

Curso Estadística IV
Sociología
Universidad
Alberto Hurtado

Profesora
Carolina Aguilera
caguilera@uahurtado.cl

Ayudantes
Vicente Díaz – vidiazam@alumnos.uahurtado.cl
Miguel Tognarelli – mtognare@alumnos.uahurtado.cl



Clase 4

3 sept

- ¿Preguntas AC simple y múltiple?
- Análisis de Componentes Principales (PCA)
- Ejercicio práctico



CONTENIDOS DEL CURSO

Método	Objetivo	Tipo de variables	Tipo de resultado	Medida base / criterio	Naturaleza del método
		Representar gráficamente asociaciones entre categorías de variables nominales	Nominales (categóricas) en tablas de contingencia	Mapa perceptual de categorías y casos en espacios 2 dimensiones	Distancia chi-cuadrado
		Reducir la dimensionalidad manteniendo la mayor varianza posible	Cuantitativas continuas (o cuantitativas ordinales tratadas como continuas)	Nuevas variables ("componentes principales") no correlacionadas	Exploratorio, descriptivo
		Identificar factores latentes que explican las correlaciones entre variables observadas	Cuantitativas continuas	Factores latentes y cargas factoriales	Exploratorio, modelo estadístico implícito
		Agrupar casos (o variales) en función de su similitud	Nominales o continuas o mixtas (según el tipo de distancia/similitud elegido)	Grupos o clústeres de casos similares entre si	Exploratorio, descriptivo o confirmatorio

Descripción general



Calcula relaciones entre variables, agrupando variables que están correlacionadas



Se usa en **análisis exploratorios cuando tengo muchas variables** -de intervalo (o con precaución ordinales)- **que están midiendo algo parecido** (están CORRELACIONADAS entre si)



Análisis exploratorio para **encontrar componentes**, que son variables nuevas que explican un fenómeno con menos variables. Se construye a partir de las correlaciones entre las variables.



Se usa para variables de intervalo (aunque se puede aplicar con precaución para variables ordinales como escalas Likert con al menos 5 categorías, idealmente 7 ó 10)



Sirve para armar índices como combinación lineal de variables, sobre un fenómeno como que es de carácter multidimensional (como por ejemplo capital social).



Las nuevas variables (componentes) permiten **observar perfiles o grupos de casos, según los pesos de las variables originales en el plano conformado por los componentes principales**

Análisis de componentes principales

Se crean nuevas variables que se llaman “componentes principales”. Estas **están correlacionadas con las variables originales, pero no correlacionadas entre sí.**

La lógica matemática descansa en el **cálculo de autovalores y autovectores** (propios) de las matrices de correlación y de covarianza

Estas nuevas variables no son observables de manera directa si no que son una construcción matemática: **son una combinación lineal de las variables originales.**

El número de componentes principales a elegir se puede determinar con pruebas estadísticas

El PCA busca **maximizar la varianza** de los datos en el espacio multidimensional.

La varianza de las variables afecta a los resultados de un PCA y siempre tendrá una mayor influencia en la generación de un componente dado aquella variable con más varianza.

Análisis de componentes principales

El PCA se basa en la **matriz de correlaciones** (o covarianzas).

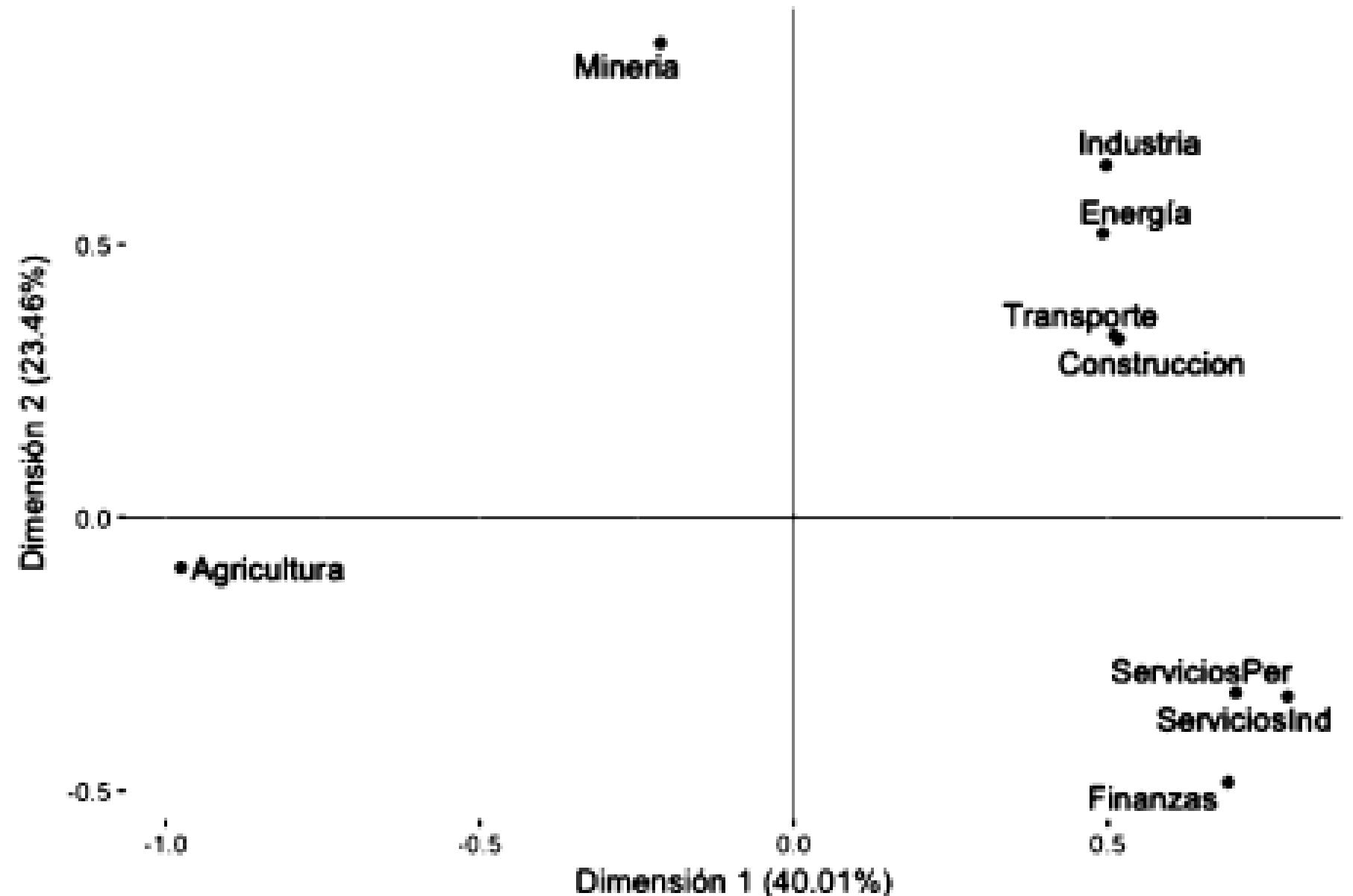
Se deben “centrar” las variables en 0 (que el promedio de cada una = 0), para las medias no influyan en los productos cruzados y no distorsionan la estimación de la covarianza.

Se deben además estandarizar las variables hace cuando las variables están en escalas diferentes (ej. ingreso en pesos vs. edad en años), porque si no se estandarizan, la variable con mayor varianza domina el análisis (varianza = 1)

Estructura sectorial del empleo en Europa

Distribución del empleo (% de personas empleadas en cada sector) en una serie de países europeos. El objetivo es determinar si existen perfiles de empleo que sirvan además para categorizar a los países en función de los mismos.

Figura 11.9.: Gráfico de las cargas sobre las dos primeras componentes



Dimensiones: El gráfico está organizado por las dos primeras componentes principales (calculadas por el modelo) (aparece el % de varianza explicada por cada dimensión)

Cargas: el gráfico muestra el “peso” (la importancia) que tiene cada VARIABLE en cada dimensión (distancia a cada eje). Esta dada por la correlación entre la variable y la dimensión.

Agricultura tiene alto peso en Dim 1 y Minería tiene poca carga en Dim 1. En Dim 1 valores coordenadas son (-). Las otras tienen valores (+)

Estructura sectorial del empleo en Europa

Distribución del empleo por sectores en una serie de países europeos. El objetivo es determinar si existen perfiles de empleo que sirvan además para categorizar a los países en función de los mismos.

Cuadro 11.8.: Variables de la base de datos de empleo

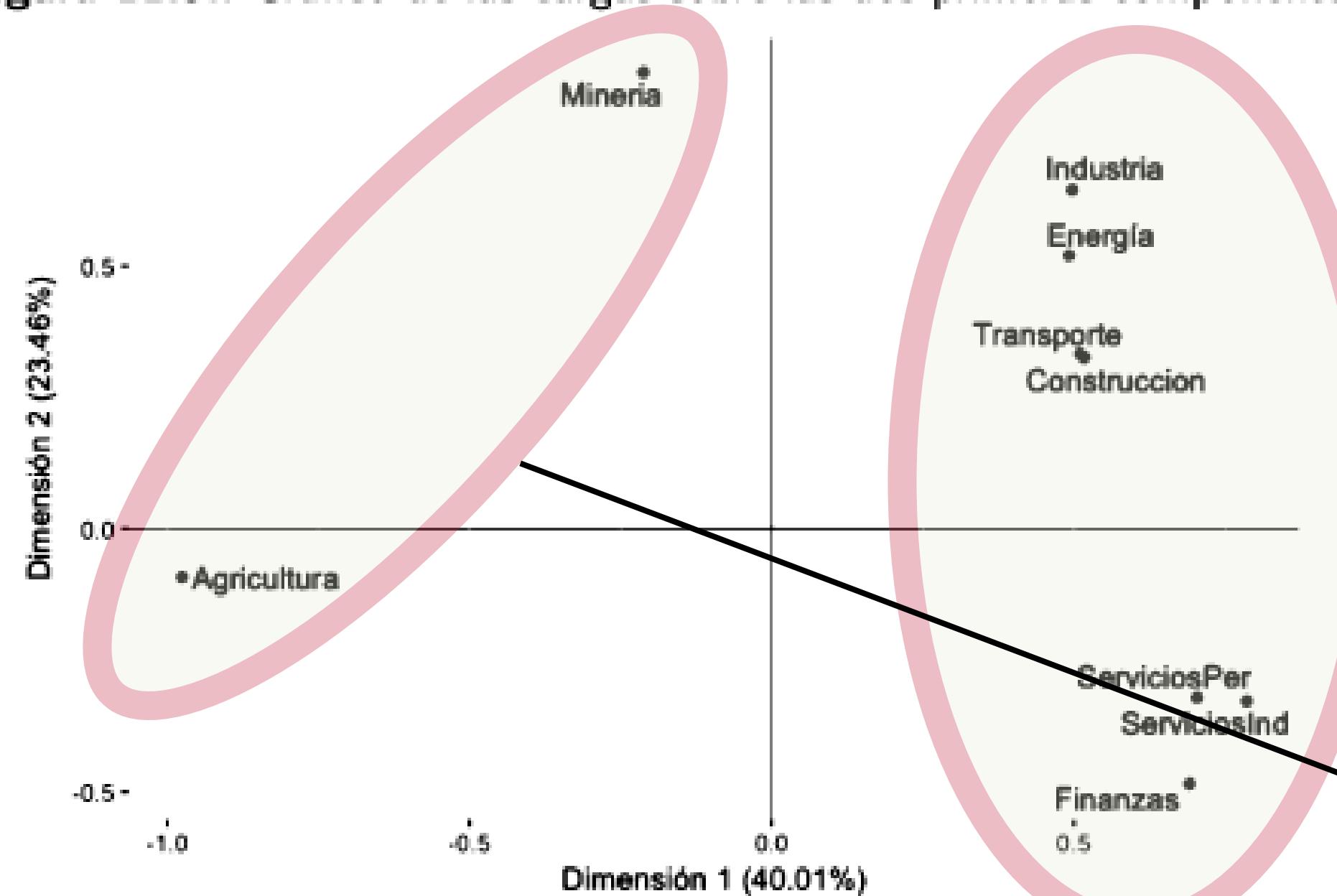
Variable	Etiqueta	Definición: porcentaje de empleados en...
x_1	Agricultura	Agricultura
x_2	Minería	Minería
x_3	Industria	Industria
x_4	Energía	Industrias de generación de energía
x_5	Construcción	Construcción
x_6	ServiciosInd	Servicios a la industria
x_7	Finanzas	Sector financiero
x_8	ServiciosPer	Servicios a la sociedad y a las personas
x_9	Transporte	Transporte y las comunicaciones

Fuente: Hand *et al.* (1994, p. 303)

Estructura sectorial del empleo en Europa

Distribución del empleo por sectores en una serie de países europeos. El objetivo es determinar si existen perfiles de empleo que sirvan además para categorizar a los países en función de los mismos.

