo, Ruiz-Barzola, & Galindo-Villardón, 2017; Rodríguez-Rosa, Gallego-Álvarez, Vicente-Galindo, & Galindo-Villardón, 2017; Tejedor-Flores, Vicente-Galindo, & Galindo-Villardón, 2017).

3 | RESULTS

3.1 | Exploratory analysis

Prior to the analysis of behaviour and prediction, it was considered appropriate to perform a descriptive analysis of the frequency with which the companies reported each of the social indicators. Table 3 allows us to observe the relative frequencies for each indicator and category.

On average, 51.2% of the companies analysed disclose the social indicators established by the GRI G4 guide. This percentage is higher for the indicators related to the dimension of labour practices and decent work (59.8%) and society (53.7%). The percentage stands at 44.1% for human rights and at 47.3% for the indicators related to product responsibility.

The social indicators most widely reported by the 201 companies in their sustainability reports (GRI-G4) are, in the society subcategory, SO4—"communication and training on anti-corruption policies and procedures"—with 80.1% disclosure; in the labour practices and decent work subcategory, indicators LA10—"programmes for skills management and lifelong learning that support the continued employability of employees and assist them in managing career endings"—with 82.6% disclosure; and LA12—"composition of governance bodies and breakdown of employees per employee category according to gender, age group, minority group membership, and other indicators of diversity"—with 81.1% disclosure. In the human rights subcategory, the most reported indicator HR10—"percentage of new suppliers that were screened using human rights criteria"—barely reaches 54.7% disclosure; finally, in product responsibility, PR5—"results of surveys measuring customer satisfaction"—achieves 68.7% disclosure.

3.2 | Logistic biplot

The heterogeneity in the dissemination of indicators highlights the need to explore in more detail why companies report on some of the social indicators in some of the categories. Evaluation of the influence of the normative forces oriented towards companies' disclosure

TABLE 3 Descriptive statistics of social indicators

Dimension	Description	Code	Reported
Labor practices and decent work (LA)	Total number and rates of new employee hires and employee turnover by age group, gender, and region	LA1	71.1
	Benefits provided to full-time employees that are not provided to temporary or part-time employees, by significant locations of operation	LA2	64.7
	Return to work and retention rates after parental leave, by gender	LA3	48.8
	Minimum notice periods regarding operational changes, including whether these are specified in collective agreements	LA4	50.2
	Percentage of total workforce represented in formal joint management-worker health and safety committees that help monitor and advise on occupational health and safety programs	LA5	49.8
	Type of injury and rates of injury, occupational diseases, lost days, absenteeism, and total number of work-related fatalities, by region and by gender	LA6	72.1
	Workers with high incidence or high risk of diseases related to their occupation	LA7	49.8
	Health and safety topics covered in formal agreements with trade unions	LA8	52.7
	Average hours of training per year per employee by gender and by employee category	LA9	72.1
	Programs for skills management and lifelong learning that support the continued employability of employees and assist them in managing career endings	LA10	82.6
	Percentage of employees receiving regular performance and career development reviews, by gender and by employee category	LA11	66.7
	Composition of governance bodies and breakdown of employees per employee category according to gender, age group, minority group membership, and other indicators of diversity	LA12	81.1
	Ratio of basic salary and remuneration of women to men by employee category, by significant locations of operation	LA13	52.2
	Percentage of new suppliers that were screened using labor practices criteria	LA14	49.8
	Significant actual and potential negative impacts for labor practices in the supply chain and actions taken	LA15	48.8
	Number of grievances about labor practices filed, addressed and resolved through formal grievance mechanisms	LA16	44.8
Total average dimension: labor practices and decent work (LA)			59.8
Human rights (HRs)	Total number and percentage of significant investment agreements and contracts that include human rights clauses or that underwent human rights screening	HR1	41.8
	Total hours of employee training on human rights policies or procedures concerning aspects of human rights that are relevant to operations, including the percentage of employees trained	HR2	49.8
	Total number of incidents of discrimination and corrective actions taken	HR3	47.8
	Operations and suppliers identified in which the right to exercise freedom of association and collective bargaining may be violated or at significant risk, and measures taken to support these rights	HR4	50.7
	Operations and suppliers identified as having significant risk for incidents of child labor, and measures taken to contribute to the effective abolition of child labor	HR5	50.7
	Operations and suppliers identified as having significant risk for incidents of forced or compulsory labor, and measures taken to contribute to the elimination of all forms of forced or compulsory labor	HR6	49.8
	Percentage of security personnel trained in the organisation's human rights policies or procedures that are relevant to operations	HR7	32.8
	Total number of incidents of violations involving rights of indigenous peoples and actions taken	HR8	31.3
	Total number and percentage of operations that have been subject to human rights reviews or impact assessments	HR9	36.8
	Percentage of new suppliers that were screened using human rights criteria	HR10	54.7
	Significant actual and potential negative human rights impacts in the supply chain and actions taken	HR11	44.3
	Number of grievances about human rights impacts filed, addressed and resolved through formal grievance mechanisms	HR12	38.8
Total average dimension: human rights (HRs)			44.1
Society (SO)	Percentage of operations with implemented local community engagement, impact assessments, and development programs	SO1	67.7
	Operations with significant actual and potential negative impacts on local communities	SO2	45.8
	Total number and percentage of operations assessed for risks related to corruption and the significant risks identified	SO3	57.7
	Communication and training on anti-corruption policies and procedures	SO4	80.1
	Confirmed incidents of corruption and actions taken	SO5	55.7
	Total value of political contributions by country and recipient/beneficiary	SO6	50.2
	Total number of legal actions for anti-competitive behaviour, anti-trust, and monopoly practices and their outcomes	SO7	49.3
	Monetary value of significant fines and total number of nonmonetary sanctions for non-compliance with laws and regulations	SO8	52.7
	Percentage of new suppliers that were screened using criteria for impacts on society	SO9	49.3

(Continues)

TABLE 3 (Continued)

Dimension	Description	Code	Reported
	Significant actual and potential negative impacts on society in the supply chain and actions taken	S010	45.0
	Number of grievances about impacts on society filed, addressed, and resolved through formal grievance mechanisms	S011	37.3
	Total average dimension: society (SO)		53.7
Product responsibility (PR)	Percentage of significant product and service categories for which health and safety impacts are assessed for improvement	PR1	59.2
	Total number of incidents of non-compliance with regulations and voluntary codes concerning the health and safety impacts of products and services during their life cycle, by type of outcomes	PR2	40.0
	Type of product and service information required by the organisation's procedures for product and service information and labeling, and percentage of significant product and service categories subject to such information requirements	PR3	50.7
	Total number of incidents of non-compliance with regulations and voluntary codes concerning product and service information and labeling, by type of outcomes	PR4	37.3
	Results of surveys measuring customer satisfaction	PR5	68.7
	Sale of banned or disputed products	PR6	34.8
	Total number of incidents of non-compliance with regulations and voluntary codes concerning marketing communications, including advertising, promotion, and sponsorship, by type of outcomes	PR7	36.8
	Total number of substantiated complaints regarding breaches of customer privacy and losses of customer data	PR8	53.2
	Monetary value of significant fines for non-compliance with laws and regulations concerning the provision and use of products and services	PR9	45.0
	Total average dimension: product responsibility (PR)		47.3
	Total average—social indicators		51.2

of the sustainability indicators of a social type uses the four traditional cultural phases proposed by Hofstede and Hofstede (2005) and Hofstede (2001): power distance, collectivism, femininity, uncertainty avoidance, and long-term orientation.

It is expected that countries with higher values in these variables will be culturally more developed and therefore show greater interest in CSR, all of which will result in greater regulatory pressure on the companies, which will lead them to be more transparent. From this basis, the values of these five variables are averaged and we obtain a numerical value that we call "culture." From the averages, a typology for the culture is established, using the P_{25} , P_{50} , and P_{75} percentiles in such a way that the countries are divided into four types of cultural development, the first type being the least developed and the fourth the most developed culturally (see Figure 2).

The goodness of global adjustment, as a percentage of correct classification in the biplot, is 76.91%; consequently, 76.91% of the presences and absences for the indicator matrix are predicted correctly. In addition, the percentage of correctly classified variables was, in most cases, over 70%, so the prediction of the absence/presence of each indicator is approximate.

In each typology, we find different cultural developments associated with different normative pressures. Thus, the quadrants on the left group the companies whose country of origin has a greater cultural development towards CSR, whereas the quadrants on the right encompass the companies that support less normative pressure, associated with lower national cultural values.

The most popular indicators are located on the left side of the graph, so that companies located further to the left show a greater

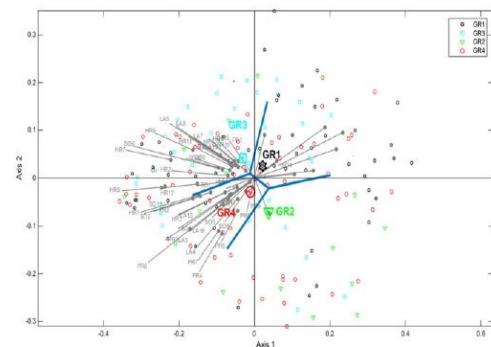


FIGURE 2 External logistic biplot: disclosure of social indicators in sustainability reports [Colour figure can be viewed at wileyonlinelibrary.com]

commitment to the dissemination of information of a social nature. The indicators that are reported less frequently are placed on the right side.

Combining the positioning of the cultural dimension and the social indicators in the plan, we can identify four types of business behaviour that we have subdivided into spaces GR1 to GR4, taking into account the centroid—that is, the midpoint of the companies of each type.

Most companies of the first type (GR1 less developed culturally) are on the right side of the figure, which indicates less commitment to the dissemination of social indicators as a result of a society supported by individualism, with high masculine values, that accepts

inequalities of power and is from a culture less open to change. The second typology (GR2) includes companies that, in general, do not assume a commitment to the disclosure of these indicators; however, they tend to report indicators related to product responsibility. The third typology (GR3) responds to companies located in the second and third quadrants, which show a greater commitment to the dissemination of social indicators in response to societies with less power distance, greater tolerance of uncertainty, and oriented towards community interests. The fourth typology (GR4) shows very dispersed companies and does not allow identification of a similar pattern of behaviour regarding the disclosure or not of social information.

The results of this representation derived from the external logistic biplot lead us to select those indicators that are reported by fewer than 50% of the companies analysed in order to predict their evolution, motivated by the regulatory pressure that each company supports. Methodologically, we observe that for this frequency there is a decoupling between the effect of normative pressure associated with cultural values and business practices related to the dissemination of social indicators.

3.3 | Prediction regions from logistic biplot

Next, we will analyse the 25 social indicators that present a percentage of disclosure below 50%. Of these 25 indicators, six are within the labour dimension, five in society, nine in human rights, and five in product responsibility (see Figure 3). The regions that predict

presence are coloured red and identify companies located in countries that have a higher normative pressure derived from higher cultural values and which, therefore, should be associated with greater transparency—that is, companies in these countries should disclose the indicator analysed. The regions that predict absence are coloured blue and identify companies in those countries with lower cultural values. In these regions, less regulatory pressure would mean a laxity for the companies located there, making them unlikely to report the indicator.

For all indicators, there is an observably high percentage of well-classified companies (red points in the region shaded in red, blue points in the region shaded in blue) and, depending on the indicator, different frequencies of companies whose disclosure practice for each indicator would not be associated with normative pressure (blue points in the red region and red points in the blue region).

The blue points in the red region indicate companies that report the indicator for reasons other than normative pressure because this is moderate in their country. The red companies located in the blue region are companies that will report on the indicator in the medium and long term due to the normative pressure they bear.

To predict the evolution of these indicators over time, in Table 4, we reflect the percentages of companies that will not increase their transparency in the social dimension of CSR, motivated by the cultural values of the country of origin of the companies as well as the effect that normative pressure could exert on the favourable evolution of these indicators.

