

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 5
10 sept

- **Repaso Análisis de Componentes Principales (PCA)**
- **Elección de número de componentes**
- **Creación de Índice**
- **Ejercicio práctico**



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 precaucion 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).

En general la **recomendación es tipificar las variables** (sobre todo cuando las variables están en escalas diferentes (ej. ingreso en pesos vs. edad en años y/o tienen varianzas muy diferentes)

La función Función “fit” del paquete FactorMineR de R studio lo hace con este código

```
library(FactoMineR)  
fit <- PCA(datos,  
scale.unit = TRUE, # estandariza variables  
ncp = ncol(datos_sel), # nº máx. de componentes
```

Análisis de componentes principales

02

Clase hoy

- repaso
- decisión sobre cuántos componentes principales seleccionar
- creación de índice

Caso simulado de países

- Repaso

Cálculo de número de componentes principales

Creación de Índice

Análisis de componentes principales

Repaso

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)

R Studio

1

Confirmar que las variables que se ingresan al modelo están correlacionadas

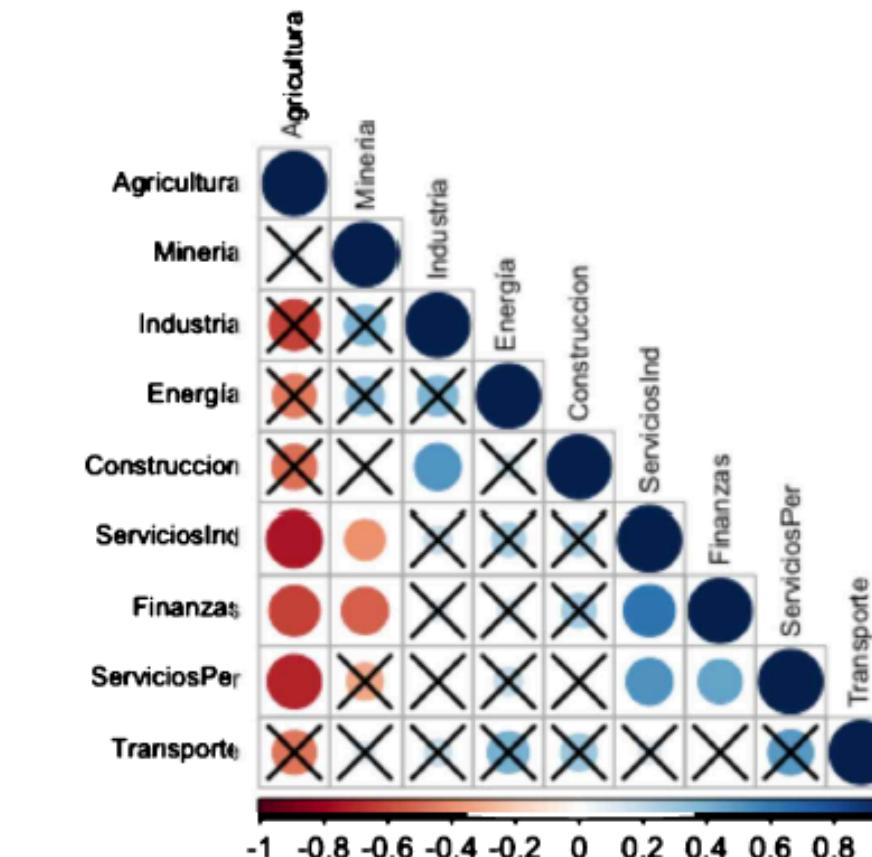
El signo da la dirección de la relación y el número la fuerza (> 0,5 hay correlación; cerca de 1 ésta es fuerte; cerca de 0, no existe)

```
# Instalar paquete corplot  
install.packages("corrplot")  
  
# Cargar librería  
library(corrplot)  
  
# Matriz de correlaciones  
matriz_cor <- cor(na.omit(datos))
```

	barrio_ideal	integrado	identifico	parte_de_mi	amigos
barrio_ideal	1.0000000	0.6970499	0.6256326	0.6708501	0.3694989
integrado	0.6970499	1.0000000	0.7423474	0.7211461	0.4676397
identifico	0.6256326	0.7423474	1.0000000	0.7256356	0.4861908
parte_de_mi	0.6708501	0.7211461	0.7256356	1.0000000	0.4511865
amigos	0.3694989	0.4676397	0.4861908	0.4511865	1.0000000
sociable	0.3603638	0.4546446	0.4768025	0.4281252	0.6548423
cordialidad	0.4458254	0.4875200	0.5220798	0.4754098	0.5432083
colaboracion	0.3518064	0.4160764	0.4426210	0.4083191	0.4777522
	sociable	cordialidad	colaboracion		
barrio_ideal	0.3603638	0.4458254	0.3518064		
integrado	0.4546446	0.4875200	0.4160764		
identifico	0.4768025	0.5220798	0.4426210		
parte_de_mi	0.4281252	0.4754098	0.4083191		
amigos	0.6548423	0.5432083	0.4777522		
sociable	1.0000000	0.6703941	0.5560530		
cordialidad	0.6703941	1.0000000	0.5714658		
colaboracion	0.5560530	0.5714658	1.0000000		

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

Figura 11.6.: Matriz de correlaciones entre las variables del caso y significatividad ($p<0.05$).