Figura 11.9.: Gráfico de las cargas sobre las dos primeras componentes



Agricultura (% de personas empleadas en Agricultura) y Minería está correlacionadas negativamente con Dim 1 (a la izquierda sobre el eje horizontal)

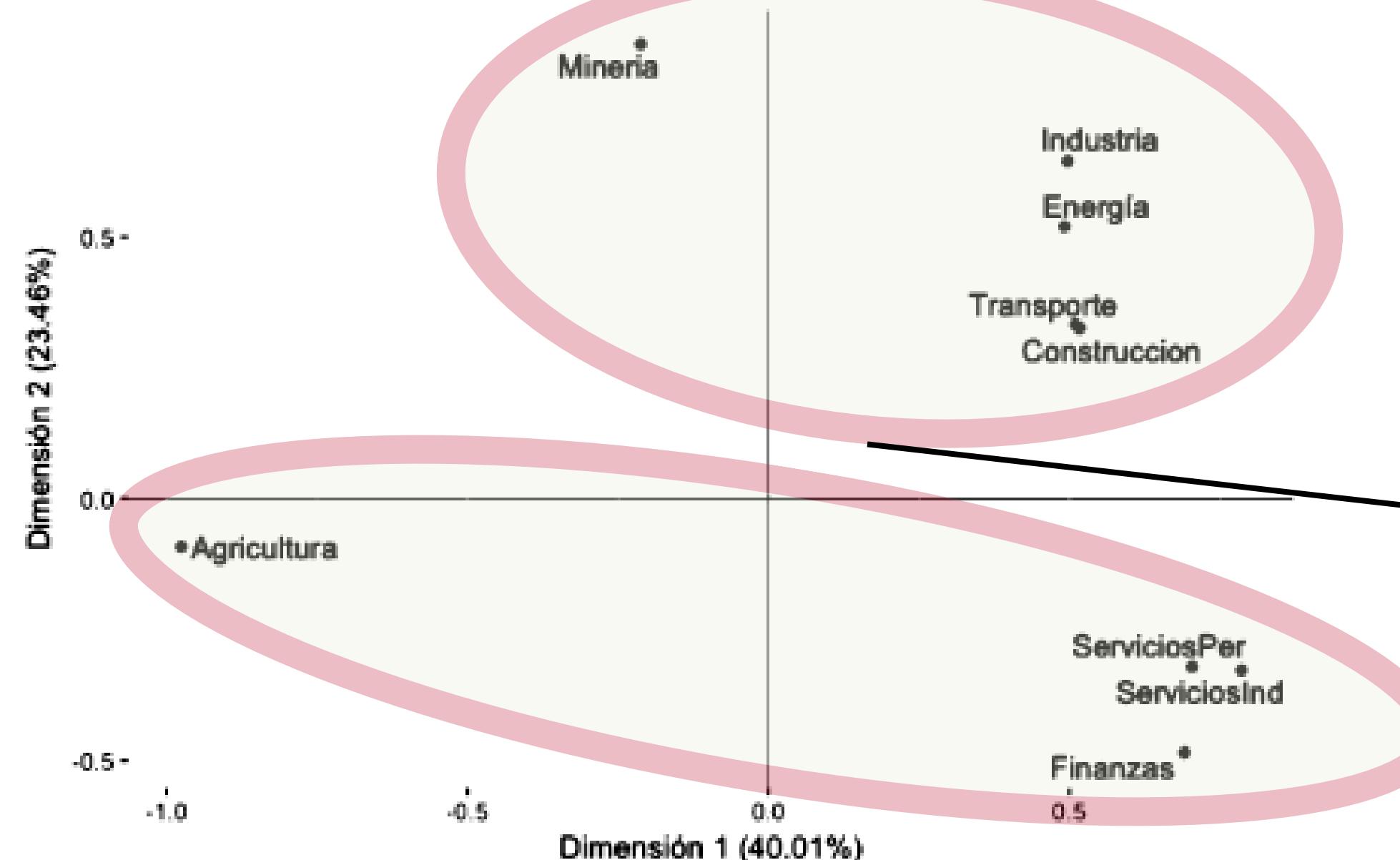
Las otras variables están correlacionadas positivamente con Dim 1 (a la derecha sobre el eje horizontal)

Interpretación: la componente principal 1 (Dim 1) distingue entre **países con economías más industriales y de servicios y países con sistemas económicos basados en el sector primario** (agr y minería)

Estructura sectorial del empleo en Europa

Distribución del empleo por sectores en una serie de países europeos. El objetivo es determinar si existen perfiles de empleo que sirvan además para categorizar a los países en función de los mismos.

Figura 11.9.: Gráfico de las cargas sobre las dos primeras componentes



Minería, Ind, Energía, Transporte y Constr están correlacionadas positivamente con Dim 2 (arriba sobre el eje vertical)

Las otras variables están correlacionadas negativamente con Dim 2 (abajo sobre el eje vertical)

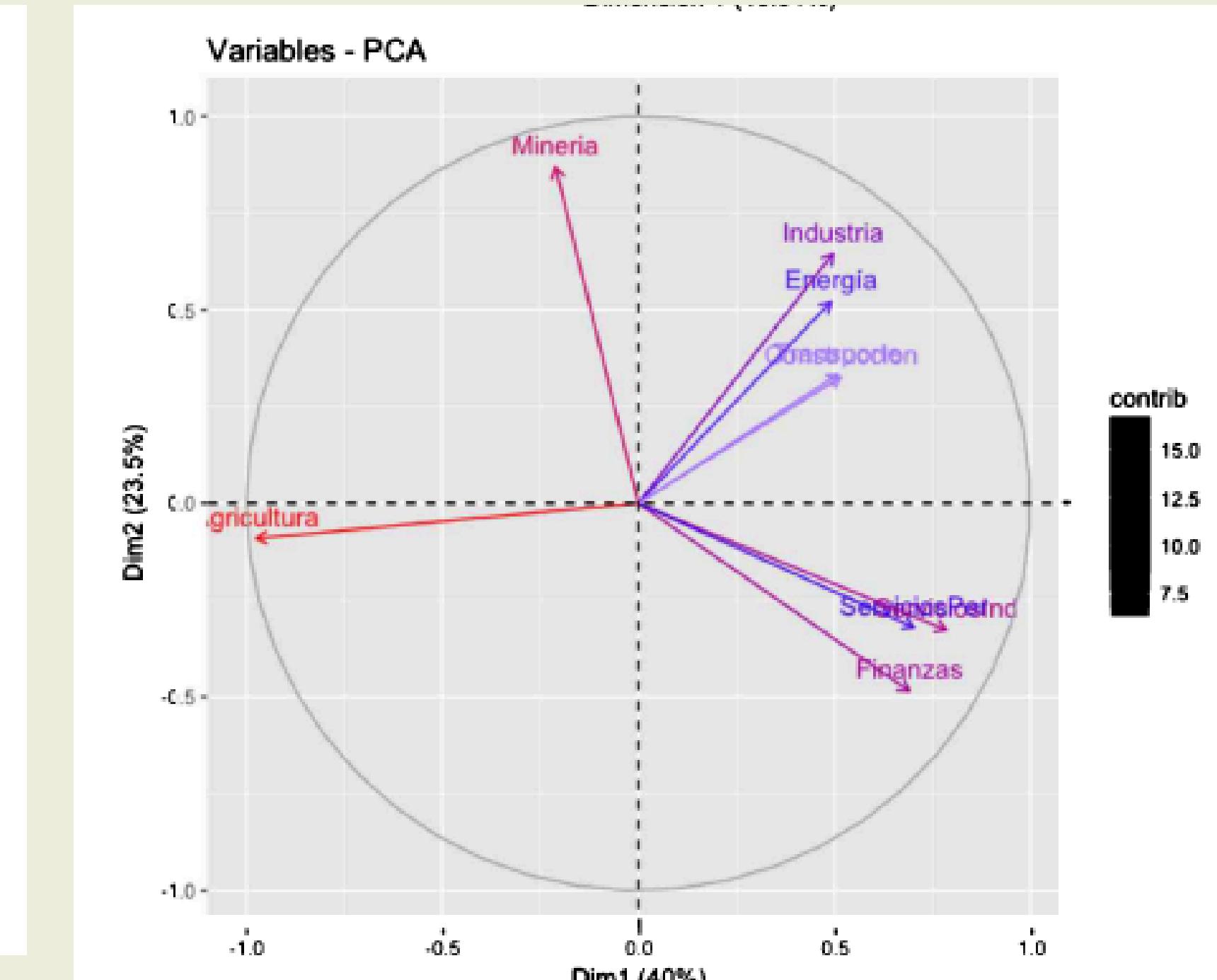
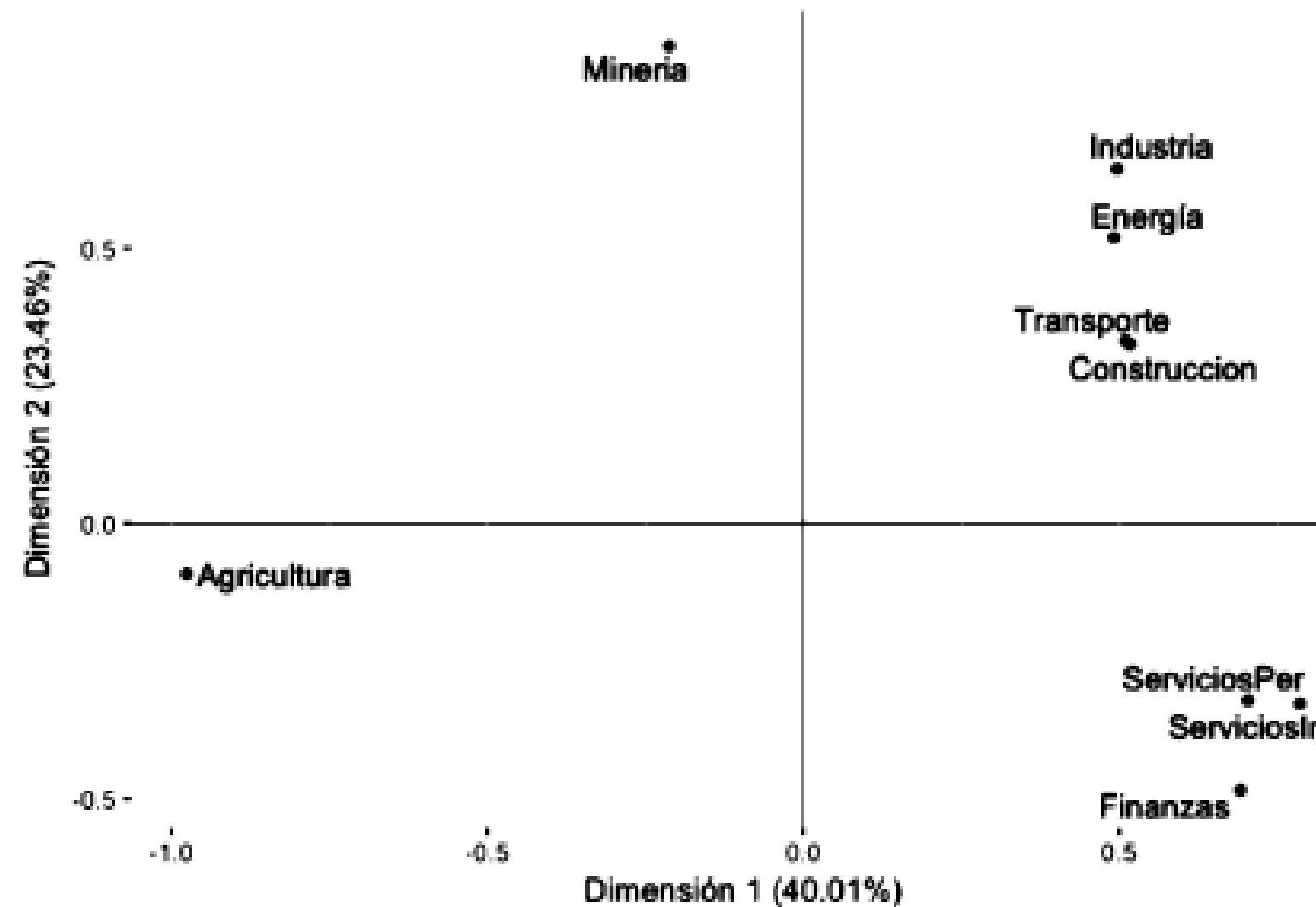
Interpretación: la componente principal 2 (Dim 2) distingue entre **países con economías con menor desarrollo en sus servicios** de aquellos **con mayor desarrollo en sus servicios**.

Estructura sectorial del empleo en Europa

Distribución del empleo por sectores en una serie de países europeos. El objetivo es determinar si existen perfiles de empleo que sirvan además para categorizar a los países en función de los mismos.

Ejemplo

Figura 11.9.: Gráfico de las cargas sobre las dos primeras componentes



Estructura sectorial del empleo en Europa

Distribución del empleo por sectores en una serie de países europeos. El objetivo es determinar si existen perfiles de empleo que sirvan además para categorizar a los países en función de los mismos.