On average, 43% of companies that do not report the social indicators analysed do so in the medium or long term due to the

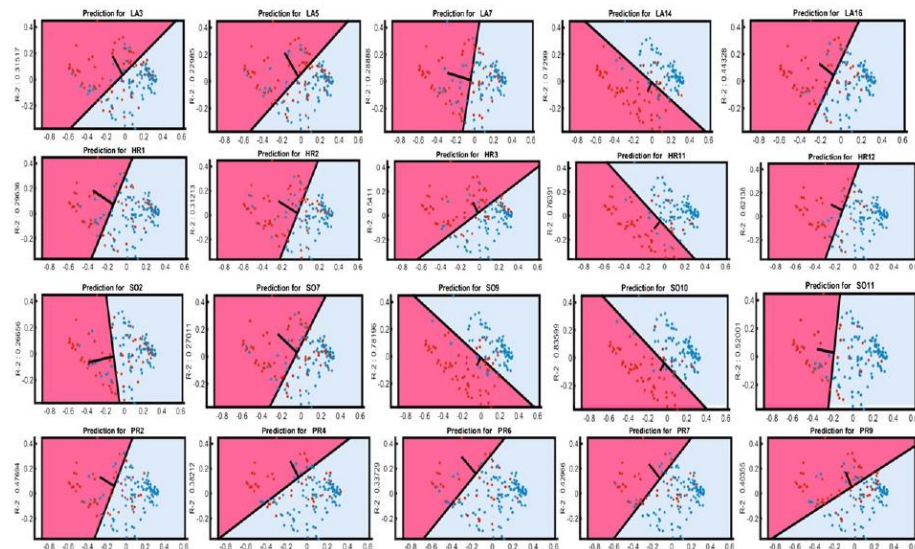


FIGURE 3 Regions of prediction of social indicators [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 4 Effect of normative pressure on the evolution of social indicators (frequencies in relation to companies that do not report)

Code indicator	R ²	Companies that do not report	
		Will report due to regulatory pressure	Will not report due to regulatory pressure
LA3	0.32	35.0	65.0
LA5	0.23	40.6	59.4
LA7	0.29	44.6	55.4
LA14	0.73	39.6	60.4
LA15	0.84	41.8	58.2
LA16	0.44	46.0	54.0
Average (LA)		41.2	58.8
HR1	0.30	40.2	59.8
HR2	0.31	42.6	57.4
HR3	0.54	52.4	47.6
HR6	0.40	42.6	57.4
HR7	0.37	48.1	51.9
HR8	0.54	50.0	50.0
HR9	0.47	38.6	61.4
HR11	0.76	47.3	52.7
HR12	0.62	20.3	79.7
Average (HR)		42.5	57.5
SO2	0.27	45.0	55.0
SO7	0.27	42.2	57.8
SO9	0.78	45.1	54.9
SO10	0.84	45.5	54.5
SO11	0.52	46.0	54.0
Average (SO)		44.7	55.3
PR2	0.48	45.0	55.0
PR4	0.38	44.4	55.6
PR6	0.34	42.0	58.0
PR7	0.43	43.3	56.7
PR9	0.40	41.8	58.2
Average (PR)		43.3	56.7
Average		43.0	57.0
Minimum		20.3	47.6
Maximum		52.4	79.7

Note. LA: labor practices and decent work; HR: human rights; PR: product responsibility; SO: society.

regulatory pressure they suffer in their country of origin. However, the remaining 57% will not do so or, if they do, it will be for reasons other than institutional force.

In the case of the dimension of labour practices and decent work, the respective percentages are 41.2% and 58.8%; for society, 44.7% and 55.3%; for human rights, 42.5% and 57.5%; and for the indicators related to product responsibility, 43.3% and 56.7%. On the other hand, indicators LA3, LA14, HR9, and HR12 associated with labour practices and decent work and human rights are expected to show the worst evolution in association with normative institutional forces, whereas indicators HR3 and HR8 will be in the opposite position.

3.4 | The HJ-biplot technique

Next, through an HJ-biplot, we will evaluate the effect or influence that each of the cultural values have on the disclosure of indicators. The application of this methodology allows us to approximate the set of indicators in a two-dimensional space, providing a useful visualisation of the structure of the countries in the sample in relation to social indicators and the variables that measure their cultural development. Therefore, we can find relationships between variables and characterise the companies according to the cultural development of their country of origin and their commitment to the disclosure of information simultaneously. The first two axes explain 55% of the variability of the data, allowing us to use the Factorial Planes 1–2 to represent the information in the following figure (Figure 4). The first eigenvalue (20.43) is significantly higher than the second eigenvalue (2.53), which means that the first (horizontal) dimension represents most of the information.

Cluster 1, located in the first quadrant, is formed by the companies with greater disclosure of sustainability in labour practices and decent work as well as in society indicators. These companies are located in countries dominated by a collectivist culture. In Cluster 2, located in the fourth quadrant, are companies whose reports are mainly characterised by the disclosure of indicators related to human rights, which is associated positively with cultures of low power distance and tolerance of uncertainty. Between the second and third quadrants, Cluster 3 brings together a group of companies that in this work are the least sustainable. It is observed that this is a group showing little activity in the reporting of social indicators; however, they are companies dominated by countries that have a feminine cultural system with a long-term vision.

The results obtained show that the favourable evolution of the indicators predicted in the previous subsection will be determined by the cultural values associated with higher levels of collectivism and tolerance of uncertainty and with less power distance. In contrast, the normative pressure associated with the cultural dimensions of femininity and long-term orientation will have no impact.

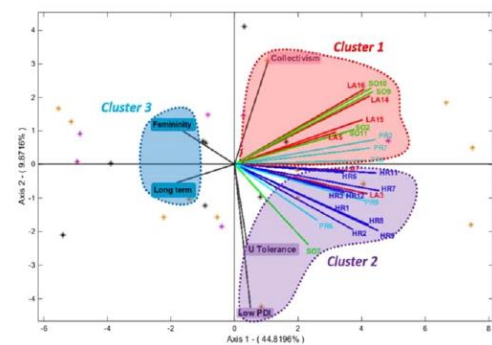


FIGURE 4 HJ-biplot with clusters [Colour figure can be viewed at wileyonlinelibrary.com]

4 | DISCUSSION

From the empirical point of view, the analyses carried out allow us to affirm that 48% of the largest companies worldwide, who should be leaders in sustainability, do not report in their sustainability reports 52% of the social indicators included in the G4 guide. In particular, revelations concerning the indicators relating to human rights are omitted, with the opacity percentage standing at 75%.

In addition, the indicators related to labour practices and decent work that companies report less frequently are related to the return to work after maternity or paternity leave, the percentage of workers whose tasks have an incidence or a high risk of illness, the presence of workers' representatives in established committees to help monitor and advise on occupational health and safety programmes, information regarding the requirements imposed on new suppliers with respect to criteria relating to labour practices, and incidents and claims that have occurred throughout the supply chain.

In the society dimension, there is a certain opacity around parameters similar to the previous ones, related to centres with a negative impact on local communities, the requirements imposed on new suppliers in relation to their impact on the local society, and incidents and associated claims that have occurred throughout the supply chain.

We find companies that show great commitment to respect human rights, labour rights, and local communities, but none report concrete mechanisms implemented to avoid violations of human rights and labour rights in all of their direct operations and through their chain of production and its commercial relations. In any case, accuracy of disclosure may add value for shareholders and stakeholders by demonstrating managerial commitment to reporting credible financial and sustainability information (García-Sánchez & Noguera-Gámez, 2017c).

The deficiencies observed for the social dimension of CSR are in line with the previous empirical evidence. Specifically, we noted that the revealed information is inferior to the recommendations established in the GRI, which limits its usefulness in decision-making processes. Our results extend the empirical evidence observed for environmental information in specific industries (i.e., Adler et al., 2017; Boiral, 2016; Kleinman et al., 2017; Leong et al., 2014; Talbot & Boiral, 2015a, 2015b) or the level of CSR information standardisation (i.e., Belal & Owen, 2015; Lock & Seele, 2016; Michelon et al., 2015).

In addition, we reinforce the findings obtained by Mio (2010), being able to affirm that even the companies that reveal greater volumes of information do not report all the necessary indicators to know the social impact of the company and to be able to evaluate the risks associated with this dimension of the CSR. Thus, according to the arguments of Boiral (2013), the fact that large companies claim that they report in accordance with GRI guidelines but do not detail all the indicators established in them legitimises the practice of camouflaging real sustainability problems, an obstacle for increasing credibility of the CSR (Hedberg & Von Malmborg, 2003). In this sense, Hoque, Rahman, Molla, Noman, and Bhuiyan (2018) finds that corporate managers practice CSR largely in a voluntary philanthropic fashion to build public image.

Our results give empirical robustness to the suggestion made by Gray et al. (2001) and those who argued that cultural dimensions should affect the outcome of CSR practices (del Mar Miras-Rodríguez, Carrasco-Gallego, & Escobar-Pérez, 2015; Scholtens & Kang, 2013). We confirm that the institutional environment is a fundamental determinant of CSR disclosure practices, not only with respect to the substantial differences observed in the quality of CSR reporting between countries (Baughn et al., 2007; Buhr & Freedman, 2001; Fekrat et al., 1996; Freedman & Stagliano, 1992; Gamble et al., 1996; Meek et al., 1995; Van der Laan Smith et al., 2005; Williams, 1999; Williams & Pei, 1999; Xiao et al., 2005), but also with regard to the amount and usefulness of reported information. Thus, strategies focused on CSR should consider not only the promotion of CSR policies but also the change in the countries' scenes (Fernandez-Feijoo, Romero, & Ruiz-Blanco, 2014).

Specifically, similar to Prado Lorenzo et al. (2013), Yusoff et al. (2014), García-Sánchez et al. (2016), and Gallén and Peraita (2017), we show that companies located in countries with communitarian cultural systems, which are feminine, more tolerant of uncertainty, and have less power distance and greater orientation to the long term, tend to disclose more relevant and comparable social information because interest groups have greater concern for the common social welfare. However, within these normative forces, it has been observed that the cultural dimensions associated with femininity and long-term orientation will not give rise to greater transparency in relation to the most controversial indicators, which companies have less interest in reporting.

5 | CONCLUSIONS

This work has the double objective of evidencing the informative practices of the large multinational companies regarding the social dimension of CSR and determining or predicting the evolution that these practices will undergo in accordance with the regulatory pressures of the companies' country of origin. It is noted that the cultural values of a society are the true drivers of information on the social impacts of business performance and of which practices companies do not report or report less frequently than other information.

For this, we use a sample of 201 companies belonging to 29 countries from different regions of the world. To demonstrate those GRI indicators related to the social dimension of business performance that are most often omitted in the CSR reports, we conducted an analysis of the content of the sustainability reports that these companies disclosed on their websites. Using biplot methodologies, we performed an analysis to predict which normative forces are decisive in their favourable evolution and to quantify their capacity or explanatory power.

The analysis affirms that in the information disclosed in sustainability reports from the largest companies in the world, based on the GRI guide, 52% of social indicators are not reported. Significantly, the analysis highlights the fact that 75% of the indicators on human rights are disclosed only by 30%–40% of the largest companies worldwide.

The results obtained correspond to the deficiencies described in previous studies on social aspects of CSR. Specifically, this work

shows that the information disclosed does not adequately reflect the recommendations established by the GRI, even though CSR assumes that companies take into account the impact of their strategies in the decision-making process. We can even say that the companies that reveal the greatest volumes of information are not reporting all the indicators associated with the social dimension of CSR. Thus, in large companies, the ambiguity between the sustainability reports according to the GRI and the actual and timely declaration of the indicators established in the guide can be interpreted as a way of hiding information and covering up important elements of disclosure. This is an attempt to pretend that the company is committed to social expectations in order to strengthen the image of the company and achieve legitimacy and competitive advantage.

On the other hand, the cultural dimensions associated with societies that are more tolerant of uncertainty, with less power distance and a community that will generate more effective regulatory pressure on companies, promote corporate transparency. Regarding evolution in the medium and long term, we predict that, on average, social indicators will be reported by 43% more companies due to the pressures they bear in relation to the cultural values of the society that characterises their country of origin, whereas 57% of companies will not report them due to the influence of these values or regulatory pressures. The indicators associated with labour practices and decent work and human rights are the worst predictors associated with normative forces.

The depth of our analysis allows us to determine that, within the normative forces analysed, the cultural dimensions associated with femininity and long-term orientation, although they are drivers of greater corporate transparency, have no impact on the revelation of more controversial indicators companies have less interest in reporting.