Nota: Las aspas (X) señalan correlaciones no significativas

R Studio

2

Conocer el porcentaje de varianza explicada por cada componente

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

Cuadro 11.11.: Aplicación del criterio del autovalor >1

\$eig	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

Interpretación:

Esta tabla nos indica los valores propios y el % de varianza explicada por cada componente principal (hay tantos componentes como variables originales). Están ordenados en orden decreciente.

El Componente 1 explica el 40,01% de la varianza total

El Componente 2 explica el 23,46% de la varianza total

En total explican el 63,47% de la varianza total

R Studio

3

Conocer las “cargas” de las variables originales sobre cada componente

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

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

Interpretación:

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.

R Studio

3

Conocer las “cargas” de las variables originales sobre cada componente

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

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

Interpretación:

Esta tabla nos indica la carga (correlación) entre las variables originales y las nuevas variables “componentes principales” (Dim 1, Dim 2, etc)

Las variables que más carga tiene (más inciden) en la dimensión 1 son Agricultura (-), ServiciosInd, Finanzas y ServiciosPer. El signo separa ambos grupos.

En la Dim 2 las que más carga y positiva tienen son Minería, Industria, Energía (que tienen carga negativa con la Dim 1).

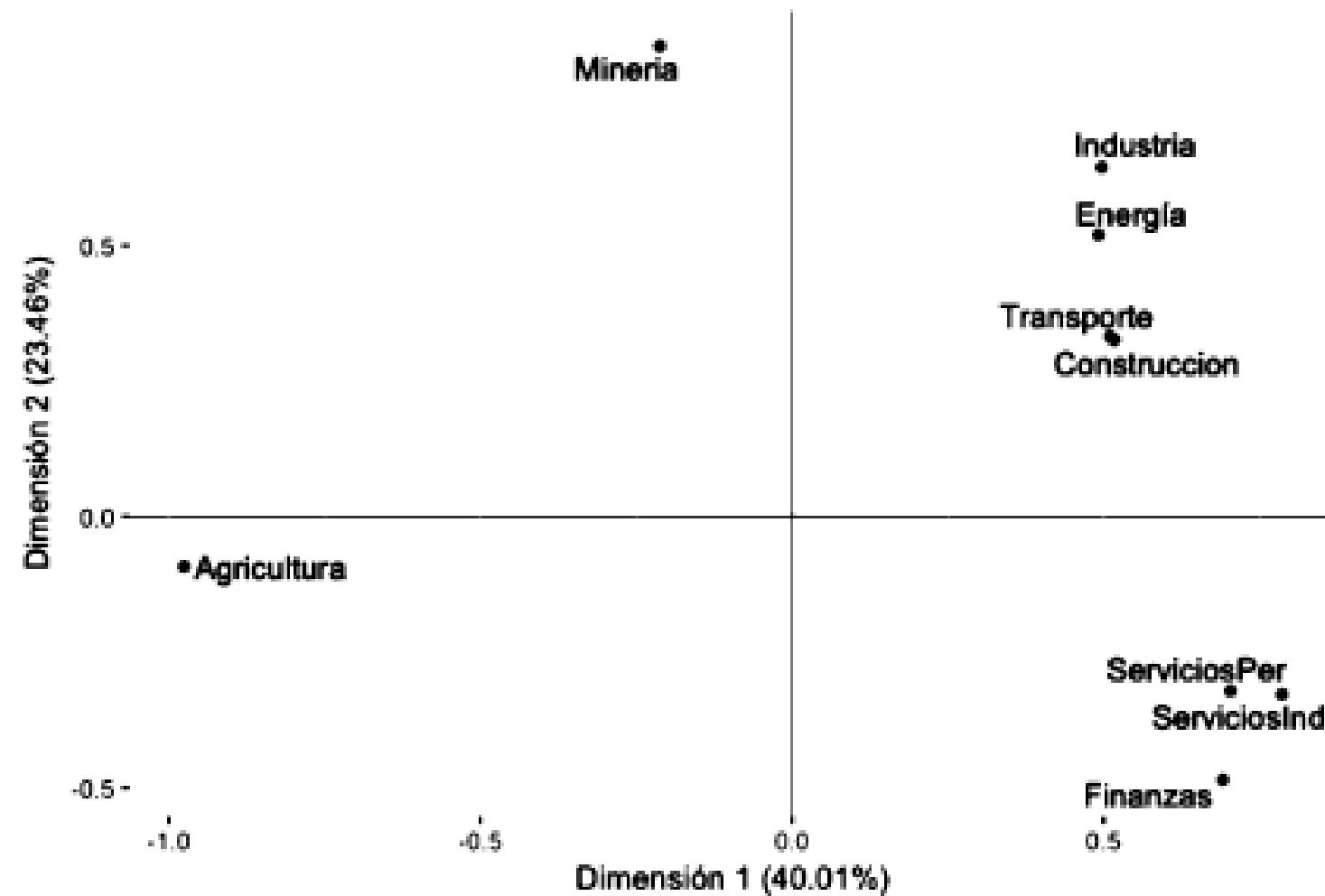
R Studio

4

Observar los gráficos poder interpretar mejor

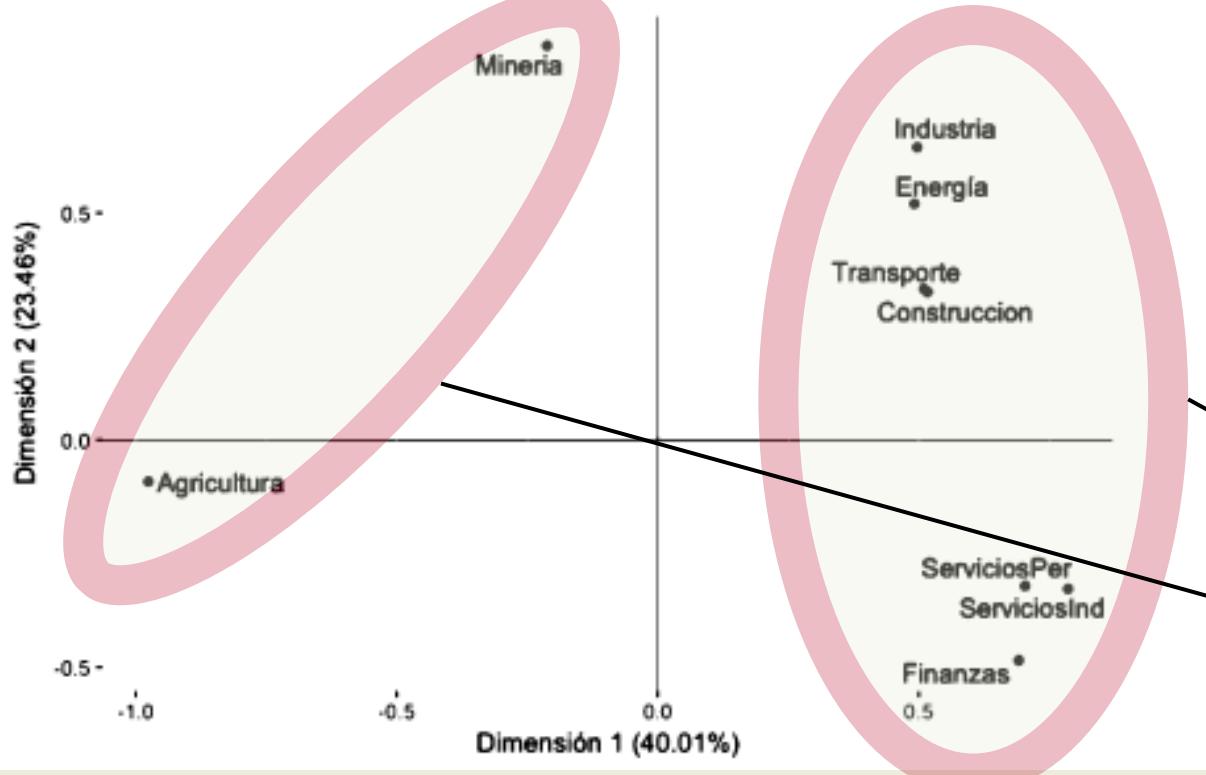
Estamos graficando las variables!!
No los casos, tampoco las categorías