Figura 11.10.: Representación gráfica de los países sobre las primeras dos componentes

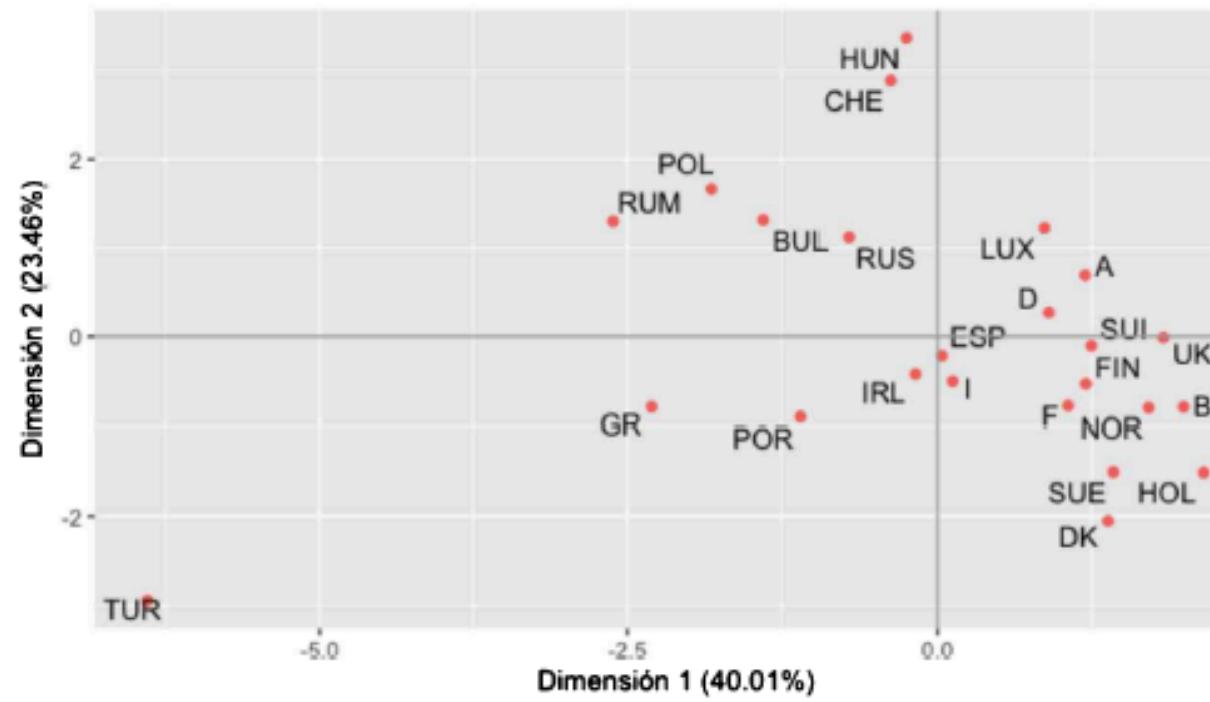
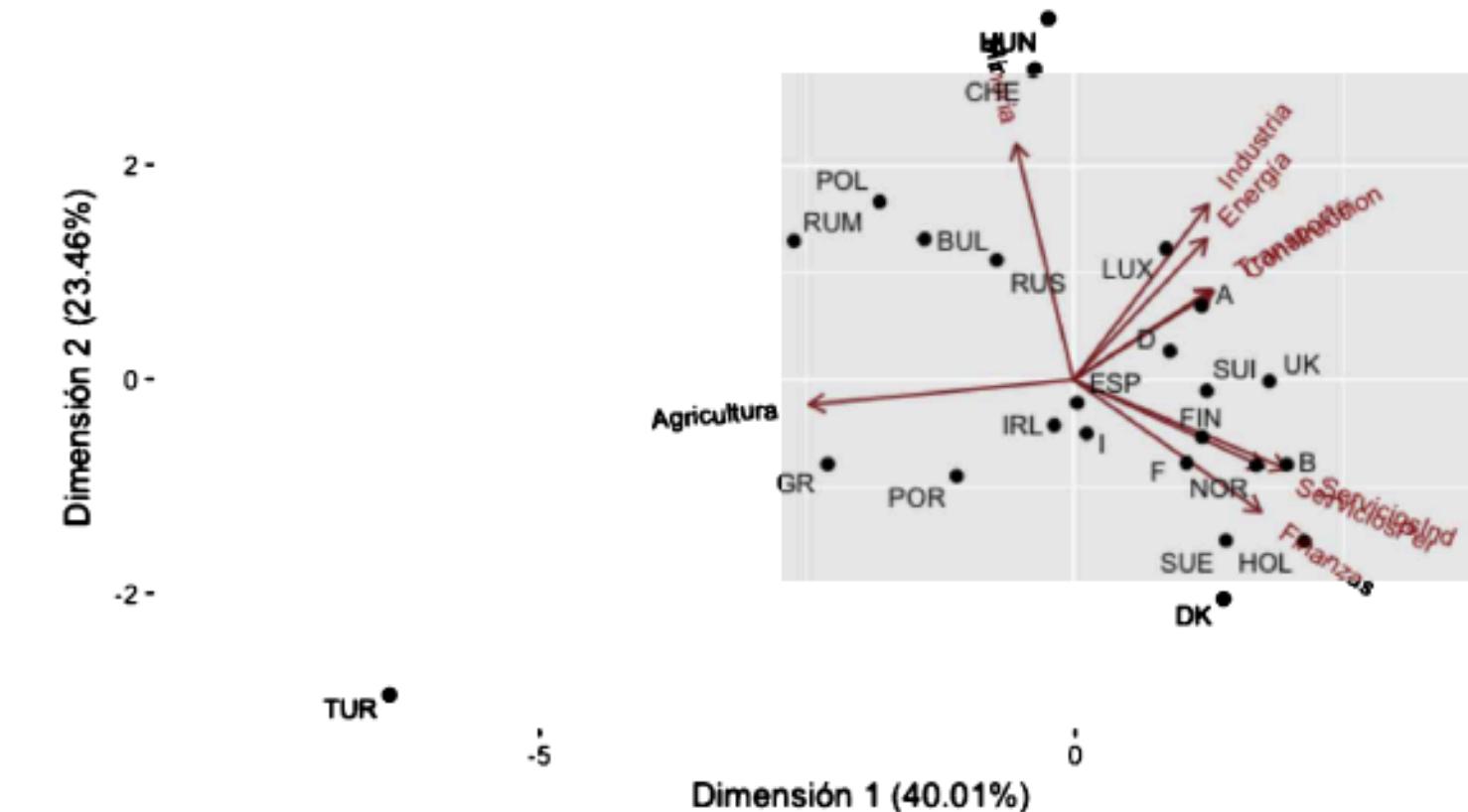


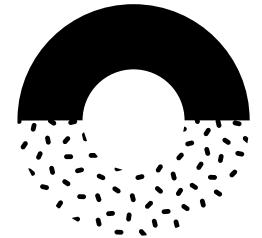
Figura 11.11.: Representación gráfica conjunta de los países y de los sectores sobre las primeras dos componentes



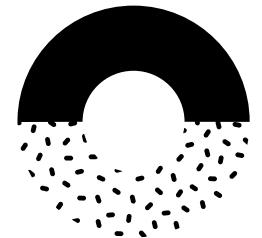
La ordenación de países sobre la componente 1 (Dim 1) va desde los más agrícolas (Turquía y Grecia) hasta los que tienen economías más industrializadas y de servicios. La Dime 2 ordena a estos últimos entre aquellos donde el peso de la industria es más fuerte (Hungría, Chequia, Rusia) y donde están los que tienen mayor desarrollo en servicios (Suecia, Holanda, Dinamarca).

¿Cómo llegamos hasta ahí?

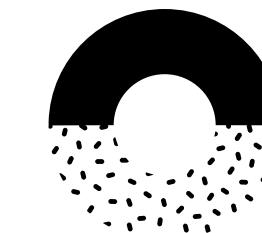
¿Cómo llegamos hasta ahí?



Contar con variables correlacionadas



Preparar las variables (estandarizarlas, tipificarlas)



Calcular las matrices de correlación, los autovalores y autovectores), los coeficientes de los componentes principales y los mapas

También se pueden construir índices a partir de una combinación lineal de variables

Análisis de componentes principales

01

Clase 1 (hoy)

- lógica geométrica
- interpretación del resultado
- gráficos
- códigos R Studio

02

Clase 2 (8 sept)

- interpretación del resultado
- determinación del número de componentes principales
- índices
- códigos R Studio

Análisis de componentes principales

01

Clase 1 (hoy)

- lógica geométrica
- interpretación del resultado
- gráficos
- códigos R Studio

Caso simulado de dos variables

- Lógica geométrica
- Cálculo matemático (y R) para dos variables

Cálculo para más variables

Análisis de componentes principales

01

Clase 1 (hoy)

- lógica geométrica
- interpretación del resultado
- gráficos
- códigos R Studio

Caso simulado de dos variables

- Lógica geométrica
- Cálculo matemático (y R) para dos variables

Cálculo para más variables

Análisis de componentes principales

Lógica geométrica

Variables

Supongamos que tenemos dos variables para 12 casos (datos originales y centrados, se resta el valor de la media).

Cuadro 11.1.: Datos originales y en desviaciones respecto a la media

Caso	x_1		x_2	
	Original	Desviación respecto media	Original	Desviación respecto media
1	16	8	8	5
2	12	4	10	7
3	13	5	6	3
4	11	3	2	-1
5	10	2	8	5
6	9	1	-1	-4
7	8	0	4	1
8	7	-1	6	3
9	5	-3	-3	-6
10	3	-5	-1	-4
11	2	-6	-3	-6
12	0	-8	0	-3
Media	8	0	3	0
Varianza	23,091	23,091	21,091	21,091

Análisis de componentes principales

Lógica geométrica

Variables

Cuadro 11.1.: Datos originales y en desviaciones respecto a la media

Caso	x_1		x_2	
	Original	Desviación respecto media	Original	Desviación respecto media
1	16	8	8	5
2	12	4	10	7
3	13	5	6	3
4	11	3	2	-1
5	10	2	8	5
6	9	1	-1	-4
7	8	0	4	1
8	7	-1	6	3
9	5	-3	-3	-6
10	3	-5	-1	-4
11	2	-6	-3	-6
12	0	-8	0	-3
Media	8	0	3	0
Varianza	23,091	23,091	21,091	21,091

$$\text{SSCP} = \begin{bmatrix} 254 & 181 \\ 181 & 232 \end{bmatrix} \quad \mathbf{S} = \begin{bmatrix} 23,091 & 16,455 \\ 16,455 & 21,091 \end{bmatrix}$$

$$\mathbf{R} = \begin{bmatrix} 1,000 & 0,746 \\ 0,746 & 1,000 \end{bmatrix}$$

Fuente: Sharma (1996, p. 59)

Varianza: desviación con respecto a la media

varianza $x_1 = 23,091$; varianza $x_2 = 21,091$; varianza total = 44,182

A la variable x_1 le corresponde el 52,26% de la varianza total y a la x_2 el 47,74%

- 52,26% = $23,091 / 44,182 \times 100$
- 47,74% = $21,091 / 44,182 \times 100$

\mathbf{R} = matriz de correlaciones:

Ambas variables están correlacionadas $p = 0,746$

Análisis de componentes principales

Lógica geométrica

Crear variables nueva (combinación lineal de x_1 y x_2) que maximice la varianza

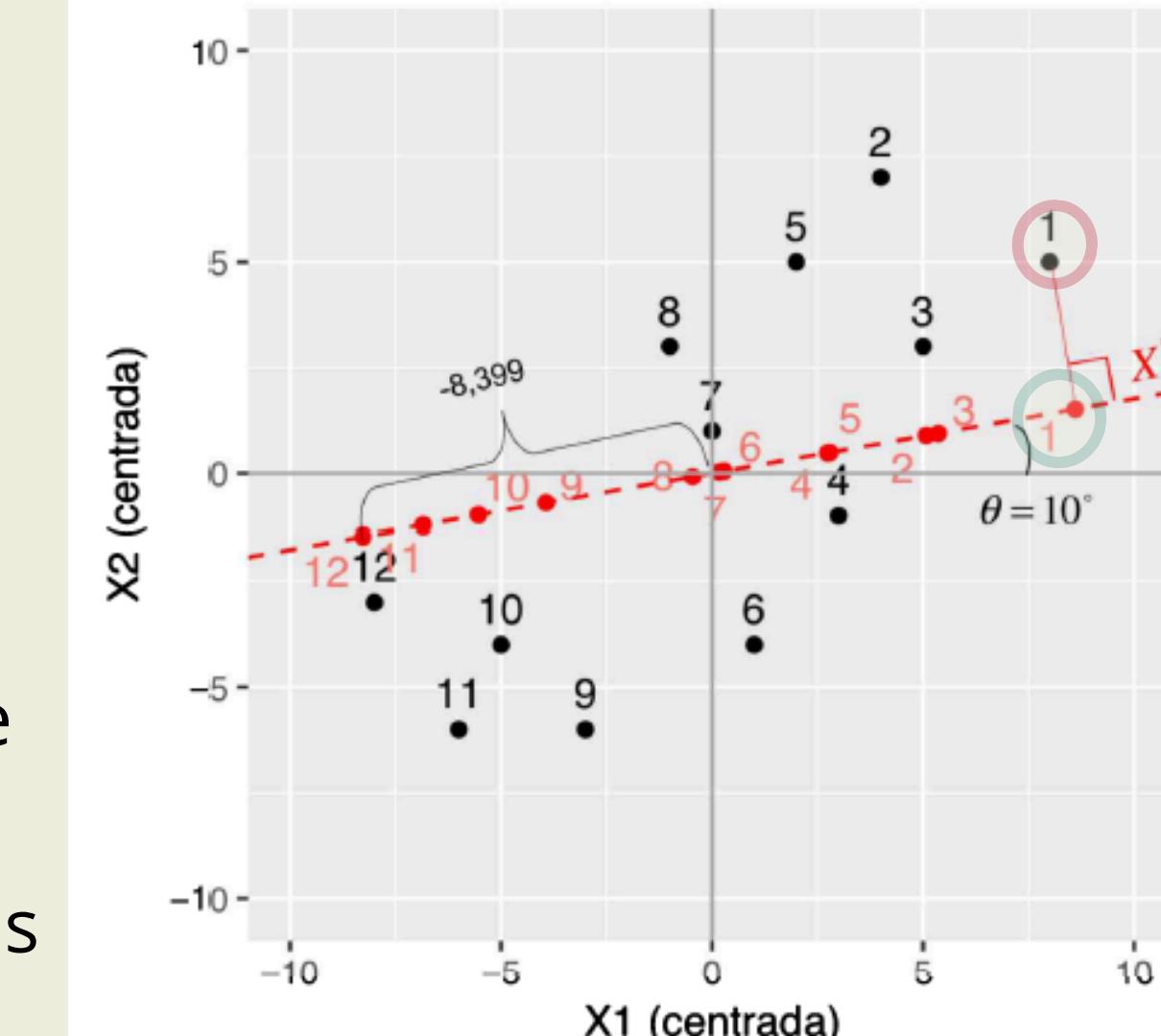
Cuadro 11.1.: Datos originales y en desviaciones respecto a la media