Finally, this work presents various contributions to literature. First is the consideration of information on social issues, especially the dimensions of labour practices and decent work, human rights, society, and product responsibility. Previous empirical evidence has been oriented mainly to the study of information on environmental issues, placing less importance on the social dimension of CSR, despite the important impact it has in certain labour-intensive sectors and to the use of other, not natural, resources, especially within the large multinationals that relocate their production to countries with looser labour legislation.

The second contribution is related to institutional theory. Analysis of normative pressure is conducted according to the cultural dimensions of collectivism, femininity, tolerance of uncertainty, power distance, and long-term vision, forces that can aid or reinforce the responsible behaviour of an organisation regarding the demands of society. However, unlike previous studies, we have observed that although initially these forces all constitute a normative pressure that promotes corporate transparency in terms of CSR, in terms of the dissemination of information on labour practices, human rights, society, and company responsibility, the cultural values of community societies with greater tolerance of uncertainty and long-term vision are the potential drivers.

The third contribution is related to the focus of the study, aimed at analysing corporate CSR information strategies, identifying their weaknesses, and determining the evolution they will undergo due to causes

beyond the control of internal decision-makers. In this sense, this work has important implications for academia, professionals, and regulators.


Considering that culture is always a collective phenomenon that is learned, not inherited, and that influences human behaviour, it seems advisable that politicians, legislators and other organisations promote campaigns, educational systems, and the like that favour the development of values oriented towards the common social welfare. Educating managers and interest groups to promote CSR practices aimed at greater sustainability and greater corporate transparency will characterise these societies.

Nonetheless, institutional pressures are the determining factors of business behaviour, though it has been stated that they are currently insufficient to increase information on the practices that companies undertake in relation to commitments to human and labour rights, the mechanisms implemented to prevent violations in all of a firm's direct operations, and repairing the negative effects caused by incidents and claims derived from their commercial relationships and systems. This lack of information does not allow a reading of corporate risk in the matter of human, labour, and considered rights that can affect the securities market of financial institutions, savings banks, and direct decisions to conduct more transparent business practices. Analysts and regulators must demand more detailed information in this regard to limit the impact of sustainability reports that do not contain this information on their predictions and the financing agreements to which these companies have access.

Finally, this paper presents several limitations such as those relating to analysing whether or not companies report the GRI indicators but not considering the level of homogeneity in the elaboration of the indicators they report. Likewise, the initial population selected corresponds to the 500 largest companies worldwide according to Fortune Global 500 for the year 2015. In this sense, it would be interesting to delve into a greater number of companies for a data paper. Moreover, in future research, it is necessary to extend the approach of this paper to current disclosure practices of environmental indicators. Likewise, it seems advisable to consider not only the effect of coercive pressures but also that of normative and mimetic forces. Methodologically, authors could use artificial intelligence techniques in order to predict business decisions about sustainability strategies.

ORCID

Mitzi Cubilla-Montilla  <https://orcid.org/0000-0002-8708-0351>

Isabel-María García-Sánchez  <https://orcid.org/0000-0003-4711-8631>

REFERENCES

- Aceituno, J. V. F., da Conceição Marques, M., & Ariza, L. R. (2013). Divulgación de información sostenible: ¿se adapta a las expectativas de la sociedad? *Revista de Contabilidad*, 16(2), 147–158. <https://doi.org/10.1016/j.rcsar.2013.07.004>
- Adams, C. A., & Kuasirikun, N. (2000). A comparative analysis of corporate reporting on ethical issues by UK and German chemical and pharmaceutical companies. *European Accounting Review*, 9(1), 53–79. <https://doi.org/10.1080/096381800407941>

- Adler, R., Mansi, M., Pandey, R., & Stringer, C. (2017). Article information: United Nations Decade on Biodiversity: A study of the reporting practices of the Australian mining industry. *Accounting, Auditing and Accountability Journal*, 30(8), 1711–1745. <https://doi.org/10.1108/09574090910954864>
- Ahmad, N. N. N., & Mohamad, N. A. (2014). Environmental disclosures by the Malaysian construction sector: Exploring extent and quality. *Corporate Social Responsibility and Environmental Management*, 21(4), 240–252. <https://doi.org/10.1002/csr.1322>
- Albertini, E. (2014). A descriptive analysis of environmental disclosure: A longitudinal study of French companies. *Journal of Business Ethics*, 121(2), 233–254. <https://doi.org/10.1007/s10551-013-1698-y>
- Allouche, J., & Laroche, P. (2005). A meta-analytical investigation of the relationship between corporate social and financial performance. *Revue de Gestion Des Ressources Humaines*, 57, 18.
- Amor-Esteban, V., Galindo-Villardón, M.-P., & García-Sánchez, I.-M. (2018). Industry mimetic isomorphism and sustainable development based on the X-STATIS and HJ-biplot methods. *Environmental Science and Pollution Research*. <https://doi.org/10.1007/s11356-018-2663-1>, 25, 26192–26208.
- Arora, P., & Dharwadkar, R. (2011). Corporate governance and corporate social responsibility (CSR): The moderating roles of attainment discrepancy and organization slack. *Corporate Governance: An International Review*, 19(2), 136–152. <https://doi.org/10.1111/j.1467-8683.2010.00843.x>
- Asif, M., Searcy, C., dos Santos, P., & Kensah, D. (2013). A review of Dutch corporate sustainable development reports. *Corporate Social Responsibility and Environmental Management*, 20(6), 321–339. <https://doi.org/10.1002/csr.1284>
- Baughn, C. C., Bodie, N. L., & McIntosh, J. C. (2007). Corporate social and environmental responsibility in Asian countries and other geographical regions. *Corporate Social Responsibility and Environmental Management*, 14(4), 189–205. <https://doi.org/10.1002/csr.160>
- Belal, A., & Owen, D. L. (2015). The rise and fall of stand-alone social reporting in a multinational subsidiary in Bangladesh: A case study. *Accounting, Auditing & Accountability Journal*, 28(7), 1160–1192. <https://doi.org/10.1108/09574090910954864>
- Boiral, O. (2013). Sustainability reports as simulacra? A counter-account of A and A+ GRI reports. *Accounting, Auditing & Accountability Journal*, 26(7), 1036–1071. <https://doi.org/10.1108/AAAJ-04-2012-00998>
- Boiral, O. (2016). Accounting for the unaccountable: Biodiversity reporting and impression management. *Journal of Business Ethics*, 135(4), 751–768. <https://doi.org/10.1007/s10551-014-2497-9>
- Buhr, N., & Freedman, M. (2001). Culture, institutional factors and differences in environmental disclosure between Canada and the United States. *Critical Perspectives on Accounting*, 12(3), 293–322. <https://doi.org/10.1006/cpac.2000.0435>
- Campbell, J. L. (2006). Institutional analysis and the paradox of corporate social responsibility. *American Behavioral Scientist*, 49(7), 925–938. <https://doi.org/10.1177/0002764205285172>
- Campbell, J. L. (2007). Why would corporations behave in socially responsible ways? An institutional theory of corporate social responsibility. *Academy of Management Review*, 32(3), 946–967. <https://doi.org/10.5465/amr.2007.25275684>
- Campbell, J. L., Hollingsworth, J. R., & Lindberg, L. N. (1991). *Governance of the American economy* (Vol. 5). Cambridge University Press. <https://doi.org/10.1017/CBO9780511664083>
- Christie, P. M. J., Kwon, I.-W. G., Stoeberl, P. A., & Baumhart, R. (2003). A cross-cultural comparison of ethical attitudes of business managers: India Korea and the United States. *Journal of Business Ethics*, 46(3), 263–287. <https://doi.org/10.1023/A:1025501426590>
- Claessens, S., & Fan, J. P. (2002). Corporate governance in Asia: A survey. *International Review of Finance*, 3(2), 71–103. <https://doi.org/10.1111/1468-2443.00034>
- Clarkson, P. M., Li, Y., Richardson, G. D., & Vasvari, F. P. (2008). Revisiting the relation between environmental performance and environmental disclosure: An empirical analysis. *Accounting, Organizations and Society*, 33(4–5), 303–327. <https://doi.org/10.1016/j.aos.2007.05.003>
- Cuadrado-Ballesteros, B., García-Sánchez, I.-M., & Martínez Ferrero, J. (2016). How are corporate disclosures related to the cost of capital? The fundamental role of information asymmetry. *Management Decision*, 54(7), 1669–1701. <https://doi.org/10.1108/MD-10-2015-0454>
- da Silva Monteiro, S. M., & Aibar-Guzmán, B. (2010). Determinants of environmental disclosure in the annual reports of large companies operating in Portugal. *Corporate Social Responsibility and Environmental Management*, 17(4), 185–204. <https://doi.org/10.1002/csr.197>
- de Mooij, M., & Hofstede, G. (2010). The Hofstede model. Applications to global branding and advertising strategy and research. *International Journal of Advertising*, 29(1), 85–110. <https://doi.org/10.2501/S026504870920104X>
- de Noronha Vaz, T., Galindo, P. V., de Noronha Vaz, E., & Nijkamp, P. (2015). Innovative firms behind the regions: Analysis of regional innovation performance in Portugal by external logistic biplots. *European Urban and Regional Studies*, 22(3), 329–344. <https://doi.org/10.1177/0969776412474675>
- Deegan, C., & Gordon, B. (1996). A study of the environmental disclosure practices of Australian corporations. *Accounting and Business Research*, 26(3), 187–199. <https://doi.org/10.1080/00014788.1996.9729510>
- del Mar Miras-Rodríguez, M., Carrasco-Gallego, A., & Escobar-Pérez, B. (2015). Are socially responsible behaviors paid off equally? A cross-cultural analysis. *Corporate Social Responsibility and Environmental Management*, 22(4), 237–256. <https://doi.org/10.1002/csr.1344>
- Demey, J. R., Vicente-Villardón, J. L., Galindo-Villardón, M. P., & Zambrano, A. Y. (2008). Identifying molecular markers associated with classification of genotypes by external logistic biplots. *Bioinformatics*, 24(24), 2832–2838. <https://doi.org/10.1093/bioinformatics/btn552>
- Dhaliwal, D., Li, O. Z., Tsang, A., & Yang, Y. G. (2014). Corporate social responsibility disclosure and the cost of equity capital: The roles of stakeholder orientation and financial transparency. *Journal of Accounting and Public Policy*, 33(4), 328–355. <https://doi.org/10.1016/j.jaccpubpol.2014.04.006>
- Diez, J. L. G., García, L. C., & Gago, R. F. (2012). Determinantes empresariales de la RSC en España. *Revista de Responsabilidad Social de La Empresa*, 12, 59–76.
- Dilling, P. F. (2010). Sustainability reporting in a global context: What are the characteristics of corporations that provide high quality sustainability reports—An empirical analysis. *International Business & Economics Research Journal*, 9(1), 19–30. <https://doi.org/10.19030/iber.v9i1.505>
- DiMaggio, P., & Powell, W. W. (1983). The iron cage revisited: Collective rationality and institutional isomorphism in organizational fields. *American Sociological Review*, 48(2), 147–160. <https://doi.org/10.2307/2095101>
- Doh, J. P., & Guay, T. R. (2006). Corporate social responsibility, public policy, and NGO activism in Europe and the United States: An institutional-stakeholder perspective. *Journal of Management Studies*, 43(1), 47–73. <https://doi.org/10.1111/j.1467-6486.2006.00582.x>
- Eckart, C., & Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3), 211–218. <https://doi.org/10.1007/BF02288367>
- Esteban, V. A., Villardón, M. P. G., & Sánchez, I. M. G. (2017). Cultural values on CSR patterns and evolution: A study from the biplot