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



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

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

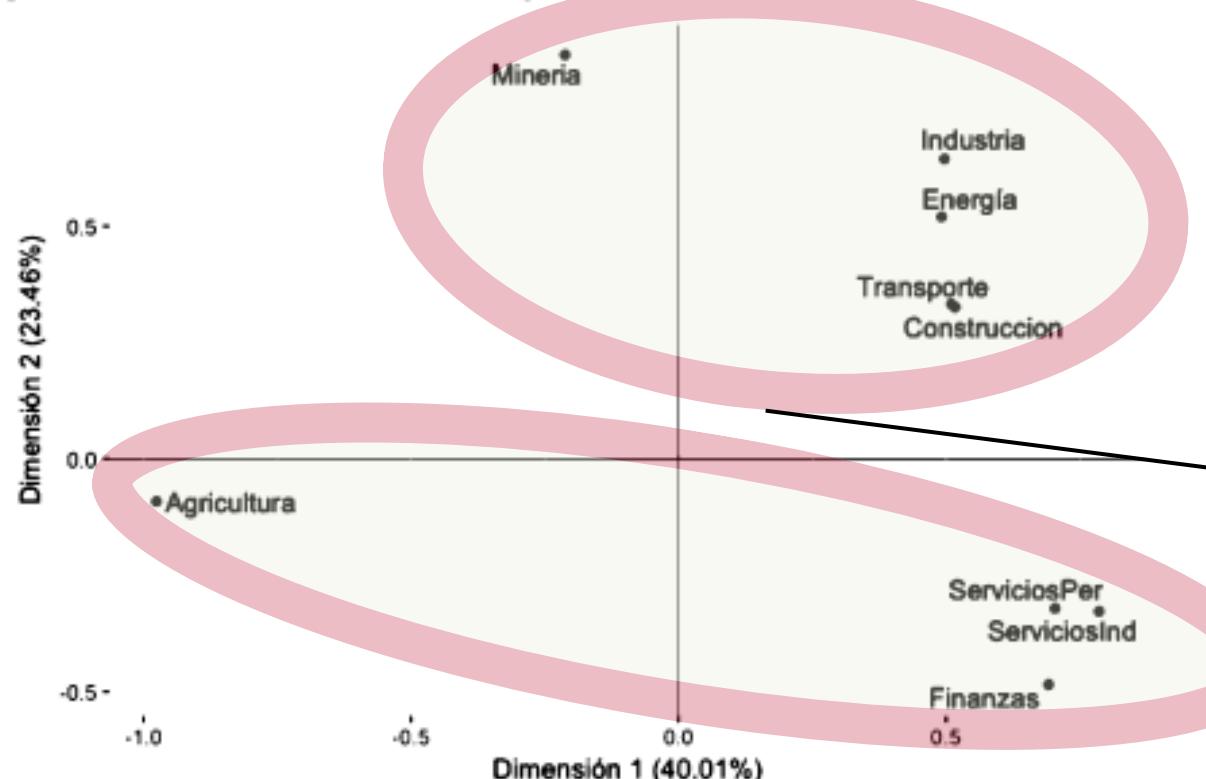


Agricultura (% de personas empleadas en Agricultura) y Minería está correlacionadas negativamente con Dim 1 (a la izquierda sobre el eje horizontal). Pero Minería con baja carga

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)

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)

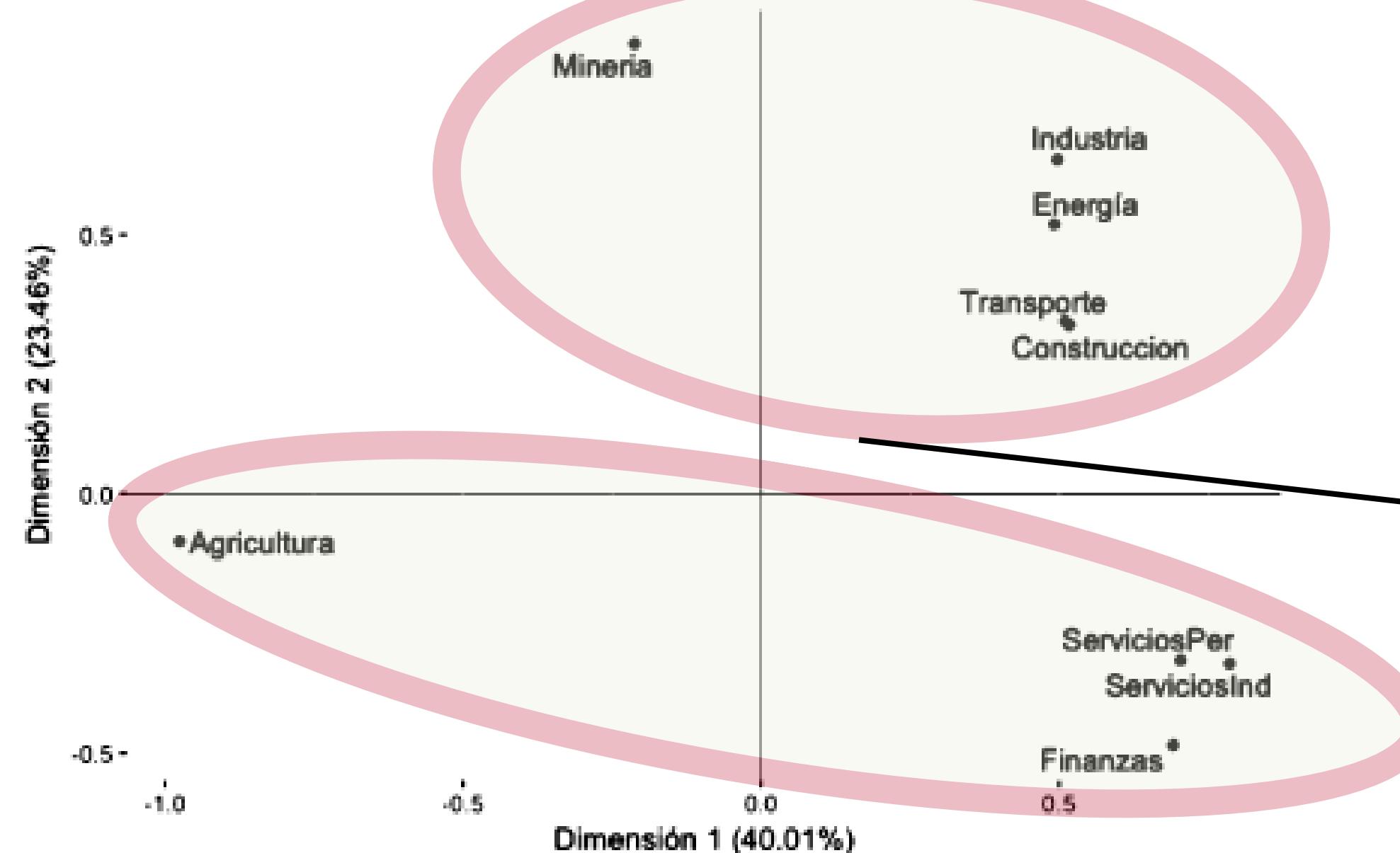
Las otras variables están correlacionadas negativamente con Dim 2 (abajo sobre el eje vertical)