Caso	x_1		x_2	
	Original	Desviación respecto media	Original	Desviación respecto media
1	16	8	8	5
2	12	4	10	7
3	13	5	6	3
4	11	3	2	-1
5	10	2	8	5
6	9	1	-1	-4
7	8	0	4	1
8	7	-1	6	3
9	5	-3	-3	-6
10	3	-5	-1	-4
11	2	-6	-3	-6
12	0	-8	0	-3
Media	8	0	3	0
Varianza	23,091	23,091	21,091	21,091



La **línea roja** es la nueva variable (componente principal), donde se “proyectan ortogonalmente” los casos (x_1, x_2)

Figura 11.1.: Datos originales centrados y su proyección sobre X_1^*



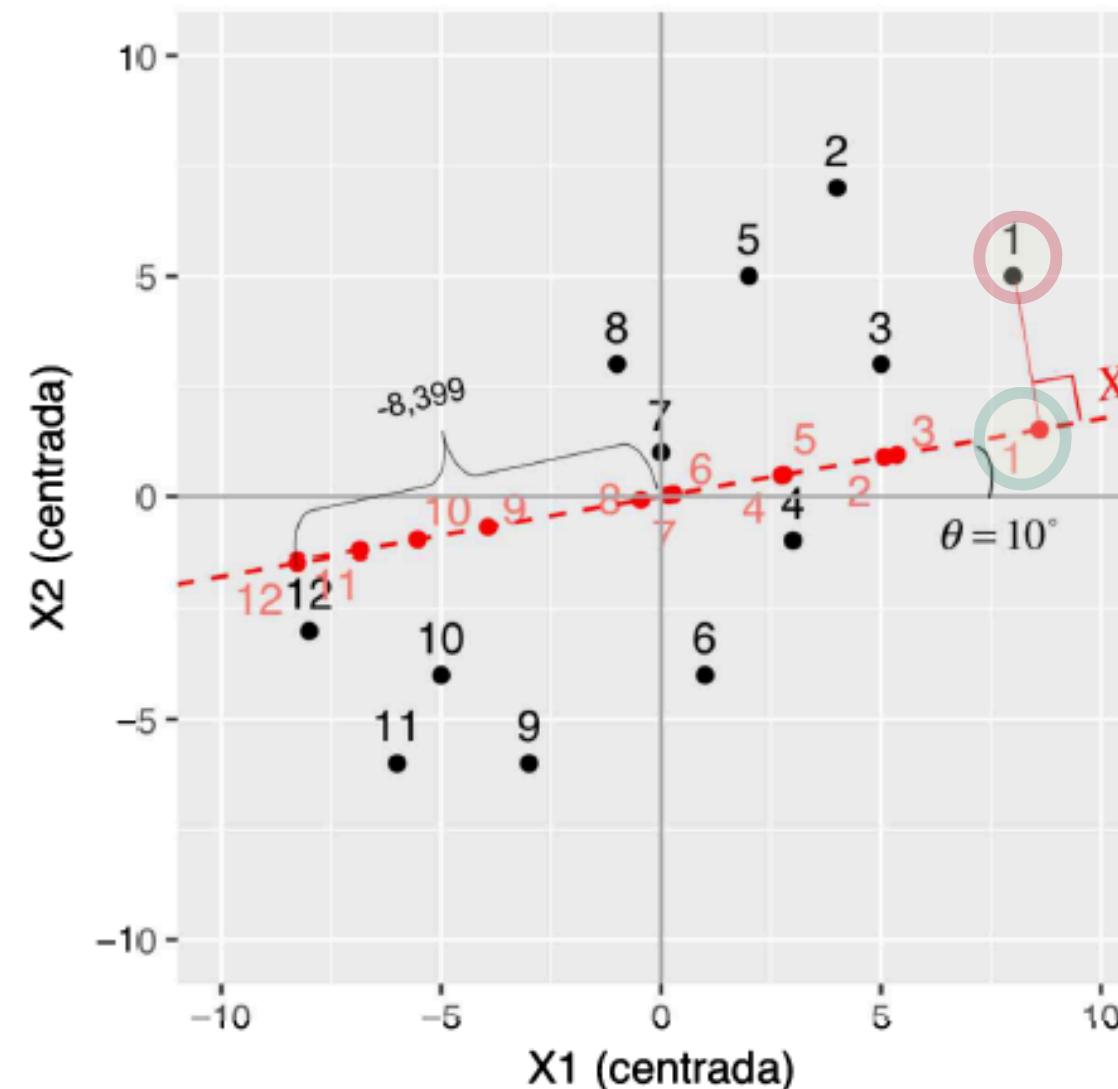
Fuente: Sharma (1996, p. 60).

Análisis de componentes principales

Lógica geométrica

¿Cómo se obtiene esa línea?

Figura 11.1.: Datos originales centrados y su proyección sobre X_1^*



Fuente: Sharma (1996, p. 60).

Se proyectan los valores de las dos variables x_1 y x_2 sobre una recta (variable) que denominaremos X^* que formará, como muestra la figura 11.1, un ángulo θ con el eje original X_1 . Por trigonometría, las proyecciones (coordenadas) de cualquier punto sobre ese nuevo eje X^* -representadas con un punto sobre la línea- vendrán dadas por la expresión:

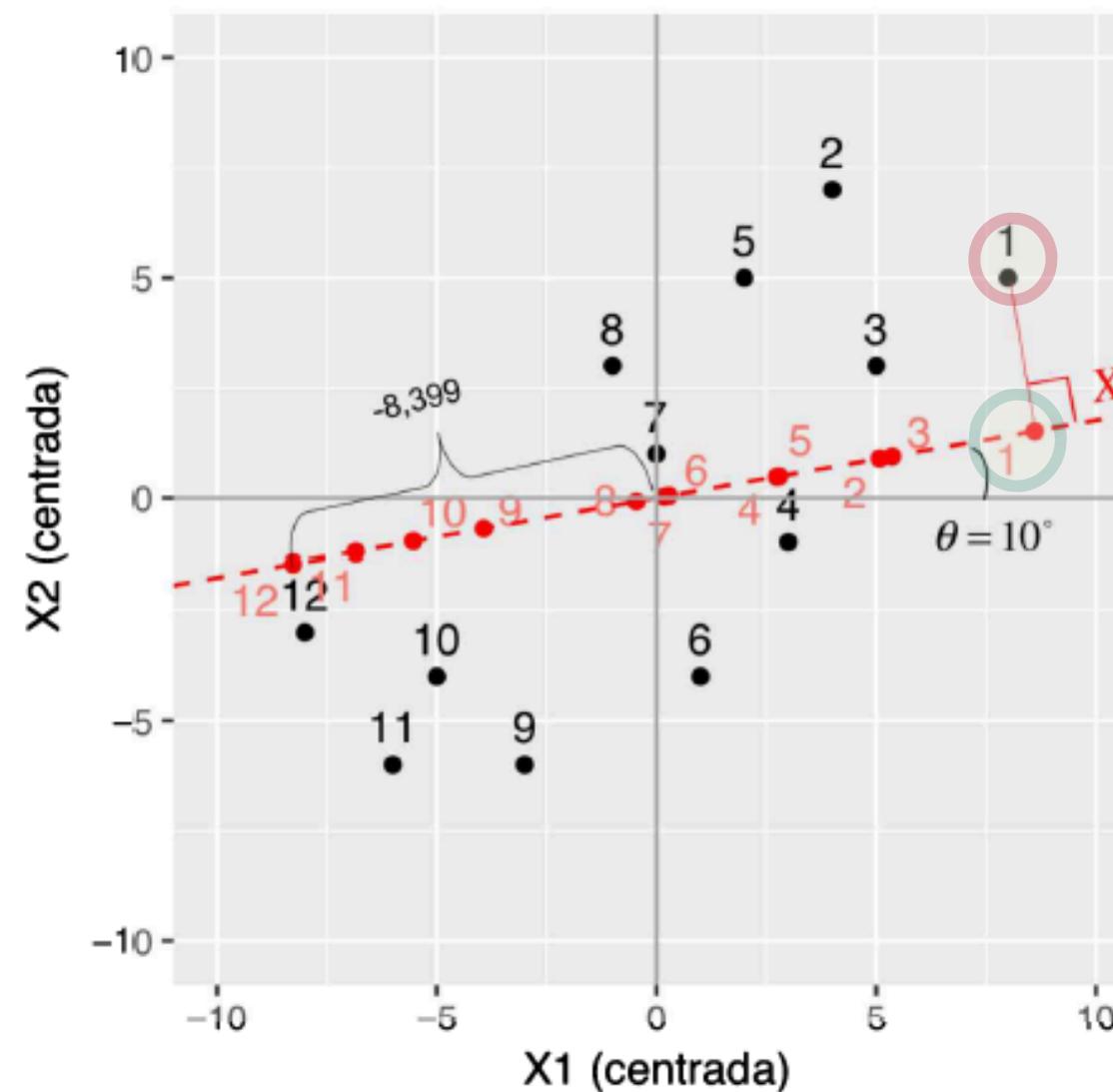
$$x_1^* = \cos \theta \times x_1 + \sin \theta \times x_2 = 0,985x_1 + 0,174x_2 \quad (\theta = 10^\circ)$$

Análisis de componentes principales

Lógica geométrica

¿Cómo se obtiene esa línea?

Figura 11.1.: Datos originales centrados y su proyección sobre X_1^*



$$x_1^* = \cos \theta \times x_1 + \sin \theta \times x_2 = 0,985x_1 + 0,174x_2 \quad (\theta = 10^\circ)$$

Para el caso $1 = (8,5) = 0,985 \times 8 + 0,174 \times 5 = 8,75$

Es decir, redujimos un valor que se expresaba con dos números [8,5] con un solo número. Bajamos de 2 a 1 variable. Esta nueva variable no contiene toda la varianza explicada pero una gran parte de ella

Análisis de componentes principales

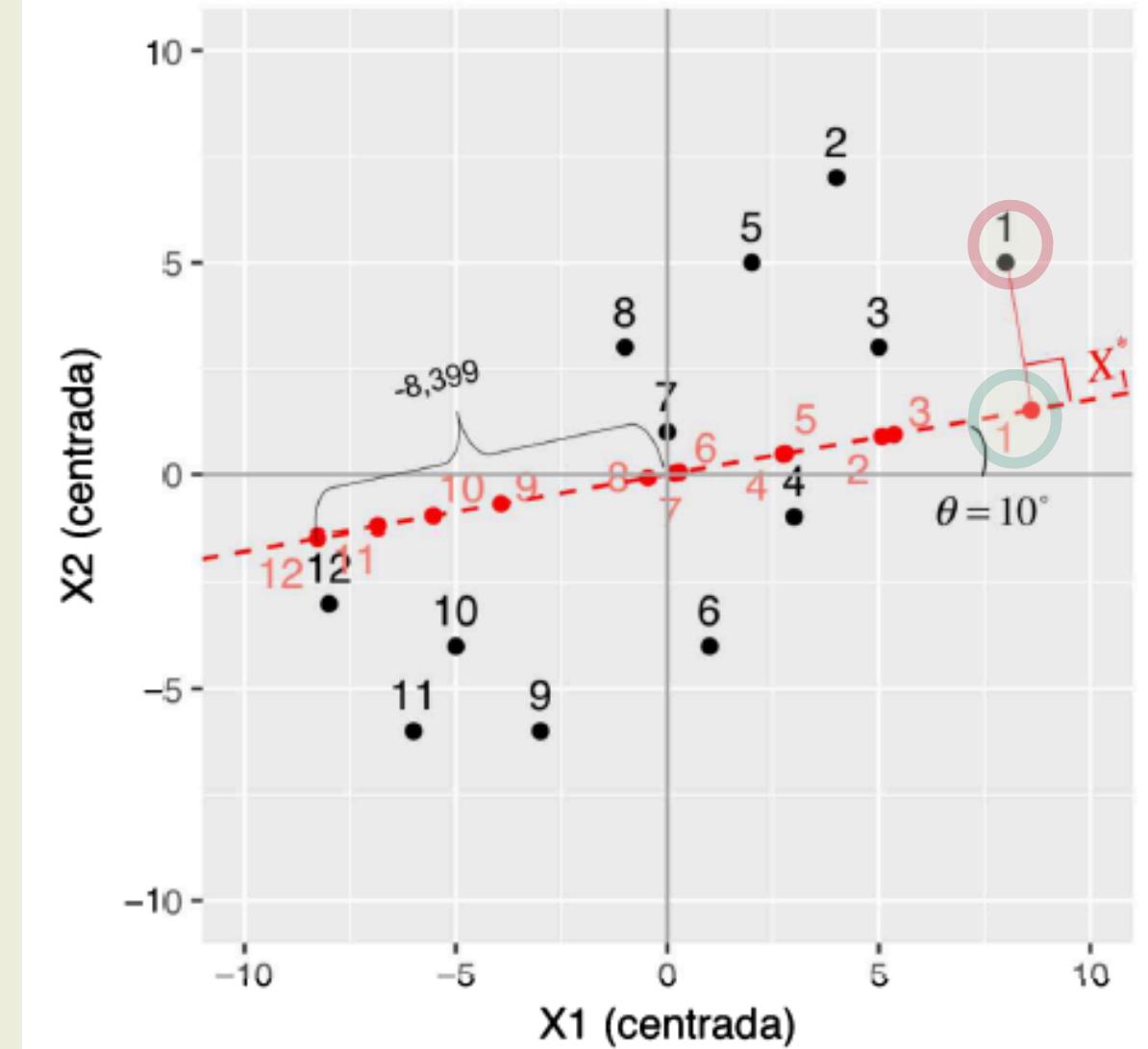
Lógica geométrica

Cuadro 11.2.: Datos originales centrados y nueva variable x_1^* para una rotación de $\theta = 10^\circ$

Caso	Datos centrados		x_1^*
	x_1	x_2	
1	8	5	8,747
2	4	7	5,155
3	5	3	5,445
4	3	-1	2,781
5	2	5	2,838
6	1	-4	0,290
7	0	1	0,173
8	-1	3	-0,464
9	-3	-6	-3,996
10	-5	-4	-5,619
11	-6	-6	-6,951
12	-8	-3	-8,399
Media	0	0	0
Varianza	23,091	21,091	28,659

Fuente: Sharma (1996, p. 61).