- representation. *Ecological Indicators*, 81, 18–29. <https://doi.org/10.1016/j.ecolind.2017.05.051>
- Fekrat, M. A., Inclan, C., & Petroni, D. (1996). Corporate environmental disclosures: Competitive disclosure hypothesis using 1991 annual report data. *The International Journal of Accounting*, 31(2), 175–195. [https://doi.org/10.1016/s0020-7063\(96\)90003-5](https://doi.org/10.1016/s0020-7063(96)90003-5)
- Fernandez-Feijoo, B., Romero, S., & Ruiz-Blanco, S. (2014). Women on boards: Do they affect sustainability reporting? *Corporate Social Responsibility and Environmental Management*, 21(6), 351–364. <https://doi.org/10.1002/csr.1329>
- Freedman, M., & Stagliano, A. J. (1992). European unification, accounting harmonization, and social disclosures. *The International Journal of Accounting*, 27(2), 112–122.
- Frías-Aceituno, J. V., Rodríguez-Ariza, L., & García-Sánchez, I. M. (2013a). Is integrated reporting determined by a country's legal system? An exploratory study. *Journal of Cleaner Production*, 44, 45–55. <https://doi.org/10.1016/j.jclepro.2012.12.006>
- Frías-Aceituno, J. V., Rodríguez-Ariza, L., & García-Sánchez, I. M. (2013b). The role of the board in the dissemination of integrated corporate social reporting. *Corporate Social Responsibility and Environmental Management*, 20(4), 219–233. <https://doi.org/10.1002/csr.1294>
- Fuente, J. A., García-Sánchez, I. M., & Lozano, M. B. (2017). The role of the board of directors in the adoption of GRI guidelines for the disclosure of CSR information. *Journal of Cleaner Production*, 141, 737–750. <https://doi.org/10.1016/j.jclepro.2016.09.155>
- Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58(3), 453–467. <https://doi.org/10.1093/biomet/58.3.453>
- Gabriel, K. R., & Odoroff, C. L. (1986). Some diagnoses of models by 3-D biplots. *Multidimensional Data Analysis*, 91–111.
- Galaskiewicz, J., & Burt, R. S. (1991). Interorganization contagion in corporate philanthropy. *Administrative Science Quarterly*, 36(1), 88–105. <https://doi.org/10.2307/2393431>
- Galindo, M. (1986). Una alternativa de representación simultánea: HJ-Biplot. *Qüestió. 1986*, Vol. 10, Núm. 1.
- Galindo, P. V., de Noronha Vaz, T., & Nijkamp, P. (2011). Institutional capacity to dynamically innovate: An application to the Portuguese case. *Technological Forecasting and Social Change*, 78(1), 3–12. <https://doi.org/10.1016/j.techfore.2010.08.004>
- Gallego-Álvarez, I., Galindo-Villardón, M. P., & Rodríguez-Rosa, M. (2015). Analysis of the sustainable society index worldwide: A study from the biplot perspective. *Social Indicators Research*, 120(1), 29–65. <https://doi.org/10.1007/s11205-014-0579-9>
- Gallego-Álvarez, I., Ortas, E., Vicente-Villardón, J. L., & Álvarez Etxeberria, I. (2017). Institutional constraints, stakeholder pressure and corporate environmental reporting policies. *Business Strategy and the Environment*, 26(6), 807–825. <https://doi.org/10.1002/bse.1952>
- Gallego-Álvarez, I., & Vicente-Villardón, J. L. (2012). Analysis of environmental indicators in international companies by applying the logistic biplot. *Ecological Indicators*, 23, 250–261. <https://doi.org/10.1016/j.ecolind.2012.03.024>
- Gallén, M. L., & Peraita, C. (2017). The relationship between femininity and sustainability reporting. *Corporate Social Responsibility and Environmental Management*, 24(6), 496–508. <https://doi.org/10.1002/csr.1423>
- Gamble, G. O., Hsu, K., Jackson, C., & Tollerson, C. D. (1996). Environmental disclosures in annual reports: An international perspective. *The International Journal of Accounting*, 31(3), 293–331. [https://doi.org/10.1016/s0020-7063\(96\)90022-9](https://doi.org/10.1016/s0020-7063(96)90022-9)
- García-Pérez, I., Muñoz-Torres, M. J., & Fernández-Izquierdo, M. Á. (2018). Microfinance institutions fostering sustainable development. *Sustainable Development*. <https://doi.org/10.1002/sd.1731>, 26, 606–619.
- García-Sánchez, I. M., Cuadrado-Ballesteros, B., & Frías-Aceituno, J.-V. (2016). Impact of the institutional macro context on the voluntary disclosure of CSR information. *Long Range Planning*, 49(1). <https://doi.org/10.1016/j.lrp.2015.02.004>, 15–35.
- García-Sánchez, I. M., & Martínez-Ferrero, J. (2017). Independent directors and CSR disclosures: The moderating effects of proprietary costs. *Corporate Social Responsibility and Environmental Management*, 24(1), 28–43. <https://doi.org/10.1002/csr.1389>
- García-Sánchez, I.-M., & Martínez-Ferrero, J. (2018). How do independent directors behave with respect to sustainability disclosure? *Corporate Social Responsibility and Environmental Management*. <https://doi.org/10.1002/csr.1481>, 25, 609–627.
- García-Sánchez, I.-M., Martínez-Ferrero, J., & García-Benau, M.-A. (2018). Integrated reporting: The mediating role of the board of directors and investor protection on managerial discretion in munificent environments. *Corporate Social Responsibility and Environmental Management*. <https://doi.org/10.1002/csr.1655>, 26, 29–45.
- García-Sánchez, I. M., & Noguera-Gámez, L. (2017a). Institutional investor protection pressures versus firm incentives in the disclosure of integrated reporting. *Australian Accounting Review*, 28(2), 199–219. <https://doi.org/10.1111/auar.12172>
- García-Sánchez, I. M., & Noguera-Gámez, L. (2017b). Integrated information and the cost of capital. *International Business Review*, 26(5), 959–975. <https://doi.org/10.1016/j.ibusrev.2017.03.004>
- García-Sánchez, I. M., & Noguera-Gámez, L. (2017c). Integrated reporting and stakeholder engagement: The effect on information asymmetry. *Corporate Social Responsibility and Environmental Management*, 24(5), 395–413. <https://doi.org/10.1002/csr.1415>
- García-Sánchez, I. M., Prado-Lorenzo, J. M., & Frías-Aceituno, J. V. (2013). Información social corporativa y sistema legal. *Revista Europea de Dirección Y Economía de La Empresa*, 22(4), 186–202. <https://doi.org/10.1016/j.redee.2012.11.003>
- García-Sánchez, I. M., Rodríguez-Ariza, L., & Frías-Aceituno, J.-V. (2013). The cultural system and integrated reporting. *International Business Review*, 22(5), 828–838. <https://doi.org/10.1016/j.ibusrev.2013.01.007>
- Gibson, K., & O'Donovan, G. (2007). Corporate governance and environmental reporting: An Australian study. *Corporate Governance: An International Review*, 15(5), 944–956. <https://doi.org/10.1111/j.1467-8683.2007.00615.x>
- Golub, G. H., & Reinsch, C. (1970). Singular value decomposition and least squares solutions. *Numerische Mathematik*, 14(5), 403–420. <https://doi.org/10.1007/BF02163027>
- Gower, J. C., & Hand, D. J. (1995). *Biplots*. (Vol. 54). CRC Press.
- Gray, R., Javad, M., Power, D. M., & Sinclair, C. D. (2001). Social and environmental disclosure and corporate characteristics: A research note and extension. *Journal of Business Finance & Accounting*, 28(3–4), 327–356. <https://doi.org/10.1111/1468-5957.00376>
- Gray, R., Kouhy, R., & Lavers, S. (1995). Methodological themes: Constructing a research database of social and environmental reporting by UK companies. *Accounting, Auditing & Accountability Journal*, 8(2), 78–101. <https://doi.org/10.1108/09513579510086812>
- Gray, S. (1988). Towards a theory of cultural influence on the development of accounting systems internationally. *Abacus*, 24(1), 1–15. <https://doi.org/10.1111/j.1467-6281.1988.tb00200.x>
- Hackston, D., & Milne, M. J. (1996). Some determinants of social and environmental disclosures in New Zealand companies. *Accounting, Auditing*

- & *Accountability Journal*, 9(1), 77–108. <https://doi.org/10.1108/09513579610109987>
- Hahn, R., & Lüf, R. (2014). Legitimizing negative aspects in GRI-oriented sustainability reporting: A qualitative analysis of corporate disclosure strategies. *Journal of Business Ethics*, 123(3), 401–420. <https://doi.org/10.1007/s10551-013-1801-4>
- Haniffa, R. M., & Cooke, T. E. (2005). The impact of culture and governance on corporate social reporting. *Journal of Accounting and Public Policy*, 24(5), 391–430. <https://doi.org/10.1016/j.jaccpubpol.2005.06.001>
- Hassan, A., & Ibrahim, E. (2012). Corporate environmental information disclosure: Factors influencing companies' success in attaining environmental awards. *Corporate Social Responsibility and Environmental Management*, 19(1), 32–46. <https://doi.org/10.1002/csr.278>
- Hedberg, C.-J., & Von Malmborg, F. (2003). The global reporting initiative and corporate sustainability reporting in Swedish companies. *Corporate Social Responsibility and Environmental Management*, 10(3), 153–164. <https://doi.org/10.1002/csr.38>
- Hess, D. (2008). The three pillars of corporate social reporting as new governance regulation: Disclosure, dialogue, and development. *Business Ethics Quarterly*, 18(4), 447–482. <https://doi.org/10.5840/beq200818434>
- Higgins, C., Milne, M. J., & Van Gramberg, B. (2015). The uptake of sustainability reporting in Australia. *Journal of Business Ethics*, 129(2), 445–468. <https://doi.org/10.1007/s10551-014-2171-2>
- Hofstede, G. (1983). The cultural relativity of organizational practices and theories. *Journal of International Business Studies*, 14(2), 75–89. <https://doi.org/10.1057/palgrave.jibs.8490867>
- Hofstede, G. (1991). *Cultures and organizations: Software of the mind*. London: McGraw-Hill Book Company.
- Hofstede, G. (2001). *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations*. Sage publications.
- Hofstede, G. (2003). What is culture? A reply to Baskerville. *Accounting, Organizations and Society*, 28(7–8), 811–813. [https://doi.org/10.1016/s0361-3682\(03\)00018-7](https://doi.org/10.1016/s0361-3682(03)00018-7)
- Hofstede, G., & Hofstede, G. J. (2005). *Cultures and organizations: Software of the mind intercultural cooperation and its importance for survival*. London: McGraw-Hill.
- Hope, O.-K. (2003). Disclosure practices, enforcement of accounting standards, and analysts' forecast accuracy: An international study. *Journal of Accounting Research*, 41(2), 235–272. <https://doi.org/10.1111/1475-679x.00102>
- Hoque, N., Rahman, A. R. A., Molla, R. I., Noman, A. H. M., & Bhuiyan, M. Z. H. (2018). Is corporate social responsibility pursuing pristine business goals for sustainable development? *Corporate Social Responsibility and Environmental Management*. <https://doi.org/10.1002/csr.1527>, 25, 1130–1142.
- Islam, M. A., & McPhail, K. (2011). Regulating for corporate human rights abuses: The emergence of corporate reporting on the ILO's human rights standards within the global garment manufacturing and retail industry. *Critical Perspectives on Accounting*, 22(8), 790–810. <https://doi.org/10.1016/j.cpa.2011.07.003>
- Jairo, G. C. (2013). Modelos financieros con Excel: Herramientas para mejorar la toma de decisiones empresariales. Ecoe ediciones.
- Khanna, T., Palepu, K. G., & Srivivasan, S. (2004). Disclosure practices of foreign companies interacting with US markets. *Journal of Accounting Research*, 42(2), 475–508. <https://doi.org/10.1111/j.1475-679x.2004.00146.x>
- Kim, Y., & Kim, S.-Y. (2010). The influence of cultural values on perceptions of corporate social responsibility: Application of Hofstede's dimensions to Korean public relations practitioners. *Journal of Business Ethics*, 91(4), 485–500. <https://doi.org/10.1007/s10551-009-0095-z>
- Kleinman, G., Kuei, C., & Lee, P. (2017). Using formal concept analysis to examine water disclosure in corporate social responsibility reports. *Corporate Social Responsibility and Environmental Management*, 24(4), 341–356. <https://doi.org/10.1002/csr.1427>
- Kolk, A., & Pinkse, J. (2010). The integration of corporate governance in corporate social responsibility disclosures. *Corporate Social Responsibility and Environmental Management*, 17(1), 15–26. <https://doi.org/10.1002/csr>
- La Porta, R., Lopez-de-Silanes, F., Shleifer, A., & Vishny, R. W. (1998). Law and finance. *Journal of Political Economy*, 106(6), 1113–1155. <https://doi.org/10.1086/250042>
- Laine, M. (2010). Towards sustaining the status quo: Business talk of sustainability in Finnish corporate disclosures 1987–2005. *European Accounting Review*, 19(2), 247–274. <https://doi.org/10.1080/09638180903136258>
- Langlois, C. C., & Schlegelmilch, B. B. (1990). Do corporate codes of ethics reflect national character? Evidence from Europe and the United States. *Journal of International Business Studies*, 21(4), 519–539. <https://doi.org/10.1057/palgrave.jibs.8490340>
- Lauwo, S. (2018). Challenging masculinity in CSR disclosures: Silencing of women's voices in Tanzania's mining industry. *Journal of Business Ethics*, 149(3), 689–706. <https://doi.org/10.1007/s10551-016-3047-4>
- Leong, S., Hazelton, J., Taplin, R., Timms, W., & Laurence, D. (2014). Mine site-level water reporting in the Macquarie and Lachlan catchments: A study of voluntary and mandatory disclosures and their value for community decision-making. *Journal of Cleaner Production*, 84(1), 94–106. <https://doi.org/10.1016/j.jclepro.2014.01.021>
- Li, Y., Gong, M., Zhang, X. Y., & Koh, L. (2018). The impact of environmental, social, and governance disclosure on firm value: The role of CEO power. *British Accounting Review*, 50(1), 60–75. <https://doi.org/10.1016/j.bar.2017.09.007>
- Lock, I., & Seele, P. (2016). The credibility of CSR (corporate social responsibility) reports in Europe. Evidence from a quantitative content analysis in 11 countries. *Journal of Cleaner Production*, 122, 186–200. <https://doi.org/10.1016/j.jclepro.2016.02.060>
- López, M. V., García, A., & Rodríguez, L. (2007). Sustainable development and corporate performance: A study based on the Dow Jones sustainability index. *Journal of Business Ethics*, 75(3), 285–300. <https://doi.org/10.1007/s10551-006-9253-8>
- Maignan, I. (2001). Consumers' perceptions of corporate social responsibilities: A cross-cultural comparison. *Journal of Business Ethics*, 30(1), 57–72. <https://doi.org/10.1023/a:1006433928640>
- Manetti, G. (2011). The quality of stakeholder engagement in sustainability reporting: Empirical evidence and critical points. *Corporate Social Responsibility and Environmental Management*, 18(2), 110–122. <https://doi.org/10.1002/csr.255>
- Martínez-Ferrero, J., García-Sánchez, I. M., & Ruiz-Barbadillo, E. (2018). The quality of sustainability assurance reports: The expertise and experience of assurance providers as determinants. *Business Strategy and the Environment*. <https://doi.org/10.1002/bse.2061>, 27, 1181–1196.
- Martínez-Ferrero, J., Ruiz-Cano, D., & García-Sánchez, I. M. (2016). The causal link between sustainable disclosure and information asymmetry: The moderating role of the stakeholder protection context. *Corporate Social Responsibility and Environmental Management*, 23(5), 319–332. <https://doi.org/10.1002/csr.1379>
- Martínez-Ferrero, J., Suárez-Fernández, O., & García-Sánchez, I. M. (2018). Obfuscation versus enhancement as corporate social responsibility disclosure strategies. *Corporate Social Responsibility and Environmental Management*. <https://doi.org/10.1002/csr.1697>
- Matten, D., & Moon, J. (2008). "Implicit" and "explicit" CSR: A conceptual framework for a comparative understanding of corporate social responsibility. *Academy of Management Review*, 33(2), 404–424. <https://doi.org/10.5465/amr.2008.31193458>