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.

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



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

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

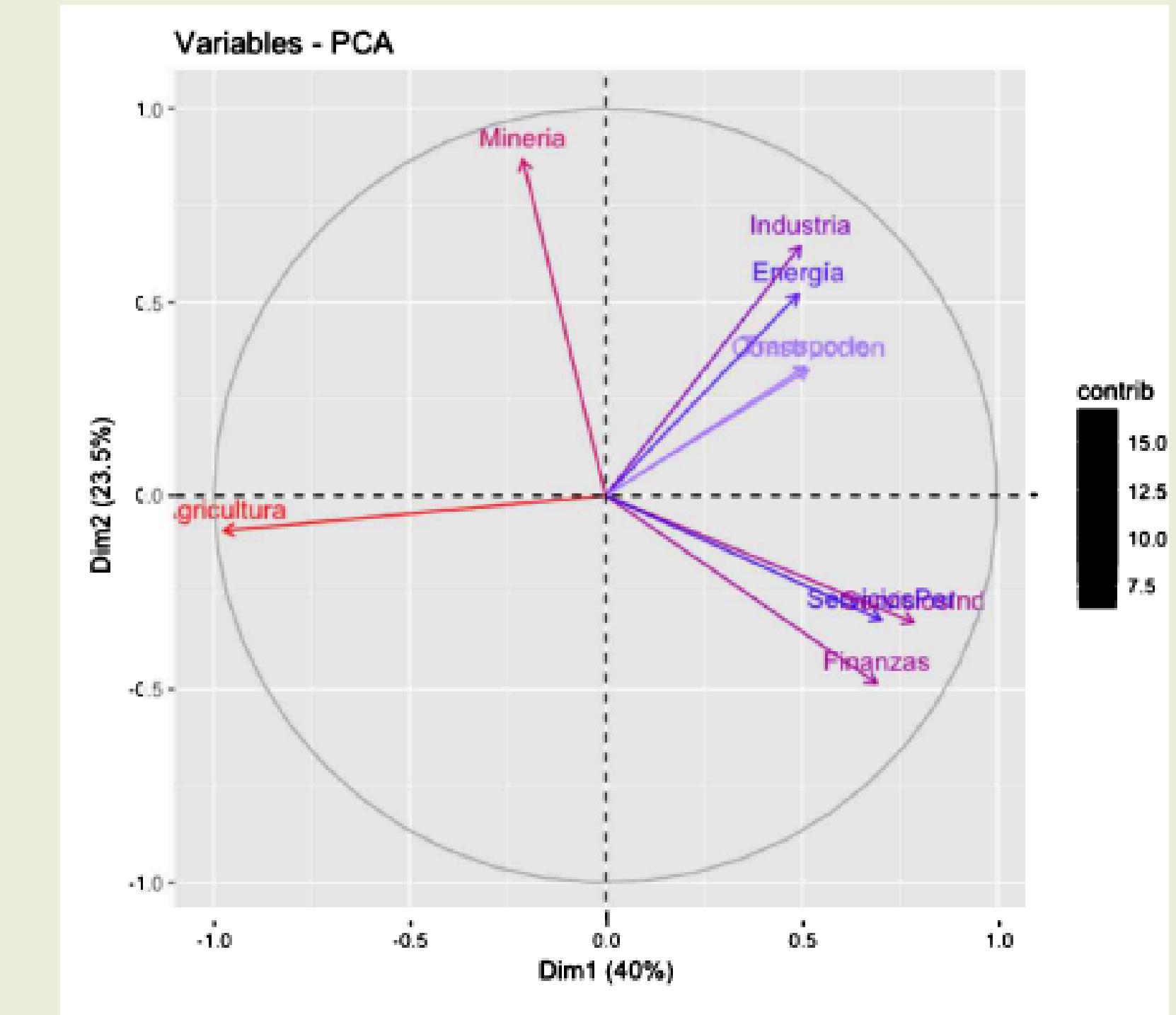
Interpretación: la componente principal 2 (Dim 2) distingue entre **países con economías con menor desarrollo en sus servicios** de aquellos **con mayor desarrollo en sus servicios**.