Figura 11.1.: Datos originales centrados y su proyección sobre X_1^*



Fuente: Sharma (1996, p. 60).

caso 1 = (8,5) = 0,985x8 + 0,174x5 = 8,75

Análisis de componentes principales

Lógica geométrica

Cuadro 11.2.: Datos originales centrados y nueva variable x_1^* para una rotación de $\theta = 10^\circ$

Caso	Datos centrados		x_1^*
	x_1	x_2	
1	8	5	8,747
2	4	7	5,155
3	5	3	5,445
4	3	-1	2,781
5	2	5	2,838
6	1	-4	0,290
7	0	1	0,173
8	-1	3	-0,464
9	-3	-6	-3,996
10	-5	-4	-5,619
11	-6	-6	-6,951
12	-8	-3	-8,399
Media	0	0	0
Varianza	23,091	21,091	28,659

Fuente: Sharma (1996, p. 61).

La varianza (28,639) de la nueva variable es mayor a la de ambas, y corresponde al 68% de la varianza total

$$= 28,639 / (23,091 + 21,091) \times 100$$

Análisis de componentes principales

Lógica geométrica

Optimizamos la recta para maximizar la varianza de x^*

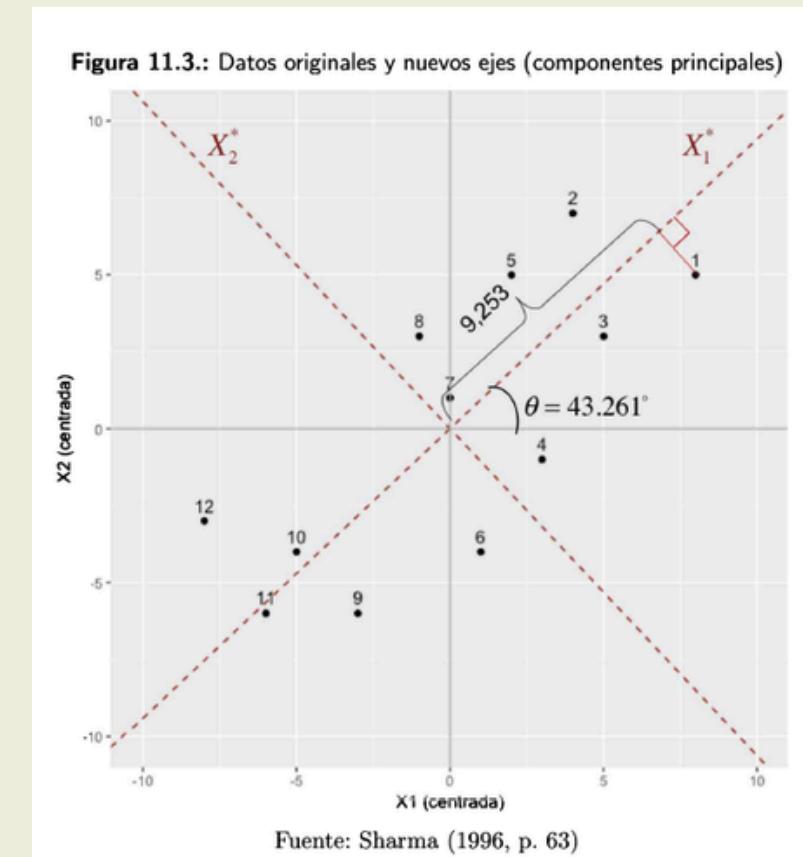
Esta es la nueva variable Componente Principal 1 para $\theta = 43^\circ$

Caso	Datos centrados		x_1^*	x_2^*
	x_1	x_2		
1	8	5	9,253	-1,841
2	4	7	7,710	2,356
3	5	3	5,697	-1,242
4	3	-1	1,499	-2,784
5	2	5	4,883	2,271
6	1	-4	-2,013	-3,598
7	0	1	0,685	0,728
8	-1	3	1,328	2,870
9	-3	-6	-6,297	-2,313
10	-5	-4	-6,382	0,514
11	-6	-6	-8,481	-0,257
12	-8	-3	-7,882	3,298
Media	0	0	0	0
Varianza	23,091	21,091	38,576	5,606

$$\text{SSCP} = \begin{bmatrix} 424,334 & 0,000 \\ 0,000 & 61,666 \end{bmatrix} \quad S = \begin{bmatrix} 38,576 & 0,000 \\ 0,000 & 5,606 \end{bmatrix}$$

$$R = \begin{bmatrix} 1,000 & 0,000 \\ 0,000 & 1,000 \end{bmatrix}$$

Fuente: Sharma (1996, p. 62).



Obtener un componente principal consiste en ir variando el ángulo θ hasta que la nueva variable x^* explique el máximo valor posible de la varianza total de las dos variables originales (44,182).

La combinación lineal para ese ángulo que maximiza la varianza explicada (minimiza la pérdida de información aunque hay una variable menos que interpretar) es la componente principal.

En este caso el valor de $\theta = 43^\circ$

Análisis de componentes principales

Lógica geométrica

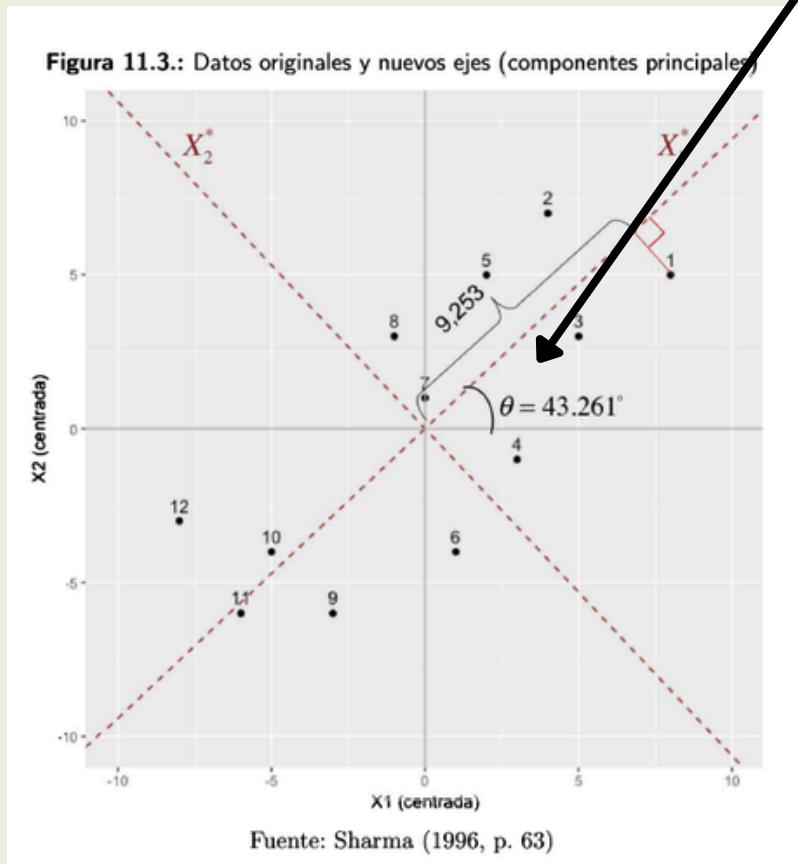
Componente Principal 1 y 2 para ángulo = 43°

Caso	Datos centrados		x_1^*	x_2^*
	x_1	x_2		
1	8	5	9,253	-1,841
2	4	7	7,710	2,356
3	5	3	5,697	-1,242
4	3	-1	1,499	-2,784
5	2	5	4,883	2,271
6	1	-4	-2,013	-3,598
7	0	1	0,685	0,728
8	-1	3	1,328	2,870
9	-3	-6	-6,297	-2,313
10	-5	-4	-6,382	0,514
11	-6	-6	-8,481	-0,257
12	-8	-3	-7,882	3,298
Media	0	0	0	0
Varianza	23,091	21,091	38,576	5,606

$$\text{SSCP} = \begin{bmatrix} 424,334 & 0,000 \\ 0,000 & 61,666 \end{bmatrix} \quad \mathbf{S} = \begin{bmatrix} 38,576 & 0,000 \\ 0,000 & 5,606 \end{bmatrix}$$

$$\mathbf{R} = \begin{bmatrix} 1,000 & 0,000 \\ 0,000 & 1,000 \end{bmatrix}$$

Fuente: Sharma (1996, p. 62).



1. A los nuevos ejes (color rojo, coordenadas con * en la tabla) se les denomina componentes principales y a las proyecciones de las variables sobre ellos, puntuaciones sobre las componentes principales.

2. Las nuevas variables son una combinación lineal de las originales y también están centradas en su media (esta es cero).

3. Las varianzas de x_1^* y x_2^* son 38,756 y 5,606, y la suma de ambas, 44,182. Esta suma es la misma que la de las variables originales, no cambia, lo que es lógico porque la orientación de los puntos en el espacio no ha cambiado.

4. Los porcentajes de la varianza recogidos por x_1^* y x_2^* son, respectivamente, el 87,31% = (38,576/44,182) y el 12,69% = (5,606/44,182) y este es el principal cambio.

La varianza recogida por la primera variable nueva es más grande que la varianza recogida por cualquiera de las variables originales.

5. La correlación entre las nuevas variables es cero, y x_1^* y x_2^* son ortogonales (están incorrelacionadas.)

Análisis de componentes principales

01

Clase 1 (hoy)

- lógica geométrica
- interpretación del resultado
- gráficos
- códigos R Studio

Caso simulado de dos variables

- Lógica geométrica
- Cálculo matemático (y R) para dos variables

Cálculo para más variables

Análisis de componentes principales

Cálculo matemático

Matrices, autovectores y autovalores (“propios”)

Uno de los conceptos claves para el PCA es el de eigenvalue y eigenvectors (o **autovalores y autovectores**)

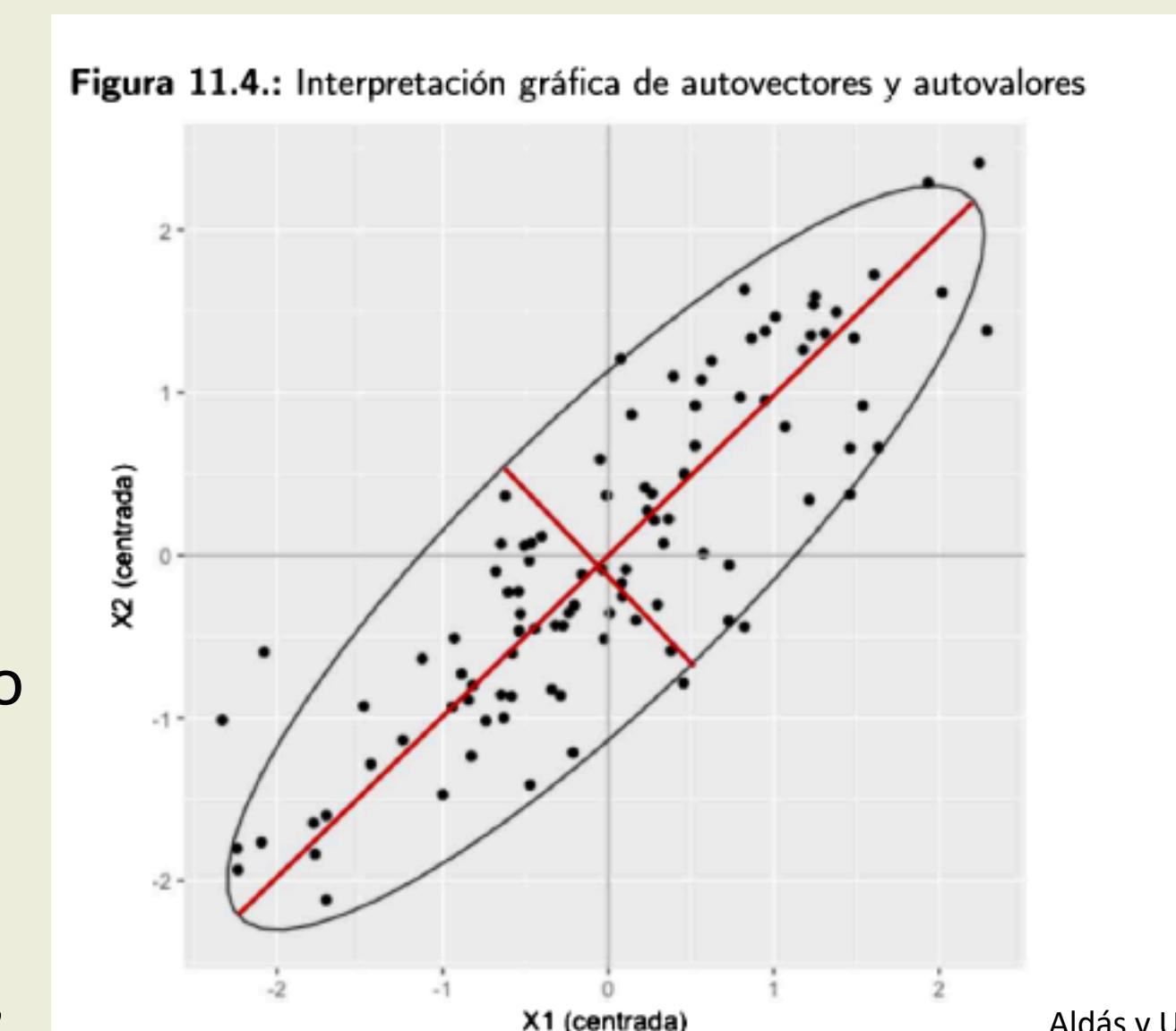
Cuando dos variables están correlacionadas, su nube de puntos de dispersión asemeja una elipse

Los **ejes ortogonales que describen la elipse** son los autovectores (en rojo en el gráfico)

Cada autovector tiene asociado un autovalor.