- McWilliams, A., & Siegel, D. (2000). Corporate social responsibility and financial performance: Correlation or misspecification? *Strategic Management Journal*, 21(5), 603–609. [https://doi.org/10.1002/\(SICI\)1097-0266\(200005\)21:5<603::AID-SMJ101>3.0.CO;2-3](https://doi.org/10.1002/(SICI)1097-0266(200005)21:5<603::AID-SMJ101>3.0.CO;2-3)
- Meek, G. K., Roberts, C. B., & Gray, S. J. (1995). Factors influencing voluntary annual report disclosures by US, UK and continental European multinational corporations. *Journal of International Business Studies*, 26(3), 555–572. <https://doi.org/10.1057/palgrave.jibs.8490186>
- Michelon, G., Pilonato, S., & Ricceri, F. (2015). CSR reporting practices and the quality of disclosure: An empirical analysis. *Critical Perspectives on Accounting*, 33, 59–78. <https://doi.org/10.1016/j.cpa.2014.10.003>
- Mio, C. (2010). Corporate social reporting in Italian multi-utility companies: An empirical analysis. *Corporate Social Responsibility and Environmental Management*, 17(5), 247–271. <https://doi.org/10.1002/csr.213>
- Morrison, L., Wilmshurst, T., & Shimeld, S. (2016). Environmental reporting through an ethical looking glass. *Journal of Business Ethics*, 150(4), 903–918. <https://doi.org/10.1007/s10551-016-3136-4>
- Moseñe, J. A., Burritt, R. L., Sanagustín, M. V., Moneva, J. M., & Tingey-Holyoak, J. (2013). Environmental reporting in the Spanish wind energy sector: An institutional view. *Journal of Cleaner Production*, 40, 199–211. <https://doi.org/10.1016/j.jclepro.2012.08.023>
- Naeem, M. A., & Welford, R. (2009). A comparative study of corporate social responsibility in Bangladesh and Pakistan. *Corporate Social Responsibility and Environmental Management*, 16(2), 108–122. <https://doi.org/10.1002/csr.185>
- Ndemanga, D. A., & Koffi, E. T. (2009). Ownership structure, industry sector and corporate social responsibility (CSR) practices: The case of Swedish listed companies.
- Neu, D., Warsame, H., & Pedwell, K. (1998). Managing public impressions: Environmental disclosures in annual reports. *Accounting, Organizations and Society*, 23(3), 265–282. [https://doi.org/10.1016/s0361-3682\(97\)00008-1](https://doi.org/10.1016/s0361-3682(97)00008-1)
- Nieto-Librero, A. B., Sierra, C., Vicente-Galindo, M. P., Ruiz-Barzola, O., & Galindo-Villardón, M. P. (2017). Clustering Disjoint HJ-Biplot: A new tool for identifying pollution patterns in geochemical studies. *Chemosphere*, 176, 389–396. <https://doi.org/10.1016/j.chemosphere.2017.02.125>
- O'Donovan, G. (2002). Environmental disclosures in the annual report: Extending the applicability and predictive power of legitimacy theory. *Accounting, Auditing & Accountability Journal*, 15(3), 344–371. <https://doi.org/10.1108/09513570210435870>
- O'Dwyer, B., Unerman, J., & Hession, E. (2005). User needs in sustainability reporting: Perspectives of stakeholders in Ireland. *European Accounting Review*, 14(4), 759–787. <https://doi.org/10.1080/09638180500104766>
- Orij, R. (2010). Corporate social disclosures in the context of national cultures and stakeholder theory. *Accounting, Auditing & Accountability Journal*, 23(7), 868–889. <https://doi.org/10.1108/09513571011080162>
- Orozco, Y. V. D., Acevedo, M. d. I. M. C., & Acevedo, J. A. R. (2014). Responsabilidad Social Empresarial: Teorías, índices, estándares y certificaciones. *Cuadernos de Administración*, 29(50), 196–206. <https://doi.org/10.25100/cdea.v29i50.55>
- Pellegrino, C., & Lodhia, S. (2012). Climate change accounting and the Australian mining industry: Exploring the links between corporate disclosure and the generation of legitimacy. *Journal of Cleaner Production*, 36, 68–82. <https://doi.org/10.1016/j.jclepro.2012.02.022>
- Perez-Batres, L. A., Doh, J. P., Miller, V. V., & Pisani, M. J. (2012). Stakeholder pressures as determinants of CSR strategic choice: Why do firms choose symbolic versus substantive self-regulatory codes of conduct? *Journal of Business Ethics*, 110(2), 157–172. <https://doi.org/10.1007/s10551-012-1419-y>
- Perez-Batres, L. A., Miller, V. V., & Pisani, M. J. (2011). Institutionalizing sustainability: An empirical study of corporate registration and commitment to the United Nations global compact guidelines. *Journal of Cleaner Production*, 19(8), 843–851. <https://doi.org/10.1016/j.jclepro.2010.06.003>
- Prado Lorenzo, J. M., García Sánchez, I. M., & Blázquez Zaballós, A. (2013). El impacto del sistema cultural en la transparencia corporativa. *Revista Europea de Dirección Y Economía de La Empresa*, 22(3), 143–154. <https://doi.org/10.1016/j.redee.2013.04.001>
- Prado-Lorenzo, J. M., & García-Sánchez, I.-M. (2010). The role of the board of directors in disseminating relevant information on greenhouse gases. *Journal of Business Ethics*, 97(3), 391–424. <https://doi.org/10.1007/s10551-010-0515-0>
- Prado-Lorenzo, J. M., García-Sánchez, I. M., & Gallego-Álvarez, I. (2009a). Características del consejo de administración e información en materia de responsabilidad social corporativa. *Spanish Journal of Finance and Accounting/Revista Española de Financiación Y Contabilidad*, 38(141), 107–135. <https://doi.org/10.1080/02102412.2009.10779664>
- Prado-Lorenzo, J. M., García-Sánchez, I. M., & Gallego-Álvarez, I. (2009b). Stakeholder engagement and corporate social responsibility reporting: The ownership structure effect. *Corporate Social Responsibility and Environmental Management*, 16(2), 94–107. <https://doi.org/10.1002/csr.189>
- Rasche, A. (2009). Toward a model to compare and analyze accountability standards—The case of the UN Global Compact. *Corporate Social Responsibility and Environmental Management*, 16(4), 192–205. <https://doi.org/10.1002/csr.202>
- Reverte, C. (2009). Determinants of corporate social responsibility disclosure ratings by Spanish listed firms. *Journal of Business Ethics*, 88(2), 351–366. <https://doi.org/10.1007/s10551-008-9968-9>
- Riahi-Belkaoui, A., & AlNajjar, F. K. (2006). Earnings opacity internationally and elements of social, economic and accounting order. *Review of Accounting and Finance*, 5(3), 189–203. <https://doi.org/10.1108/14757700610686408>
- Richerson, P. J., & Boyd, R. (2005). *Not by genes alone: How culture transformed human evolution*. Chicago: University of Chicago Press.
- Ringov, D., & Zollo, M. (2007). The impact of national culture on corporate social performance. *Corporate Governance: The International Journal of Business in Society*, 7(4), 476–485. <https://doi.org/10.1108/14720700710820551>
- Rodero, H. M., Sanz-Valero, J., & Galindo-Villardón, M. P. (2018). The methodological quality of systematic reviews indexed in the MEDLINE database: A multivariate approach. *The Electronic Library*, 36, 146–158. <https://doi.org/10.1108/EL-01-2017-0002>
- Rodrigue, M. (2014). Contrasting realities: Corporate environmental disclosure and stakeholder-released information. *Accounting, Auditing & Accountability Journal*, 27(1), 119–149. <https://doi.org/10.1108/AAJ-04-2013-1305>
- Rodríguez-Ariza, L., Aceituno, J. V. F., & Rubio, R. G. (2014). El consejo de administración y las memorias de sostenibilidad. *Revista de Contabilidad*, 17(1), 5–16. <https://doi.org/10.1016/j.rcsar.2013.02.002>
- Rodríguez-Rosa, M., Gallego-Álvarez, I., Vicente-Galindo, M. P., & Galindo-Villardón, M. P. (2017). Are social, economic and environmental well-being equally important in all countries around the world? A study by income levels. *Social Indicators Research*, 131(2), 543–565. <https://doi.org/10.1007/s11205-016-1257-x>
- Romolini, A., Fissi, S., & Gori, E. (2014). Scoring CSR reporting in listed companies—Evidence from Italian best practices. *Corporate Social Responsibility and Environmental Management*, 21(2), 65–81. <https://doi.org/10.1002/csr.1299>
- Russo-Spena, T., Tregua, M., & De Chiara, A. (2018). Trends and drivers in CSR disclosure: A focus on reporting practices in the automotive