R Studio

4

Observar los gráficos poder interpretar mejor

```
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()
```

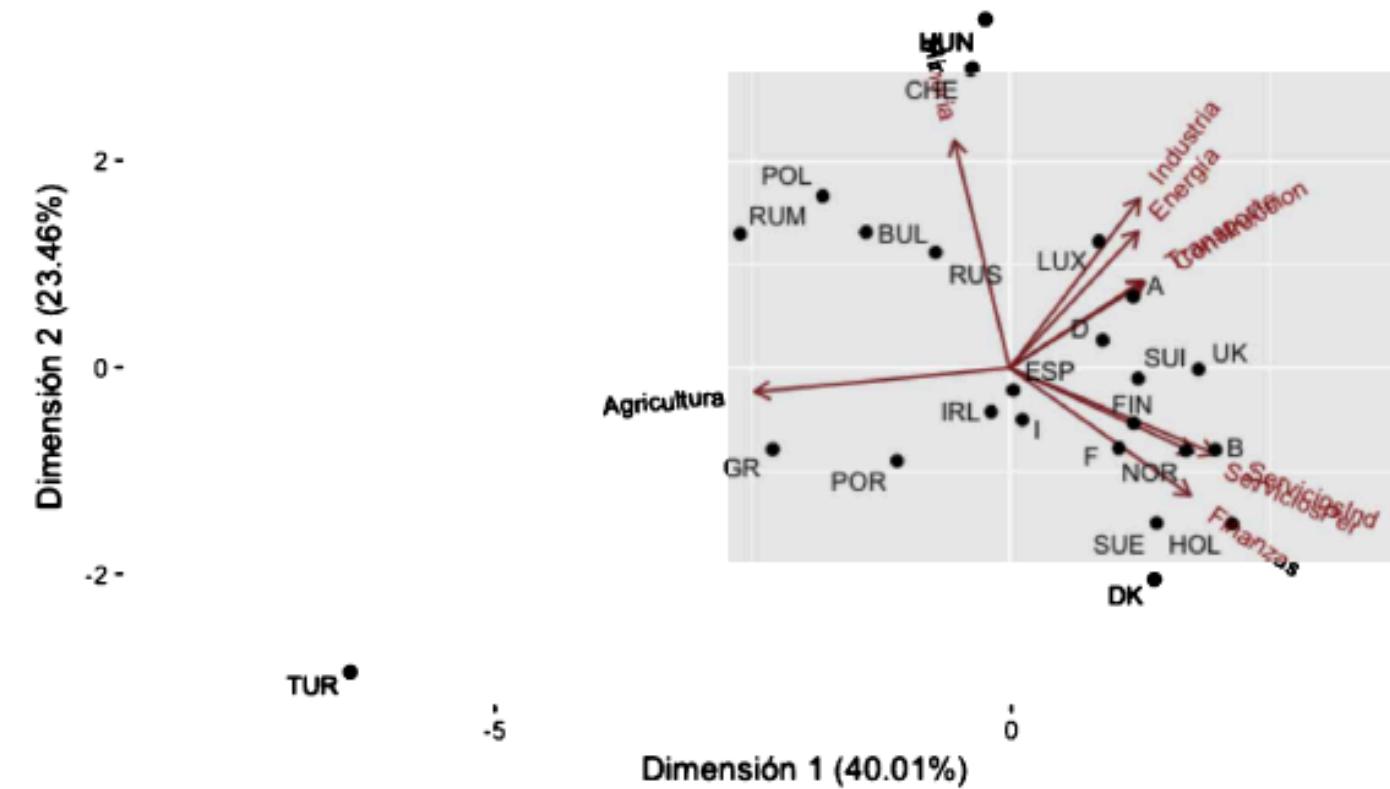


R Studio

4

Observar los gráficos poder interpretar mejor

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



```
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)
```

Graficamos los casos cuando son pocos (ej países) y su distribución en el espacio nos ayuda a interpretar porque conocemos las características de cada caso (país)

Análisis de componentes principales

02

Clase hoy

- repaso
- decisión sobre cuántos componentes principales seleccionar

Caso simulado de dos variables

- Repaso

Cálculo de número de componentes principales

Creación de índice

R Studio

5

¿Cuántos componentes principales debemos usar?

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

Cuadro 11.11.: Aplicación del criterio del autovalor >1

	\$eig	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	

R Studio

5

¿Cuántos componentes principales debemos usar?

Criterio 1

El autovalor de la Componente > 1

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	0.8551719711	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

Nos quedaríamos con los tres primeros componentes principales

R Studio

5

¿Cuántos componentes principales debemos usar?