El **autovalor nos da una medida de la longitud** del autovector y observando ese valor podemos tener una idea clara de lo homogénea o heterogéneamente que los datos están distribuidos.

Si añadiéramos una tercera variable, la elipse se convertiría en algo parecido a un balón de rugby y tendríamos tres líneas perpendiculares, tres autovectores, y así sucesivamente.



Análisis de componentes principales

Cálculo matemático

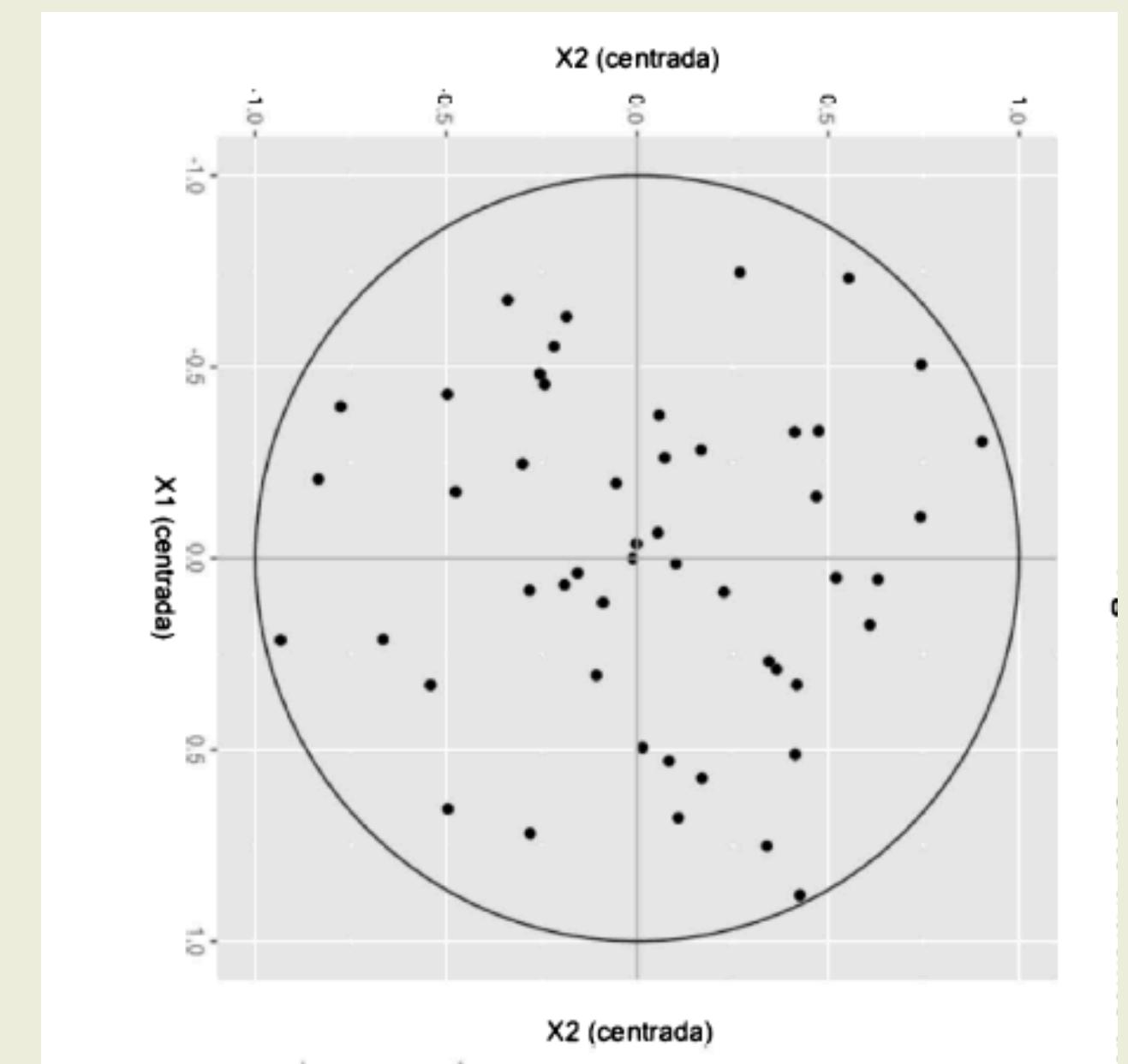
Matrices, autovectores y autovalores (“propios”)

figura 11.5

Cuando no hay relación (figura 11.5), el gráfico de dispersión mostrará más o menos un círculo.

La altura y la longitud -los autovectores- son los mismos, y esa relación daría 1.

En el caso de correlación perfecta, el gráfico de dispersión forma una línea recta, la altura de la elipse será muy pequeña (0 en el caso extremo) y la división entre los autovalores tenderá hacia infinito.



Análisis de componentes principales

Cálculo matemático

Matrices, autovectores y autovalores (“propios”)

Cuadro 11.1.: Datos originales y en desviaciones respecto a la media

Caso	x_1		x_2	
	Original	Desviación respecto media	Original	Desviación respecto media
1	16	8	8	5
2	12	4	10	7
3	13	5	6	3
4	11	3	2	-1
5	10	2	8	5
6	9	1	-1	-4
7	8	0	4	1
8	7	-1	6	3
9	5	-3	-3	-6
10	3	-5	-1	-4
11	2	-6	-3	-6
12	0	-8	0	-3
Media	8	0	3	0
Varianza	23,091	23,091	21,091	21,091

$$\text{SSCP} = \begin{bmatrix} 254 & 181 \\ 181 & 232 \end{bmatrix} \quad S = \begin{bmatrix} 23,091 & 16,455 \\ 16,455 & 21,091 \end{bmatrix}$$

$$R = \begin{bmatrix} 1,000 & 0,746 \\ 0,746 & 1,000 \end{bmatrix}$$

Fuente: Sharma (1996, p. 59)

El PCA requiere **calcula los autovalores y los autovectores de la matriz de correlación o de covarianzas**

En general se prefiere de correlación para que no afecten las unidades

S: La matriz de varianzas y covarianzas

R = matriz de correlaciones

Análisis de componentes principales

Cálculo matemático

Matrices, vectores y valores propios

Cuadro 11.1.: Datos originales y en desviaciones respecto a la media

Caso	x_1		x_2	
	Original	Desviación respecto media	Original	Desviación respecto media
1	16	8	8	5
2	12	4	10	7
3	13	5	6	3
4	11	3	2	-1
5	10	2	8	5
6	9	1	-1	-4
7	8	0	4	1
8	7	-1	6	3
9	5	-3	-3	-6
10	3	-5	-1	-4
11	2	-6	-3	-6
12	0	-8	0	-3
Media	8	0	3	0
Varianza	23,091	23,091	21,091	21,091

$$\text{SSCP} = \begin{bmatrix} 254 & 181 \\ 181 & 232 \end{bmatrix} \quad \mathbf{S} = \begin{bmatrix} 23,091 & 16,455 \\ 16,455 & 21,091 \end{bmatrix}$$

$$\mathbf{R} = \begin{bmatrix} 1,000 & 0,746 \\ 0,746 & 1,000 \end{bmatrix}$$

Fuente: Sharma (1996, p. 59)

Calculo en R de los autovalores y los autovectores de la matriz de covarianzas.

```
# Datos originales
x1<-c(16,12,13,11,10,9,8,7,5,3,2,0)
x2<-c(8,10,6,2,8,-1,4,6,-3,-1,-3,0)
# Datos centrados
x1_m<-x1-mean(x1) x2_m<-x2-mean(x2)
# Matriz datos originales
X<-matrix(c(x1,x2),ncol=2,nrow=12)
# Matriz datos centrados
Xm<-matrix(c(x1_m,x2_m),ncol=2,nrow=12)
# SSCP, S y R (datos centrados)
SSCP<-t(Xm) %*% (Xm)
S<-SSCP/(12-1)
R<-cov2cor(S)
```

`svd(X)`

Análisis de componentes principales

Cálculo matemático

Matrices, vectores y valores propios

Cuadro 11.5.: Autovalores y autovectores

```
> svd(S)  
$d  
[1] 38.575813 5.606005
```

```
$u  
[,1] [,2]  
[1,] -0.7282381 -0.6853242  
[2,] -0.6853242 0.7282381
```

```
$v  
[,1] [,2]  
[1,] -0.7282381 -0.6853242  
[2,] -0.6853242 0.7282381
```

Caso	Datos centrados		x_1^*	x_2^*
	x_1	x_2		
1	8	5	9,253	-1,841
2	4	7	7,710	2,356
3	5	3	5,697	-1,242
4	3	-1	1,499	-2,784
5	2	5	4,883	2,271
6	1	-4	-2,013	-3,598
7	0	1	0,685	0,728
8	-1	3	1,328	2,870
9	-3	-6	-6,297	-2,313
10	-5	-4	-6,382	0,514
11	-6	-6	-8,481	-0,257
12	-8	-3	-7,882	3,298
Media	0	0	0	0
Varianza	23,091	21,091	38,576	5,606

$$\text{SSCP} = \begin{bmatrix} 424,334 & 0,000 \\ 0,000 & 61,666 \end{bmatrix} \quad \mathbf{S} = \begin{bmatrix} 38,576 & 0,000 \\ 0,000 & 5,606 \end{bmatrix}$$

$$\mathbf{R} = \begin{bmatrix} 1,000 & 0,000 \\ 0,000 & 1,000 \end{bmatrix}$$

Fuente: Sharma (1996, p. 62).

d\$

Arroja los autovalores de cada dimensión (autovectores)

Dim 1 = 38, 576

Explica el 87,3% de la varianza total (= 38, 576 / (38, 576+5, 606))

Dim 2 = 5,6 (13% de la varianza total explicada)

Por lo tanto, si decidiéramos realizar el análisis con una sola variable en lugar de con dos (reducción de datos) perderíamos apenas un 13% de la información total.

El autovalor es la varianza de cada componente principal: cuanto mayor es su valor, más correlacionadas están las proyecciones en torno a su eje.

Análisis de componentes principales

Cálculo matemático

Matrices, vectores y valores propios

Cuadro 11.5.: Autovalores y autovectores

```
> svd(S)  
$d  
[1] 38.575813 5.606005
```

```
$u  
[,1] [,2]  
[1,] -0.7282381 -0.6853242  
[2,] -0.6853242 0.7282381
```

```
$v  
[,1] [,2]  
[1,] -0.7282381 -0.6853242  
[2,] -0.6853242 0.7282381
```

\$u son los autovectores

$$\mathbf{u}_1 = \begin{bmatrix} u_{11} \\ u_{12} \end{bmatrix} = \begin{bmatrix} -0,728 \\ -0,685 \end{bmatrix}$$

$$\mathbf{u}_2 = \begin{bmatrix} u_{21} \\ u_{22} \end{bmatrix} = \begin{bmatrix} -0,685 \\ 0,728 \end{bmatrix}$$

Los coeficientes de los vectores \mathbf{u}_1 y \mathbf{u}_2 son los coeficientes que hay que aplicar a las variables originales para obtener las componentes principales.

Las componentes principales se expresan de la siguiente forma:

$$x^*_1 = u_{11}X_1 + u_{12}X_2$$

$$x^*_2 = u_{21}X_1 + u_{22}X_2$$

En este ejemplo se trabaja con la matriz S de covarianzas, pero se recomienda trabajar con la matriz de correlaciones, que está tipificada (desviación standard = 1)

Análisis de componentes principales

Cálculo matemático

Matrices, vectores y valores propios

Cuadro 11.5.: Autovalores y autovectores

```
> svd(S)  
$d  
[1] 38.575813 5.606005
```

```
$u  
[,1] [,2]  
[1,] -0.7282381 -0.6853242  
[2,] -0.6853242 0.7282381
```

```
$v  
[,1] [,2]  
[1,] -0.7282381 -0.6853242  
[2,] -0.6853242 0.7282381
```

\$u son los autovectores

$$\mathbf{u}_1 = \begin{bmatrix} u_{11} \\ u_{12} \end{bmatrix} = \begin{bmatrix} -0,728 \\ -0,685 \end{bmatrix}$$

$$\mathbf{u}_2 = \begin{bmatrix} u_{21} \\ u_{22} \end{bmatrix} = \begin{bmatrix} -0,685 \\ 0,728 \end{bmatrix}$$

Variables componentes principales

$$x^*_1 = -0,728 x_1 - 0,685 x_2$$

$$x^*_2 = -0,685 x_1 + 0,728 x_2$$

Análisis de componentes principales

01

Clase 1 (hoy)

- lógica geométrica
- interpretación del resultado
- gráficos
- códigos R Studio

Caso simulado de dos variables

- Lógica geométrica
- Cálculo matemático (y R) para dos variables

Cálculo para más variables

Estructura sectorial del empleo en Europa

Distribución del empleo por sectores en una serie de países europeos. El objetivo es determinar si existen perfiles de empleo que sirvan además para categorizar a los países en función de los mismos.