- industry. *Journal of Business Ethics*, 151(2), 563–578. <https://doi.org/10.1007/s10551-016-3235-2>
- Salter, S. B., & Niswander, F. (1995). Cultural influence on the development of accounting systems internationally: A test of Gray's [1988] theory. *Journal of International Business Studies*, 26(2), 379–397. <https://doi.org/10.1057/palgrave.jibs.8490179>
- Scholten, B., & Kang, F.-C. (2013). Corporate social responsibility and earnings management: Evidence from Asian economies. *Corporate Social Responsibility and Environmental Management*, 20(2), 95–112. <https://doi.org/10.1002/csr.1286>
- Scott, W. R. (2008). Approaching adulthood: The maturing of institutional theory. *Theory and Society*, 37(5), 427–442. <https://doi.org/10.1007/s11186-008-9067-z>
- Secchi, D. (2006). The Italian experience in social reporting: An empirical analysis. *Corporate Social Responsibility and Environmental Management*, 13(3), 135–149. <https://doi.org/10.1002/csr.96>
- Singhapakdi, A., Vitell, S. J., & Leelakulthanit, O. (1994). A cross-cultural study of moral philosophies, ethical perceptions and judgements: A comparison of American and Thai marketers. *International Marketing Review*, 11(6), 65–78. <https://doi.org/10.1108/02651339410073015>
- Skouloudis, A., Evangelinos, K., & Kourmoussis, F. (2010). Assessing non-financial reports according to the Global Reporting Initiative guidelines: Evidence from Greece. *Journal of Cleaner Production*, 18(5), 426–438. <https://doi.org/10.1016/j.jclepro.2009.11.015>
- Su, S.-H. (2006). Cultural differences in determining the ethical perception and decision-making of future accounting professionals: A comparison between accounting students from Taiwan and the United States. *Journal of American Academy of Business*, 9(1), 147–158.
- Talbot, D., & Boiral, O. (2015a). GHG reporting and impression management: An assessment of sustainability reports from the energy sector. *Journal of Business Ethics*, 147(2), 1–17. <https://doi.org/10.1007/s10551-015-2979-4>
- Talbot, D., & Boiral, O. (2015b). Strategies for climate change and impression management: A case study among Canada's large industrial emitters. *Journal of Business Ethics*, 132(2), 329–346. <https://doi.org/10.1007/s10551-014-2322-5>
- Tejedor-Flores, N., Galindo-Villardón, P., & Vicente-Galindo, P. (2016). Sustainability multivariate analysis based on the global reporting initiative (GRI) framework, using as a case study: Brazil compared to Spain and Portugal. *Urban Regeneration & Sustainability*, 12, 307. <https://doi.org/10.2495/SDP-V12-N4-667-677/032>
- Tejedor-Flores, N., Vicente-Galindo, P., & Galindo-Villardón, P. (2017). Sustainability multivariate analysis of the energy consumption of Ecuador using MuSIASEM and biplot approach. *Sustainability*, 9(6), 984. <https://doi.org/10.3390/su9060984>
- Torres-Salinas, D., Robinson-García, N., Jiménez-Contreras, E., Herrera, F., & López-Cózar, E. D. (2013). On the use of biplot analysis for multivariate bibliometric and scientific indicators. *Journal of the American Society for Information Science and Technology*, 64(7), 1468–1479. <https://doi.org/10.1002/asi.22837>
- Trotman, K. T., & Bradley, G. W. (1981). Associations between social responsibility disclosure and characteristics of companies. *Accounting, Organizations and Society*, 6(4), 355–362. [https://doi.org/10.1016/0361-3682\(81\)90014-3](https://doi.org/10.1016/0361-3682(81)90014-3)
- Tsakumis, G. T. (2007). The influence of culture on accountants' application of financial reporting rules. *Abacus*, 43(1), 27–48. <https://doi.org/10.1111/j.1467-6281.2007.00216.x>
- Tsang, S., Welford, R., & Brown, M. (2009). Reporting on community investment. *Corporate Social Responsibility and Environmental Management*, 16(3), 123–136. <https://doi.org/10.1002/csr.178>
- Ullmann, A. A. (1985). Data in search of a theory: A critical examination of the relationships among social performance, social disclosure, and economic performance of US firms. *Academy of Management Review*, 10(3), 540–557. <https://doi.org/10.2307/258135>
- Van der Laan Smith, J., Adhikari, A., & Tonkar, R. H. (2005). Exploring differences in social disclosures internationally: A stakeholder perspective. *Journal of Accounting and Public Policy*, 24(2), 123–151. <https://doi.org/10.1016/j.jaccpubpol.2004.12.007>
- Vicente Galindo, P., Vaz, E., & de Noronha, T. (2015). How corporations deal with reporting sustainability: assessment using the multicriteria logistic biplot approach. *Systems*, 3(1), 6–26. <https://doi.org/10.3390/systems3010006>
- Vicente-Villardón, J. L. (2010). MultiBiplot: A package for multivariate analysis using biplots. In *Computer Software*. Spain: Departamento de Estadística. University of Salamanca. <http://Http://Biplot.Usal.Es/ClassicalBiplot/Index.Html>
- Vicente-Villardón, J. L., Galindo-Villardón, M. P., & Blázquez-Zaballos, A. (2006). *Logistic biplots. Multiple correspondence analysis and related methods* (pp. 503–521). London: Chapman & Hall. <https://doi.org/10.1201/9781420011319.ch23>
- Vitell, S. J., Paolillo, J. G., & Thomas, J. L. (2003). The perceived role of ethics and social responsibility: A study of marketing professionals. *Business Ethics Quarterly*, 13(1), 63–86. <https://doi.org/10.5840/beq20031315>
- Welford, R. (2004). Corporate social responsibility in Europe and Asia: Critical elements and best practice. *Journal of Corporate Citizenship*, 13, 31–47. <https://doi.org/10.9774/gleaf.4700.2004.sp.00007>
- White, A. L. (2006). Why we need global standards for corporate disclosure. *Law and Contemporary Problems*, 69(3), 167–186.
- Williams, C. A. (1999). The securities and exchange commission and corporate social transparency. *Harvard Law Review*, 112(6), 1197–1311. <https://doi.org/10.2307/1342384>
- Williams, G., & Zinkin, J. (2008). The effect of culture on consumers' willingness to punish irresponsible corporate behaviour: Applying Hofstede's typology to the punishment aspect of corporate social responsibility. *Business Ethics: A European Review*, 17(2), 210–226. <https://doi.org/10.1111/j.1467-8608.2008.00532.x>
- Williams, S. M., & Pei, C.-A. H. W. (1999). Corporate social disclosures by listed companies on their web sites: An international comparison. *The International Journal of Accounting*, 34(3), 389–419. [https://doi.org/10.1016/s0020-7063\(99\)00016-3](https://doi.org/10.1016/s0020-7063(99)00016-3)
- Xiao, J. Z., Gao, S. S., Heravi, S., & Cheung, Y. C. (2005). The impact of social and economic development on corporate social and environmental disclosure in Hong Kong and the UK. *Advances in International Accounting*, 18, 219–243. [https://doi.org/10.1016/s0897-3660\(05\)18011-8](https://doi.org/10.1016/s0897-3660(05)18011-8)
- Yadava, R. N., & Sinha, B. (2016). Scoring sustainability reports using GRI 2011 guidelines for assessing environmental, economic, and social dimensions of leading public and private Indian companies. *Journal of Business Ethics*, 138(3), 549–558. <https://doi.org/10.1007/s10551-015-2597-1>
- Yusoff, H., Othman, R., & Yatim, N. (2014). Culture and accountants' perceptions of environmental reporting practice. *Business Strategy and the Environment*, 23(7), 433–446. <https://doi.org/10.1002/bse.1793>

How to cite this article: Cubilla-Montilla M, Nieto-Librero A-B, Galindo-Villardón MP, Vicente Galindo MP, Garcia-Sanchez JM. Are cultural values sufficient to improve stakeholder engagement human and labour rights issues? *Corp Soc Resp Env Ma*. 2019;1–18. <https://doi.org/10.1002/csr.1733>

ARTÍCULOS SOMETIDOS

Trabajos de investigación en proceso de Publicación:

#1

Title: Dynamic analysis of worldwide quality of life based on the level of IDH. A study using multivariate techniques

Authors: Mitzi Cubilla-Montilla
Purificación Galindo-Villardón
Ana Belén Nieto-Librero
Isabel García-Sánchez

Journal: Journal of Cleaner Production

#2

Title: What the companies do not tell us about their environmental behaviour and the role that the institutional pressures could perform

Authors: Mitzi Cubilla-Montilla
Ana Belén Nieto-Librero
Purificación Galindo-Villardón
Purificación Vicente Galindo
Isabel García-Sánchez

Journal: Social Indicators Research

#3

Title: HJ Biplot from a Sparse Optimization Viewpoint

Authors: Mitzi Cubilla-Montilla
Purificación Galindo-Villardón
Ana Belén Nieto-Librero

Journal: Por determinar

PARTICIPACIÓN EN CONGRESOS NACIONALES E INTERNACIONALES (2015-2019)

Año	CONGRESO	FECHA	Título de la Comunicación o Póster	Modalidad
2015	XXX Foro Internacional de Estadística. Acapulco, México	14 al 18 de septiembre-2015	Análisis de la competitividad en el sector de viajes y turismo de América, a través del GGBiplot	Comunicación Oral
2016	XXVI Simposio Internacional de Estadística. Sucre, Colombia	8 al 12 de agosto-2016.	La competitividad global en América Latina: un análisis integral a través del MULTIBIPLLOTGUI	Comunicación Oral
2016	XXXVI Congreso Nacional de estadística e investigación operativa. Toledo, España	5 al 7 de septiembre-2016	Biplot dinámico en el estudio del índice de competitividad global en América Latina.	Formato Póster
2017	IV Conferencia Española de Biometría. CEB 2017. Sevilla, España	13 al 15 de septiembre-2017	Sustainable Tourism and Competitiveness. An analysis from the Biplot perspective to model the interaction countries-by-competitive factors	Formato Póster
2018	I Congreso Internacional "Políticas Públicas en defensa de la inclusión, la diversidad y el género". Salamanca, España	22 al 24 de julio-2018	¿Igualdad de género?: La realidad a través de los datos	Comunicación Oral
2019	IV International Workshop on Proximity Data, Multivariate Analysis and Classification. USAL. España	25 y 26 de abril-2019	Contributions to biplot Analysis: Disjoint Solutions and Sparse HJ Biplot	Formato Póster
2019	II Congreso Internacional "Políticas Públicas en defensa de la inclusión, la diversidad y el género". Salamanca, España	15 y 16 de julio-2019	Perfil estadístico de la desigualdad de género	Comunicación Oral

BIBLIOGRAFÍA

- Aitchison, J., Barceló-Vidal, C., Egozcue, J. J., & Pawlowsky-Glahn, V. (2002). A concise guide for the algebraic-geometric structure of the simplex, the sample space for compositional data analysis. In *Proceedings of Eighth Annual Conference of the International Association for Mathematical Geology*. (Vol. 2, pp. 387-392).
- Aitchison, J. (1982). The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2), 139–160. <https://doi.org/10.1111/j.2517-6161.1982.tb01195.x>
- Aitchison, J. (1997). The one-hour course in compositional data analysis or compositional data analysis is simple. In *in: V. Pawlowsky-Glahn (Ed.), Proceedings of IAMG'97 – The III Annual Conference of the International Association for Mathematical Geology*. (Vols. I, II and addendum, p. 3-35). Barcelona: International Center for Numerical Methods in Engineering (CIMNE).
- Aitchison, J., & Greenacre, M. (2002). Biplots of compositional data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 51(4), 375–392. <https://doi.org/10.1111/1467-9876.00275>
- Amaratunga, D., & Cabrera, J. (2016). High-dimensional data. *Journal of the National Science Foundation of Sri Lanka*, 44(1).
- Amaro, I. R. (2001). *Manova biplot para diseños con varios factores, basado en modelos lineales generales multivariantes*. [Tesis doctoral]. Universidad de Salamanca, España.
- Amaro, I. R., Vicente-Villardón, J. L. y, & Galindo-Villardón, P. (2004). MANOVA BILOT para arreglos de tratamientos con dos factores basado en modelos lineales generales multivariantes. *Interciencia*, 29(1), 26–32.
- Baccalá, N. (2004). *Contribuciones al Análisis de Matrices de Datos Multivía: tipología de las variables*. [Tesis doctoral]. Universidad de Salamanca, España, Spain.
- Becker, H. C. (1981). Correlations among some statistical measures of phenotypic stability. *Euphytica*, 30(3), 835–840.
- Benzécri, J. P. (1973). *L'analyse des données* (Vol. 2). Paris: Dunod.