Criterio 2

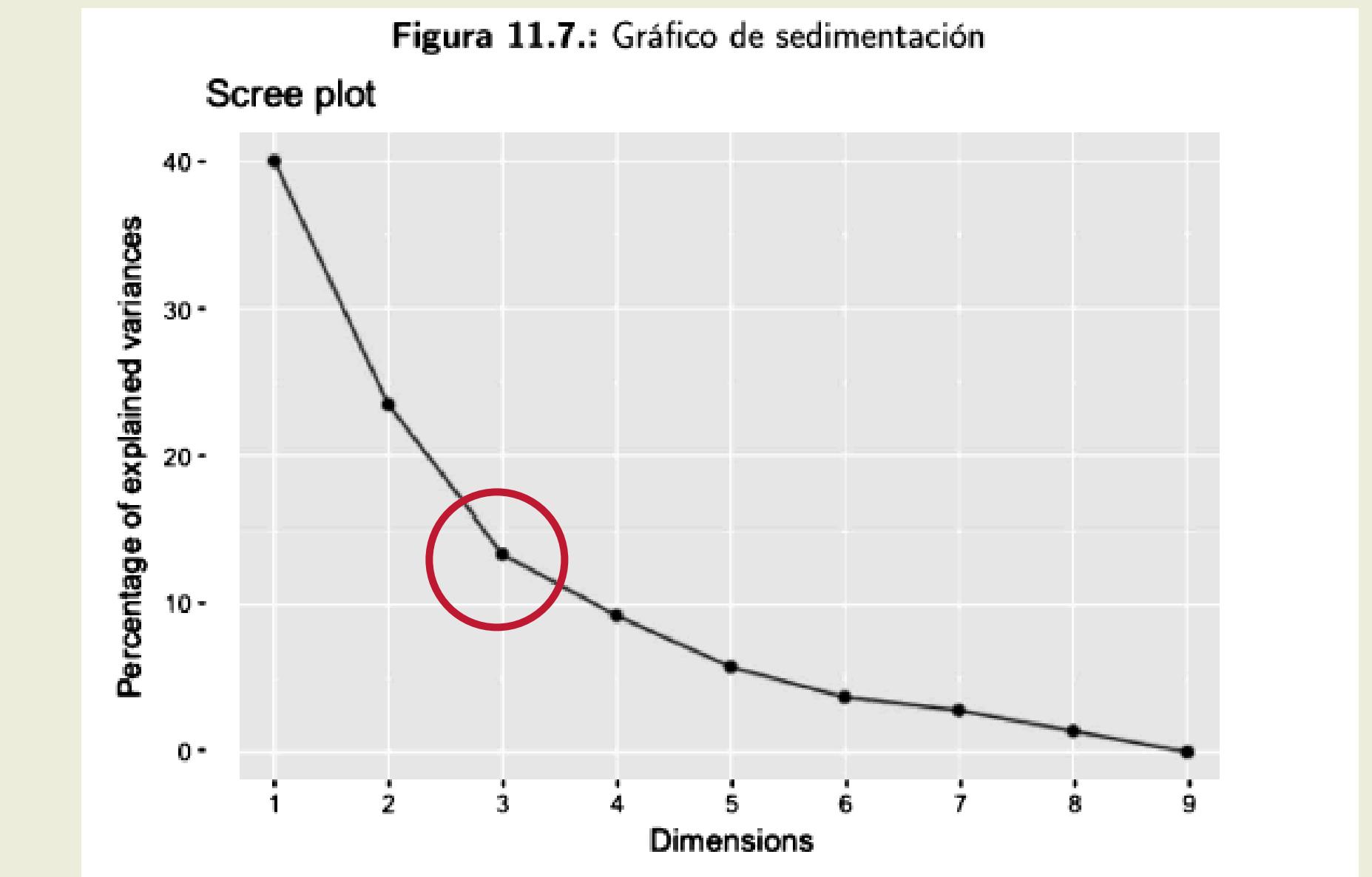
Codo o gráfico de sedimentación

Se hace a partir de la representación gráfica de los autovalores de las componentes.

Se retienen aquellas que (en el gráfico) muestran que la línea tiene una pendiente fuerte.

```
library (factoextra)
```

```
fviz_eig(fit, geom="line") + theme_grey()
```



R Studio

5

¿Cuántos componentes principales debemos usar?

Criterio 3 “Parallel Analysis”

Objetiva el criterio del codo mediante la función paran (paran) de R.
Solo se retienen los componentes con autovalores “ajustados”
superiores a 1.

```
library (paran)
paran(na.omit(proc_data[2:8]),
      iterations = 5000,
      graph = TRUE,
      color = FALSE)
```

Cuadro 11.12.: Análisis paralelo de Horn (1965)
Results of Horn's Parallel Analysis for component retention
5000 iterations, using the mean estimate

Component	Adjusted Eigenvalue	Unadjusted Eigenvalue	Estimated Bias
1	2.520492	3.600935	1.080443
2	1.447376	2.111346	0.663969

Análisis de componentes principales

02

Clase hoy

- repaso
- decisión sobre cuántos componentes principales seleccionar

Caso simulado de dos variables

- Repaso

Cálculo de número de componentes principales

Creación de índice

Uno de los usos frecuentes que se le da al PCA es quedarse con el primer componente principal, cuando representa un porcentaje muy grande de la varianza total explicada, y tratarlo como una nueva variable que toma de la forma de un índice

Ventajas:

Resume en una sola variable la información de todas las variables del modelo. La ponderación que se da a cada variable depende de la correlación de las variables entre si, y de éstas con la Componente Principal.