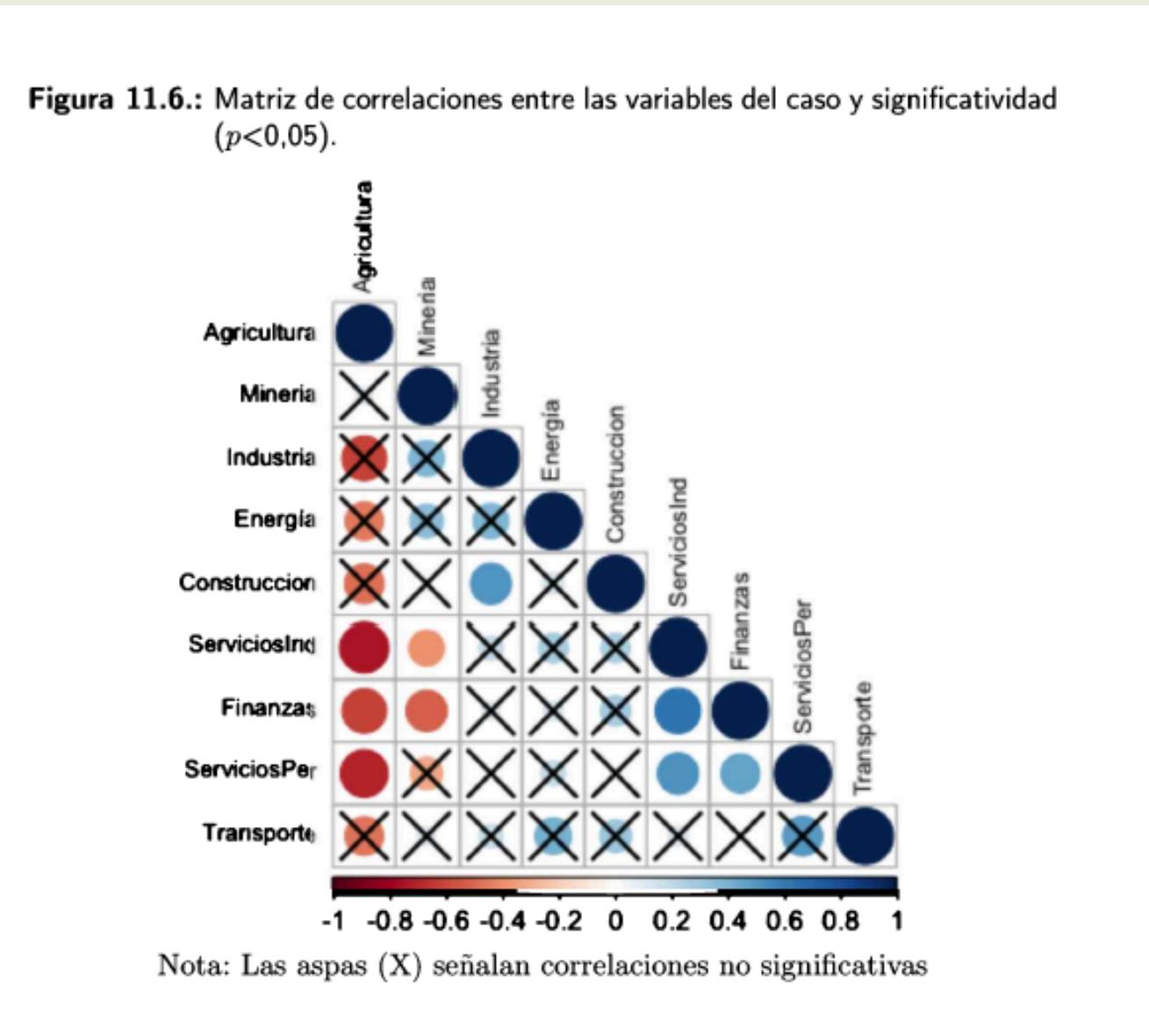
Cuadro 11.8.: Variables de la base de datos de empleo

Variable	Etiqueta	Definición: porcentaje de empleados en...
x_1	Agricultura	Agricultura
x_2	Minería	Minería
x_3	Industria	Industria
x_4	Energía	Industrias de generación de energía
x_5	Construcción	Construcción
x_6	ServiciosInd	Servicios a la industria
x_7	Finanzas	Sector financiero
x_8	ServiciosPer	Servicios a la sociedad y a las personas
x_9	Transporte	Transporte y las comunicaciones

Fuente: Hand *et al.* (1994, p. 303)

Estructura sectorial del empleo en Europa

Distribución del empleo por sectores en una serie de países europeos. El objetivo es determinar si existen perfiles de empleo que sirvan además para categorizar a los países en función de los mismos.



Para graficar la matriz R de correlaciones

```
library(corrplot)
```

```
corrplot(R, type = "lower")
```

R = matriz de correlaciones

Estructura sectorial del empleo en Europa

Distribución del empleo por sectores en una serie de países europeos. El objetivo es determinar si existen perfiles de empleo que sirvan además para categorizar a los países en función de los mismos.

Cuadro 11.11.: Aplicación del criterio del autovalor >1

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	3.6009354526	4.001039e+01	40.01039
comp 2	2.1113460064	2.345940e+01	63.46979
comp 3	1.2103485274	1.344832e+01	76.91811
comp 4	0.8351719711	9.279689e+00	86.19780
comp 5	0.5207887756	5.786542e+00	91.98434
comp 6	0.3368747540	3.743053e+00	95.72739
comp 7	0.2552492568	2.836103e+00	98.56350
comp 8	0.1292414244	1.436016e+00	99.99951
comp 9	0.0000438317	4.870189e-04	100.00000

Para obtener los autovalores y el % de variance (incluye la tipificación de las variables) y el gráfico con los componentes principales

```
library(FactoMineR)

fit <- PCA(datos,
            scale.unit = TRUE, # estandariza variables
            ncp = ncol(datos_sel), # n° máx. de componentes
head (fit)
```

Estructura sectorial del empleo en Europa

Distribución del empleo por sectores en una serie de países europeos. El objetivo es determinar si existen perfiles de empleo que sirvan además para categorizar a los países en función de los mismos.

Supongamos que elegimos dos componentes principales (próxima clase veremos cómo se selecciona el número)

Para interpretar las dos componentes extraídas es necesario fijarse en la contribución de cada variable.

Cuanto mayor es la carga, mayor es la influencia que ha tenido esa variable en la formación de la componente.

Por lo tanto podemos analizar cuáles son las cargas más altas y usar las variables a las que corresponden para dar una interpretación al eje.

Si nos fijamos en el cuadro 11.14 vemos que la primera componente está muy correlacionada de manera negativa con la agricultura y en menor medida con la minería.

Cuadro 11.14.: Correlaciones de las variables con las componentes (cargas)
\$var\$cor

		Dim.1	Dim.2
	Agricultura	-0.9755452	-0.0909600
	Mineria	-0.2119704	0.8680503
	Industria	0.4967865	0.6451644
	Energía	0.4913828	0.5202499
	Construccion	0.5166372	0.3259305
	ServiciosInd	0.7857710	-0.3276420
	Finanzas	0.6915111	-0.4852026
	ServiciosPer	0.7029305	-0.3210492
	Transporte	0.5093258	0.3325506

```
proc_data_pca <- proc_data %>%
  select(barrio_ideal, integrado, identifico, parte_de_mi,
         amigos, sociable, cordialidad, colaboracion)

fit <- PCA(proc_data_pca,
            scale.unit = TRUE,
            ncp = ncol(proc_data_pca),
            graph = TRUE)

# Cargas de las variables sobre cada componente
fit$var$coord
```

Estructura sectorial del empleo en Europa

Distribución del empleo por sectores en una serie de países europeos. El objetivo es determinar si existen perfiles de empleo que sirvan además para categorizar a los países en función de los mismos.

Supongamos que elegimos dos componentes principales (próxima clase veremos cómo se selecciona el número)

La segunda componente correlaciona de manera positiva la minería, la industria, la energía, la construcción y el transporte y, de manera negativa, los servicios, ya sean a la industria, personales o financieros, por lo que nos llevaría a interpretarla como aquella que contrapone países con un mayor o menor desarrollo en su sector servicios.

Cuadro 11.14.: Correlaciones de las variables con las componentes (cargas)
\$var\$cor

		Dim.1	Dim.2
Agricultura	-0.9755452	-0.0909600	
Mineria	-0.2119704	0.8680503	
Industria	0.4967865	0.6451644	
Energía	0.4913828	0.5202499	
Construccion	0.5166372	0.3259305	
ServiciosInd	0.7857710	-0.3276420	
Finanzas	0.6915111	-0.4852026	
ServiciosPer	0.7029305	-0.3210492	
Transporte	0.5093258	0.3325506	

```
proc_data_pca <- proc_data %>%
  select(barrio_ideal, integrado, identifico, parte_de_mi,
         amigos, sociable, cordialidad, colaboracion)

fit <- PCA(proc_data_pca,
            scale.unit = TRUE,
            ncp = ncol(proc_data_pca),
            graph = TRUE)

# Cargas de las variables sobre cada componente
fit$var$coord
```

Estructura sectorial del empleo en Europa

Distribución del empleo por sectores en una serie de países europeos. El objetivo es determinar si existen perfiles de empleo que sirvan además para categorizar a los países en función de los mismos.

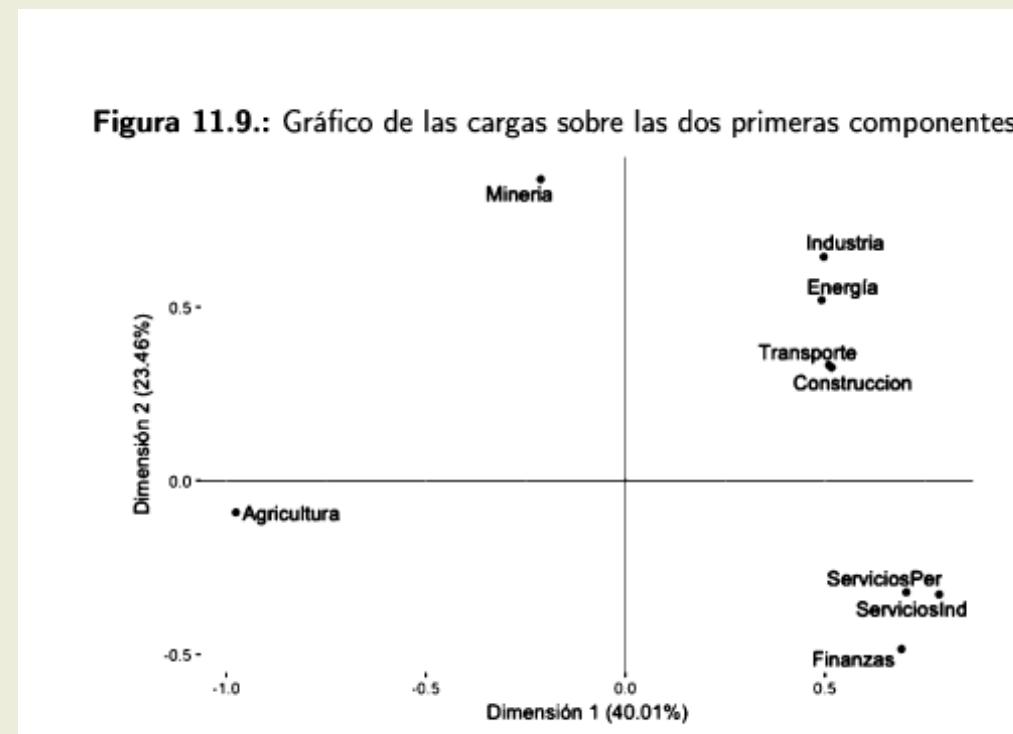
Puede facilitarse la interpretación representando las coordenadas estandarizadas de las variables sobre el mapa que marcan las componentes.

La función `PCA{FactoMineR}` lo hace por defecto

Puede generarse un mapa más visual mediante el módulo de gráficos `ggplot2`, que extrae los datos directamente del objeto `fit` en el que se ha guardado la estimación.

También la función `fviz_pca_var{factoextra}` permite combinar el sentido de la correlación (carga) con la intensidad de la contribución de cada variable.