- Berman, J. J. (2013). *Principles of big data: preparing, sharing, and analyzing complex information*. Newnes.
- Billheimer, D., Guttorp, P., & Fagan, W. F. (2001). Statistical Interpretation of Species Composition. *Journal of the American Statistical Association*, 96(456), 1205–1214. <https://doi.org/10.1198/016214501753381850>
- Blázquez, A. (1998). *Análisis biplot basado en modelos lineales generalizados*. [Tesis doctoral]. Universidad de Salamanca, España.
- Bodor, A., Csabai, I., Mahoney, M. W., & Solymosi, N. (2012). rCUR: an R package for CUR matrix decomposition. *BMC Bioinformatics*, 13(1), 103. <https://doi.org/10.1186/1471-2105-13-103>
- Braak, C. J. Ter, & Looman, C. W. (1994). Biplots in reduced-rank regression. *Biometrical Journal*, 36(8), 983–1003.
- Bradu, D., & Gabriel, K. R. (1974). Simultaneous statistical inference on interactions in two-way analysis of variance. *Journal of the American Statistical Association*, 69(346), 428–436.
- Bradu, D., & Gabriel, K. R. (1978). The biplot as a diagnostic tool for models of two-way tables. *Technometrics*, 20(1), 47–68.
- Breiman, L. (1995). Better Subset Regression Using the Nonnegative Garrote. *Technometrics*, 37(4), 373–384. <https://doi.org/10.1080/00401706.1995.10484371>
- Cadima, J., & Jolliffe, I. T. (1995). Loading and correlations in the interpretation of principle compenents. *Journal of Applied Statistics*, 22(2), 203–214. <https://doi.org/10.1080/757584614>
- Cárdenas, O. C., & Galindo, M. P. (2004). Biplot con información externa basado en modelos bilineales generalizados.
- Cárdenas, O., Noguera, C., Galindo, M. P., & Vicente-Villardón, J. L. (2003). El uso de información externa en aproximaciones Biplot. *Revista Venezolana de Análisis de Coyuntura*, 9(2), 257–276.
- Cárdenas, O., Noguera, C., Galindo, P., & Vicente-Villardón, J. L. (2006). An alternative to principal components regression based on Regression Biplot. *INTERCIENCIA*, 31(3), 160–167.
- Carlier, A., & Kroonenberg, P. M. (1996). Decompositions and biplots in three-

- way correspondence analysis. *Psychometrika*, 61(2), 355–373.
- Chessel, D., Dufour, A.B., Dray, S., Jombart, T., Lobry, J.R., Ollier, S. y, & Thioulouse, J. (2013). *ade4*. R package version 1.5-2: analysis of ecological data: exploratory and Euclidean methods in environmental sciences. cran.r-project.org/package=ade4, 2013.
- Chessel, D., Dufour, A., & Thioulouse, J. (2004). The *ade4* package-I-One-table methods. *Pdfs.Semanticscholar.Org*.
- Choulakian, V. (1996). Generalized bilinear models. *Psychometrika*, 61(2), 271–283.
- Cortés-Rodríguez, M., & Sánchez-Barba, M. (2013). Biplot de datos composicionales: una herramienta útil en el estudio de test psicológicos. Universidad de Salamanca, España.
- Crossa, J., Cornelius, P. L., & Yan, W. (2002). Biplots of linear-bilinear models for studying crossover genotypex environment interaction. *Crop Science*, 42(2), 619–633.
- Crossa, J., Gauch, H. G., & Zobel, R. W. (1990). Additive main effects and multiplicative interaction analysis of two international maize cultivar trials. *Crop Science*, 30, 493–500.
- Cubilla-Montilla, M., Nieto-Librero, A.-B., Galindo-Villardón, M. P., Vicente Galindo, M. P., & Garcia-Sanchez, I.-M. (2019). Are cultural values sufficient to improve stakeholder engagement human and labour rights issues? *Corporate Social Responsibility and Environmental Management*, 26(4), 938–955. <https://doi.org/10.1002/csr.1733>
- d'Aspremont, A., El Ghaoui, L., Jordan, M. I., & Lanckriet, G. R. G. (2007). A Direct Formulation for Sparse PCA Using Semidefinite Programming. *SIAM Review*, 49(3), 434–448. <https://doi.org/10.1137/050645506>
- De Falguerolles, A. (1996). Generalized bilinear models and generalized biplots: some examples. *Publications Du Laboratoire de Statistique et Probabilités*.
- Demey, J. R. (2008). Diversidad genética en bancos de Germoplasma: un enfoque Biplot.
- Denis, J. B. (1991). Ajustements de modèles linéaires et bilinéaires sous contraintes linéaires avec données manquantes. *Revue de Statistique*

Appliquée, 39(2), 5–24.

- Donoho, D. L., & Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3), 425–455. <https://doi.org/10.1093/biomet/81.3.425>
- Dray, S., Dufour, A.B., Chessel, D. (2007). The ade4 package-II: Two-table and K-table methods. *R Journal*, 7(2), 47–52.
- Drineas, P., Kannan, R., & Mahoney, M. (2006). Fast Monte Carlo algorithms for matrices III: Computing a compressed approximate matrix decomposition. *SIAM, Journal on Computing*, 36(1), 184–206.
- Drineas, P., Mahoney, M. W., & Muthukrishnan, S. (2008). Relative-Error CUR Matrix Decompositions. *SIAM Journal on Matrix Analysis and Applications*, 30(2), 844–881. <https://doi.org/10.1137/07070471X>
- Eckart, C., & Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3), 211–218.
- Eckart, C., & Young, G. (1939). A principal axis transformation for non-Hermitian matrices. *Bulletin of the American Mathematical Society*, 45(2), 118–121.
- Efron, B., & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, 37(1), 36–48.
- Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32(2), 407–499.
- Efron, B., & Tibshirani, R. (1994). An introduction to the bootstrap.
- Efron, Bradley. (1979). Computers and the Theory of Statistics: Thinking the Unthinkable. *SIAM Review*, 21(4), 460–480. <https://doi.org/10.1137/1021092>
- Egido, J. (2014). dynBiplotGUI. R package version 1.0.1. cran.r-project.org/web/packages/dynBiplotGUI.
- Erichson, N., Zheng, P., Manohar, K., Brunton, S., Kuetz, J., & Aravkin, A. (2018). Sparse Principal Component Analysis via Variable Projection. *ArXiv Preprint ArXiv:1804.00341*.
- Fan, J., & Li, R. (2001). Variable Selection via Nonconcave Penalized

- Likelihood and its Oracle Properties. *Journal of the American Statistical Association*, 96(456), 1348–1360. <https://doi.org/10.1198/016214501753382273>
- Farcomeni, A. (2009). An exact approach to sparse principal component analysis. *Computational Statistics*, 24(4), 583–604. <https://doi.org/10.1007/s00180-008-0147-3>
- Faria, J. C., & Demetrio, C. G. B. (2012). Biplot of multivariate data based on principal components analysis. R package version 1.02. URL <http://cran.r-project.org/package=bpca>.
- Fernández-Gómez, M. J. (1995). *Contribuciones al análisis multivariante directo del gradiente mediante estudio combinado de configuraciones espaciales*. [Tesis doctoral]. Universidad de Salamanca, España.
- Fisher, R. A. (1936). The use of mutiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2), 179–188. <https://doi.org/10.1111/j.1469-1809.1936.tb02137.x>
- Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1), 1.
- Frieze, A., Kannan, R., & Vempala, S. (2004). Fast Monte-Carlo algorithms for finding low-rank approximations. *Journal of the ACM (JACM)*, 51(6), 1025–1041.
- Frutos, E., & Galindo, M. P. (2013). GGEBiplotGUI. R package version 1.0-6: interactive GGE biplots in R. cran.r-project.org/package=GGEBiplotGUI,.
- Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58(3), 453–467.
- Gabriel, K. R. (1972). Analysis of Meteorological Data by Means of Canonical Decomposition and Biplots. *Journal of Applied Meteorology*, 11(7), 1071–1077. [https://doi.org/10.1175/1520-0450\(1972\)011<1071:AOMDBM>2.0.CO;2](https://doi.org/10.1175/1520-0450(1972)011<1071:AOMDBM>2.0.CO;2)
- Gabriel, K. R. (1995). “MANOVA biplots for two-way contingency tables.” In: *W.J. Krzanowski (Ed.), Recent Advances in Descriptive Multivariate Analysis*, Oxford: Clarendon Press.
- Gabriel, K. R. (1998). Generalised bilinear regression. *Biometrika*, 85(3), 689–

700.

- Gabriel, K. R., & Odoroff, C. L. (1990). Biplots in biomedical research. *Statistics in Medicine*, 9(5), 469–485. <https://doi.org/10.1002/sim.4780090502>
- Galindo, M. P. (1986). Una alternativa de representacion simultanea: HJ-Biplot. *Qüestió: Quaderns d'estadística i Investigació Operativa*, 10(1), 13–23.
- Galindo, M. P., & Cuadras, C. M. (1986). Una extensión del método Biplot y su relación con otras técnicas. *Publicaciones de Bioestadística y Biomatemática*, 17.
- Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2), 137–144.
- Gauch, H. G. (1988). Model selection and validation for yield trials with interaction. *Biometrics*, 705–715.
- Gauch, H. G., & Zobel, R. W. (1989). Accuracy and selection success in yield trial analyses. *Theoretical and Applied Genetics*, 77(4), 473–481. <https://doi.org/10.1007/BF00274266>
- Gollob, H. F. (1968). A statistical model which combines features of factor analytic and analysis of variance techniques. *Psychometrika*, 33(1), 73–115.
- Goreinov, S. ., Tyrtshnikov, E. E., & Zamarashkin, N. L. (1997). A theory of pseudoskeleton approximations. *Linear Algebra and Its Applications*, 261(1–3), 1–21.
- Goreinov, S. A., & Tyrtshnikov, E. E. (2001). The maximal-volume concept in approximation by low-rank matrices. *Contemporary Mathematics*, 280, 47–52.
- Gower, J. C. (1992). Generalized biplots. *Biometrika*, 79(3), 475–493.
- Gower, J. C., & Hand, D. J. (1995). *Biplots* (Vol. 54). CRC Press.
- Gower, J. C., & Hand, D. J. (1996). *Biplots*. Chapman&Hall, London UK.
- Gower, J. C., & Harding, S. A. (1988). Nonlinear biplots. *Biometrika*, 75(3), 445–455.
- Graffelman, J. (2012). Calibrate: calibration of scatterplot and biplot axes. R

- package version 1(1). URL cran.r-project.org/package=calibrate, 2012.
- Greenacre, M. J. (1984). *Correspondence analysis*. London: Academic Press.
- Greenacre, M. J. (1993). Biplots in correspondence analysis. *Journal of Applied Statistics*, 20(2), 251–269. <https://doi.org/10.1080/02664769300000021>
- Greenacre, M., & Nenadic, O. (2012). The ca R package version 0.53: simple, multiple and joint correspondence analysis. cran.r-project.org/package=ca.
- Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2015). The rise of “big data” on cloud computing: Review and open research issues. *Information Systems*, 47, 98–115.
- Hausman, R. E. (1982). Constrained multivariate analysis. In: Zanakis SH, Rustagi JS (eds). *Optimisation in Statistics*. North-Holland, Amsterdam, 137–151.
- Hernández, J. C., & Vicente-Villardón, J. L. (2013a). NominalLogisticBiplot. R package version 0.1: biplot representations of categorical data. cran.r-project.org/web/packages/NominalLogisticBiplot/index.html.
- Hernández, J. C., & Vicente-Villardón, J. L. (2013b). OrdinalLogisticBiplot. R package version 0.2: ordinal logistic biplots. cran.r-project.org/web/packages/OrdinalLogisticBiplot/index.html.
- Hernández, S. (2005). *Biplots robustos*. [Tesis doctoral]. Universidad de Salamanca, España.
- Hernández, S., & Galindo-Villardón, M. P. (2006). BIPROB: UN MÉTODO PARA OBTENER UN BILOT ROBUSTO. *Investigación Operacional*, 27(3), 287–299.
- Hernández Sánchez, J. C. (2016). *Biplot logístico para datos nominales y ordinales*. [Tesis Doctoral]. Universidad de Salamanca, España.
- Hernández Suárez, M., Molina Pérez, D., Rodríguez-Rodríguez, E., Díaz Romero, C., Espinosa Borreguero, F., & Galindo-Villardón, P. (2016). The Compositional HJ-Biplot—A New Approach to Identifying the Links among Bioactive Compounds of Tomatoes. *International Journal of Molecular Sciences*, 17(11), 1828.
- Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for