Dos opciones para traficar



```
library(ggplot2)

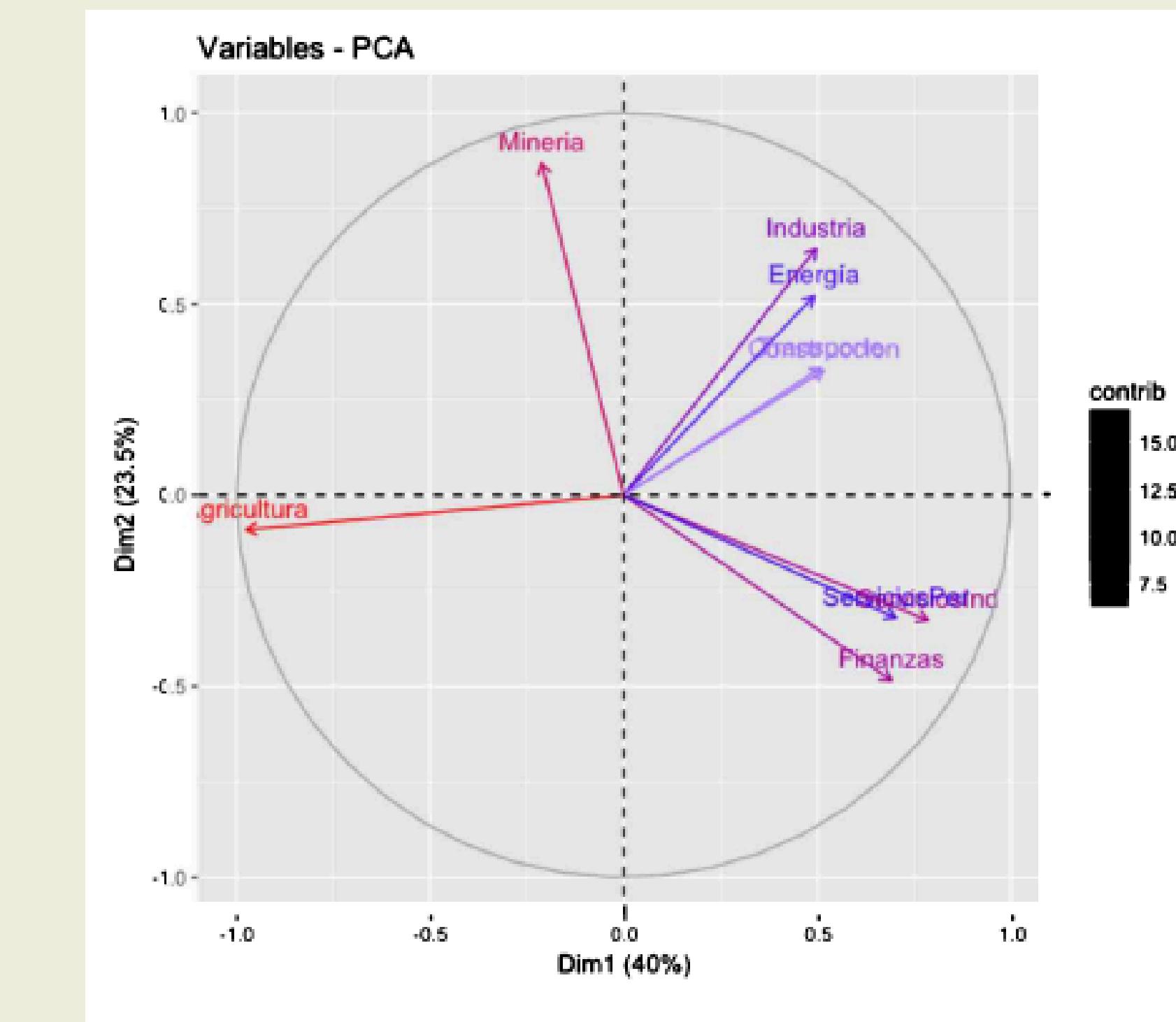
datos.grafico2<- data.frame(fit$var$coord[,1:2])

ggplot(datos.grafico2)+  
  geom_point (aes (x=Dim. 1, y=Dim. 2,  
 colour="darkred"))+  
  geom_text_repel (aes(x=Dim.1, y=Dim. 2),  
 label rownames (datos.grafico2))+  
  geom_vline(xintercept = 0, colour="darkgray") +  
  geom_hline (yintercept = 0, colour="darkgray") +  
  labs (x="Dimension 1 (40.01%)", y="Dimension 2  
(23.46%)") + theme (legend.position="none")
```

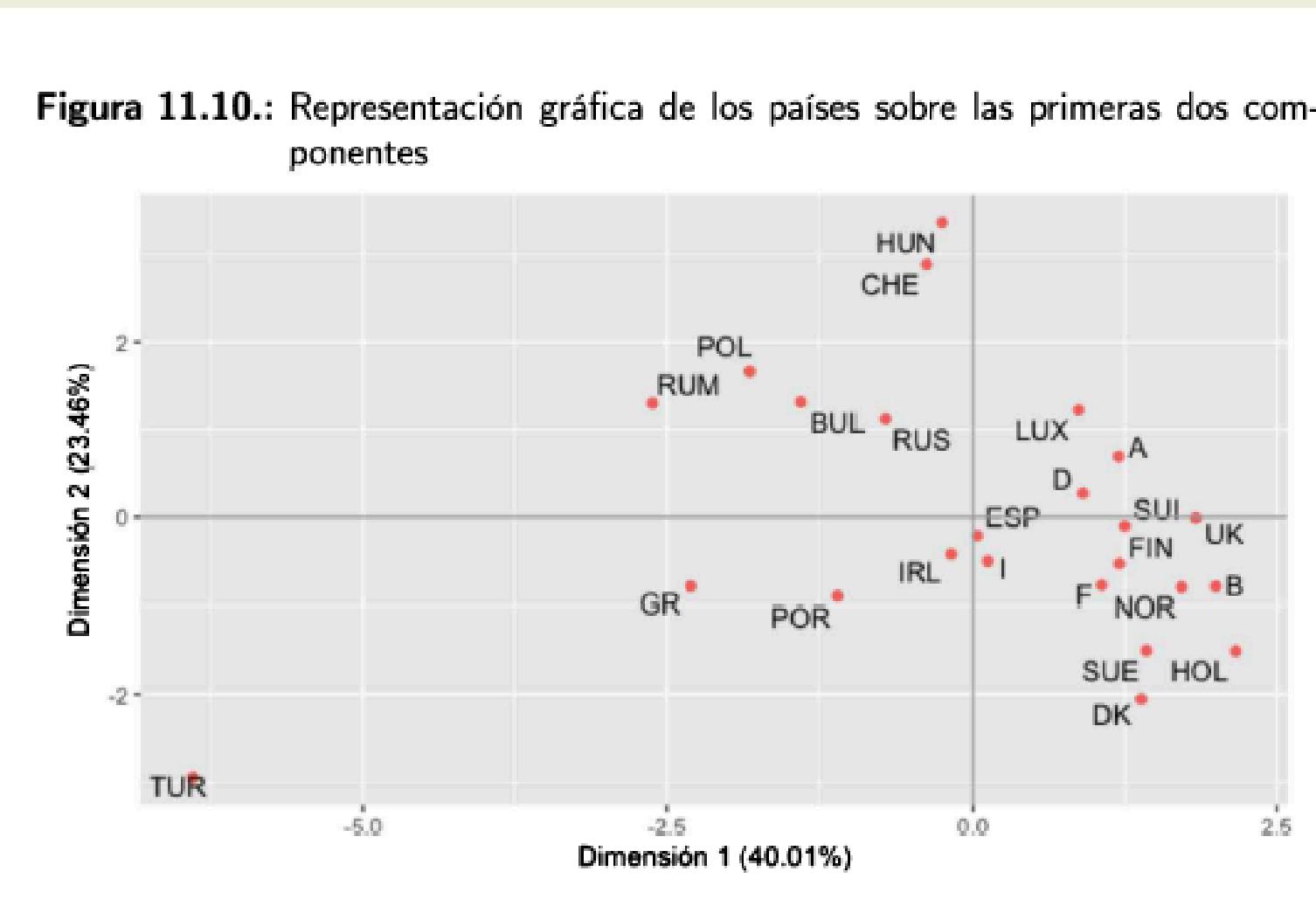
Ejemplo

Graficar las variables en el plan multidimensional (flchas indican dirección y carga)

```
library(ggplot2)  
  
library(factoextra)  
fviz_pca_var(fit, col.var="contrib") +  
  scale_color_gradient2 (low="white",  
mid="blue",high="red", midpoint=10.0) +  
  theme_gray()
```



Graficar casos (cuando son pocos y tiene sentido) en el plano multidimensional



```
datos.grafico<-
  data.frame(fit$ind$coord[,1:2],datos$Pais)

  colnames(datos.grafico)<-c("Dim. 1",
  "Dim. 2","pais")
  ggplot(datos.grafico)+  

    geom_point (aes (x=Dim. 1, y=Dim. 2,  

  colour="darkred"))+  

    geom_text_repel (aes (x=Dim. 1, y=Dim.  

  2), label=datos.grafico$pais)+  

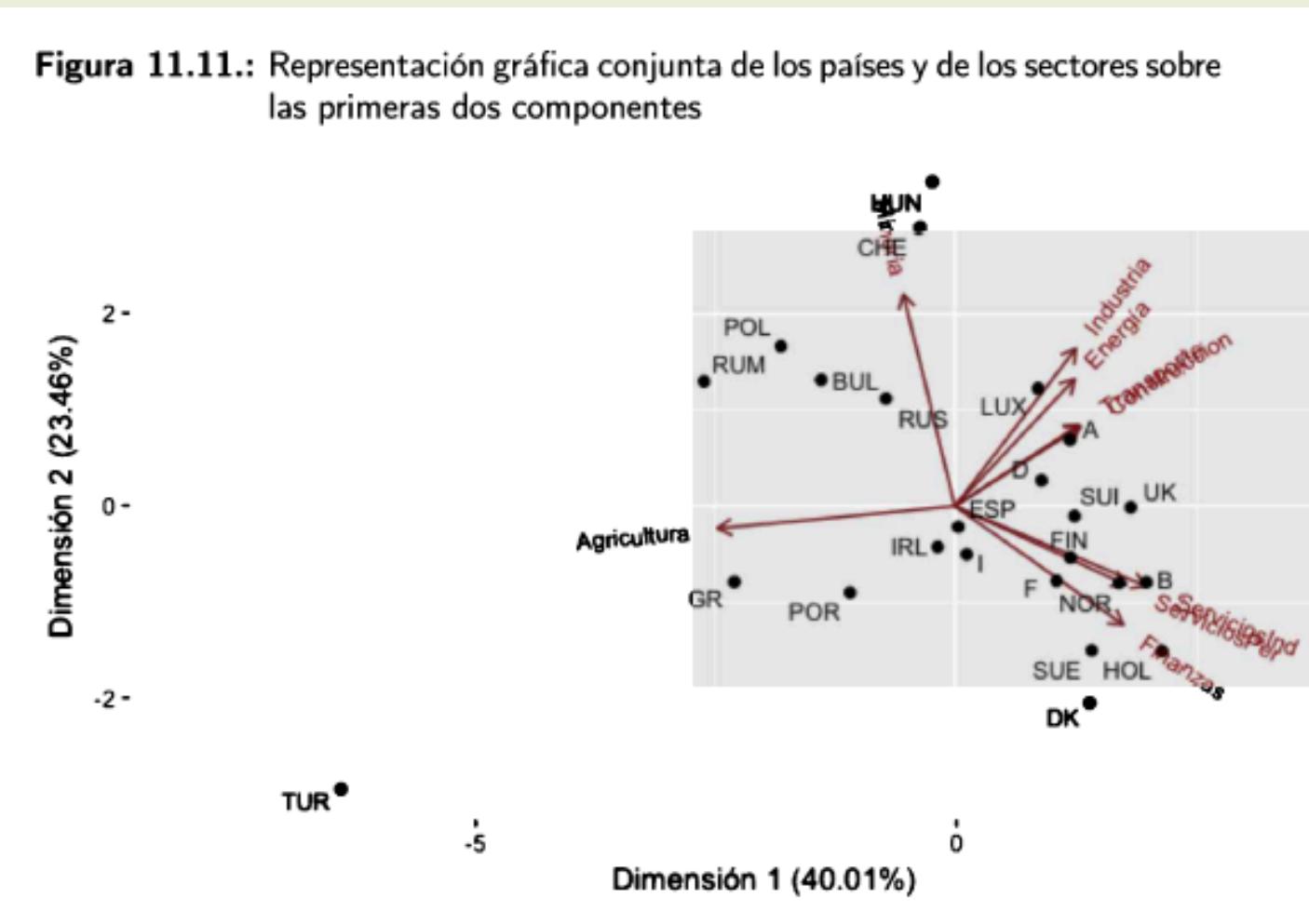
    geom_vline (xintercept = 0,  

  colour="darkgray")+
  geom_hline (yintercept = 0,  

  colour="darkgray")+
  labs (x="Dimension 1 (40.01%)",
  y="Dimension 2 (23.46%)")+ theme  

  (legend.position="none")
```

Graficar en conjunto casos y variables (cuando tiene sentido graficar casos)



```
library(ggbiplot)
ggbiplot (fit, obs.scale = 1,
var.scale = 1)+
  scale_color_discrete (name = '')+
  expand_limits (x=c(-8,4), y=c(-2.5,
2.5))+
  labs (x="Dimension 1 (40.01%)",
y="Dimension 2 (23.46%)")+
  geom_text_repel
(aes(x=datos.grafico$Dim.1,
y=datos.grafico$Dim.2),
label=datos.grafico$pais, size=3)
```

Módulo 2 Cálculo e interpretación PCA

Realizar un cálculo de PCA siguiendo el ejercicio cargado en U Campus. Puede usar la base ELSOC u otra base (bases cargadas en “enlaces” en U Campus)

Entregar script con:

- cálculo de la matriz de correlaciones con interpretación (0,1 pto)
- calculo de los autovalores y autovectores con interpretación (0,1 pto)
- opcional: gráficos
- **una interpretación general (0,1 pto)**