- Nonorthogonal Problems. *Technometrics*, 12(1), 55–67.
<https://doi.org/10.1080/00401706.1970.10488634>
- Hofstede, G. (2011). Dimensionalizing cultures: The Hofstede model in context. *Readings in Psychology and Culture*, 2(1), 8.
- Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6), 417.
- Jolliffe, I. T. (1995). Rotation of principal components: choice of normalization constraints. *Journal of Applied Statistics*, 22(1), 29–35.
<https://doi.org/10.1080/757584395>
- Jolliffe, I. T., Trendafilov, N. T., & Uddin, M. (2003). A Modified Principal Component Technique Based on the LASSO. *Journal of Computational and Graphical Statistics*, 12(3), 531–547.
<https://doi.org/10.1198/1061860032148>
- Jolliffe, I. T., & Uddin, M. (2000). The Simplified Component Technique: An Alternative to Rotated Principal Components. *Journal of Computational and Graphical Statistics*, 9(4), 689–710.
<https://doi.org/10.1080/10618600.2000.10474908>
- Kempton, R. A. (1984). The use of biplots in interpreting variety by environment interactions. *The Journal of Agricultural Science*, 103(1), 123–135.
- La Grange, A. M., Le Roux, N. J., Rousseeuw, I. R., & Tukey, J. W. (2009). BiplotGUI: interactive biplots in R. R package version 0.0-7. cran.r-project.org/package=BiplotGUI.
- Mahoney, M. W., & Drineas, P. (2009). CUR matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 106(3), 697–702.
<https://doi.org/10.1073/pnas.0803205106>
- Mahoney, M. W., Maggioni, M., & Drineas, P. (2008). Tensor-CUR Decompositions for Tensor-Based Data. *SIAM Journal on Matrix Analysis and Applications*, 30(3), 957–987. <https://doi.org/10.1137/060665336>
- Mandel, J. (1971). A New Analysis of Variance Model for Non-additive Data. *Technometrics*, 13(1), 1–18.
<https://doi.org/10.1080/00401706.1971.10488751>
- Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate Analysis*. Press

Inc. London.

- Markos, A. (2012). Package caGUI. The GUI ca R package version 0.1-4: a Tcl/Tk GUI for the functions. cran.r-project.org/package=caGUI, 2012.
- Martín-Rodríguez, J. (1996). *Contribuciones a la integración de subespacios desde una perspectiva biplot*. [Tesis doctoral]. Universidad de Salamanca, España.
- McCabe, G. P. (1984). Principal Variables. *Technometrics*, 26(2), 137–144. <https://doi.org/10.1080/00401706.1984.10487939>
- Moghaddam, B., Weiss, Y., & Avidan, S. (2006). Spectral bounds for sparse PCA: Exact and greedy algorithms. *Advances in Neural Information Processing Systems*.
- Nenadic, O., & Greenacre, M. (2007). Correspondence analysis in R, with two- and three-dimensional graphics: The ca package. *Journal of Statistical Software*, 20(3), 1–13.
- Nieto-Librero, A. B. (2015). *Versión inferencial de los métodos Biplot basada en remuestreo Bootstrap y su aplicación a tablas de tres vías*. [Tesis doctoral]. Universidad de Salamanca, España.
- Nieto-Librero, A. B., Baccalá, N., & Galindo, M. P. (2012). MultibiplotGUI. R package version 0.0-1: Multibiplot Analysis. cran.r-project.org/package=multibiplotGUI.
- Nieto-Librero, A. B., & Galindo-Villardón, P. (2015). biplotbootGUI. R package version 1.0: Bootstrap on Classical Biplots and Clustering Disjoint Biplot. cran.r-project.org/web/packages/biplotbootGUI/.
- Nieto-Librero, A. B., Sierra, C., Vicente-Galindo, M. P., Ruíz-Barzola, O., & Galindo-Villardón, M. P. (2017). Clustering Disjoint HJ-Biplot: A new tool for identifying pollution patterns in geochemical studies. *Chemosphere*, 176, 389–396.
- Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., Minchin, P. R., ... Wagner, H. (2013). Package vegan. *Comunity Ecology Package, Version 2.9*. R package version 2.0-8. cran.r-project.org/package=vegan.
- Pawlowsky-Glahn, V., & Egozcue, J. J.-S. (2001). Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment*, 15(5), 384–398.

- Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11), 559–572.
- Qi, X., Luo, R., & Zhao, H. (2013). Sparse principal component analysis by choice of norm. *Journal of Multivariate Analysis*, 114:127-150. <https://doi.org/10.1016/j.jmva.2012.07.004>
- Quenouille, M. H. (1950). An application of least squares to family diet surveys. *Econometrica: Journal of the Econometric Society*, 27–44.
- R-TEAM. (2014). A language and environment for statistical computing. R foundation for statistical computing. Vienna, Austria, 2014. Retrieved from www.R-project.org.
- Rodríguez-Mazahua, L., Rodríguez-Enríquez, C.-A., Sánchez-Cervantes, J. L., Cervantes, J., García-Alcaraz, J. L., & Alor-Hernández, G. (2016). A general perspective of Big Data: applications, tools, challenges and trends. *The Journal of Supercomputing*, 72(8), 3073–3113.
- Shen, H., & Huang, J. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis*, 99(6), 1015–1034.
- Stewart, G. W. (1999). Four algorithms for the the efficient computation of truncated pivoted QR approximations to a sparse matrix. *Numerische Mathematik*, 83(2), 313–323. <https://doi.org/10.1007/s002110050451>
- Talia, D. (2013). Clouds for scalable big data analytics. *Computer*, 46(5), 98–101.
- Ter Braak, C. J. (1986). Canonical Correspondence Analysis: a new eigenvector technique for Multivariate Direct Gradient Analysis. *Ecology*, 67(5), 1167–1179.
- Ter Braak, C. J. (1990). Interpreting canonical correlation analysis through biplots of structure correlations and weights. *Psychometrika*, 55(3), 519–531.
- Thioulouse, J., & Dray, S. (2007). Interactive multivariate data analysis in R with the ade4 and ade4TkGUI packages. *Journal of Statistical Software*, 22(5), 1–14.
- Thioulouse, J., & Dray, S. (2012). ade4TkGUI. R package version 0.2-6: ade4

- Tcl/Tk graphical user interface. cran.r-project.org/package=ade4TkGUI, 2012.
- Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267–288. <https://doi.org/10.1111/j.2517-6161.1996.tb02080.x>
- Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3), 273–282.
- Torgerson, W. S. (1952). Multidimensional scaling: I. Theory and method. *Psychometrika*, 17(4), 401–419. <https://doi.org/10.1007/BF02288916>
- Trendafilov, N. T. (2014). From simple structure to sparse components: A review. *Computational Statistics*, 29(3–4), 431–454.
- Tukey, J. (1958). Bias and confidence in not quite large samples. *Ann. Math. Statist.*, 29, 614.
- Vairinhos, V. M. (2003). Desarrollo de un sistema para minería de datos basado en los métodos Biplot.
- Van Eeuwijk, F. A. (1995). Multiplicative interaction in generalized linear models. *Biometrics*, 1017–1032.
- Vicente-Tavera, S. (1992). *Las técnicas de representación de datos multidimensionales en el estudio del Índice de Producción Industrial en la CEE* (Tesis Doctoral). Universidad de Salamanca, España.
- Vicente-Villardón, J. L. (1992). *Una alternativa a las técnicas factoriales clásicas basada en una generalización de los métodos BIPLLOT. [Tesis doctoral]. Universidad de Salamanca, España.*
- Vicente-Villardón, J. L. (2001). Biplot for binary data based on logistic response surfaces. In *Salamanca Statistics Seminar IV: Advances in Multivariate Analysis*. Salamanca, Spain.
- Vicente-Villardón, J. L. (2010). MULTBIPLLOT: package for multivariate analysis using biplots. Departamento de Estadística. Universidad de Salamanca. Retrieved from <http://biplot.usal.es/multbiplot/introduction.html>
- Vicente-Villardón, J. L. (2017). MultBiplotR: Multivariate Analysis using Biplot. R package version 0.1.0. <http://biplot.usal.es/classicalbiplot/multbiplot-in>

r/.

- Vicente-Villardón, J. L., Galindo-Villardón, M. P., & Blázquez-Zavallos, A. (2004). Constrained Logistic Biplots. *In SALAMANCA STATISTICS SEMINAR V. Advances in Descriptive Multivariate Analysis . Universidad de Salamanca, España.*
- Vicente-Villardón, J. L., Galindo-Villardón, M. P., & Blázquez-Zavallos, A. (2006). Logistic biplots. *Multiple Correspondence Analysis and Related Methods*. London: Chapman & Hall. 503-521.
- Vichi, M., & Saporta, G. (2009). Clustering and disjoint principal component analysis. *Computational Statistics & Data Analysis*, 53(8), 3194–3208.
- Vines, S. K. (2000). Simple principal components. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 49(4), 441–451. <https://doi.org/10.1111/1467-9876.00204>
- Witten, D. M., Tibshirani, R., & Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3), 515–534.
- Yan, W. (2001). GGEbiplot—a Windows application for graphical analysis of multi-environment trial data and other types of two-way data. *Agronomy*, 93(5), 1111–1118.
- Yan, W. (2002). Singular-value partitioning in biplot analysis of multi-environment trial data. *Agronomy Journal*, 94(5), 990–996.
- Yan, W., Cornelius, P. L., Crossa, J., & Hunt, L. A. (2001). Two types of GGE biplots for analyzing multi-environment trial data. *Crop Science*, 41, 656–663.
- Yan, W., & Hunt, L. A. (2002). Biplot analysis of diallel data. *Crop Science*, 42(1), 21–30.
- Yan, W., Hunt, L. A., Sheng, Q., & Szlavnic, Z. (2000). Cultivar evaluation and mega-environment investigation based on the GGE biplot. *Crop Science*, 40, 597–605.
- Yan, W., & Kang, M. S. (2002). *GGE biplot analysis: A graphical tool for breeders, geneticists, and agronomists*. CRC Press.
- Yan, W., & Tinker, N. A. (2006). Biplot analysis of multi-environment trial data:

- Principles and applications. *Canadian Journal of Plant Science*, 86(3), 623–645.
- Yang, J., Rübél, O., Prabhat, Mahoney, M. W., & Bowen, B. P. (2015). Identifying Important Ions and Positions in Mass Spectrometry Imaging Data Using CUR Matrix Decompositions. *Analytical Chemistry*, 87(9), 4658–4666. <https://doi.org/10.1021/ac5040264>
- Young, G., & Householder, A. S. (1938). Discussion of a set of points in terms of their mutual distances. *Psychometrika*, 3(1), 19–22.
- Zhang, Z., Xu, Y., Yang, J., Li, X., & Zhang, D. (2015). A survey of sparse representation: algorithms and applications. *IEEE*, 490–530.
- Zobel, R. W., Wright, M. J., & Gauch, H. G. (1988). Statistical analysis of a yield trial. *Agronomy Journal*, 80, 388–393.
- Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society*., 67(2), 301–320.
- Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse Principal Component Analysis. *Journal of Computational and Graphical Statistics*, 15(2), 265–286. <https://doi.org/10.1198/106186006X113430>