Índice = PC1 = combinación lineal de todas las variables originales

$$PC_1 = w_1X_1 + w_2X_2 + w_3X_3 + \dots + w_pX_p$$

- X_1, X_2, \dots, X_p = variables originales (estandarizadas si corresponde).
- w_1, w_2, \dots, w_p = cargas factoriales asociados a cada variable en ese componente

Ejemplo (solo válido para fines del ejercicio)

Índice = PC1 = combinación lineal de todas las variables originales.

Tomamos 8 variables de escala Likert del módulo Territorial de la Encuesta ELSOC:
(solo se toma para fines de ejercicio, porque la escala va de 1 a 4 y es deseable en general usar escalas de rango más amplio para este modelo)

t02_01	Grado de acuerdo: Este es el barrio ideal para mi
t02_02	Grado de acuerdo: Me siento integrado/a en este barrio
t02_03	Grado de acuerdo: Me identifico con la gente de este barrio
t02_04	Grado de acuerdo: Este barrio es parte de mi
t03_01	Grado de acuerdo: En este barrio es facil hacer amigos
t03_02	Grado de acuerdo: La gente en este barrio es sociable
t03_03	Grado de acuerdo: La gente en este barrio es cordial
t03_04	Grado de acuerdo: La gente en este barrio es colaboradora

Ejemplo (solo válido para fines del ejercicio)

Índice = PC1 = combinación lineal de todas las variables originales.

Calculamos la matriz de correlación

```
matriz_cor <- cor(na.omit(proc_data))
matriz_cor

##           barrio_ideal integrado identifico parte_de_mi    amigos sociable
## barrio_ideal      1.0000000 0.6970499  0.6256326  0.6708501 0.3694989 0.3603638
## integrado        0.6970499 1.0000000  0.7423474  0.7211461 0.4676397 0.4546446
## identifico       0.6256326 0.7423474  1.0000000  0.7256356 0.4861908 0.4768025
## parte_de_mi      0.6708501 0.7211461  0.7256356  1.0000000 0.4511865 0.4281252
## amigos           0.3694989 0.4676397  0.4861908  0.4511865 1.0000000 0.6548423
## sociable          0.3603638 0.4546446  0.4768025  0.4281252 0.6548423 1.0000000
## cordialidad      0.4458254 0.4875200  0.5220798  0.4754098 0.5432083 0.6703941
## colaboracion     0.3518064 0.4160764  0.4426210  0.4083191 0.4777522 0.5560530
##           cordialidad colaboracion
## barrio_ideal     0.4458254   0.3518064
## integrado        0.4875200   0.4160764
## identifico       0.5220798   0.4426210
## parte_de_mi      0.4754098   0.4083191
## amigos           0.5432083   0.4777522
## sociable          0.6703941   0.5560530
## cordialidad      1.0000000   0.5714658
## colaboracion     0.5714658   1.0000000
```

Ejemplo (solo válido para fines del ejercicio)

Índice = PC1 = combinación lineal de todas las variables originales.

Calculamos las cargas factoriales de las variables

```
fit$var$coord

##           Dim.1      Dim.2      Dim.3      Dim.4      Dim.5
## barrio_ideal 0.7444588 -0.4321671  0.07382953 -0.17440822  0.44637941
## integrado    0.8256050 -0.3548389 -0.03191452  0.04667929 -0.04191985
## identifico   0.8312220 -0.2823344 -0.02653115  0.08641030 -0.31929597
## parte_de_mi   0.8059986 -0.3687355 -0.01307085  0.08003877 -0.13171815
## amigos        0.7170664  0.3630342 -0.45808825  0.26414336  0.17975301
## sociable       0.7432279  0.4789087 -0.16513341 -0.14680486 -0.05148399
## cordialidad   0.7627293  0.3487999  0.12388898 -0.44317734 -0.08832106
## colaboracion  0.6730152  0.3926693  0.53593419  0.31602450  0.07518876
##           Dim.6      Dim.7      Dim.8
## barrio_ideal -0.034476405  0.009909944  0.150340249
## integrado     -0.200178742 -0.212207990 -0.320011255
## identifico    -0.005189705 -0.168665084  0.301267080
## parte_de_mi    0.180652976  0.384555521 -0.099775719
## amigos         0.185143684 -0.088162291 -0.006587265
## sociable        -0.353363402  0.196734538  0.056720168
## cordialidad    0.244284347 -0.109167308 -0.073722928
## colaboracion  -0.010120143 -0.002475324  0.001601984
```

Ejemplo (solo válido para fines del ejercicio)

Índice = PC1 = combinación lineal de todas las variables originales.

Calculamos los autovalores y la varianza explicada de cada componente principal

```
fit$eig
```

```
##          eigenvalue percentage of variance cumulative percentage of variance
## comp 1  4.6776837      58.471047             58.47105
## comp 2  1.1653557      14.566946             73.03799
## comp 3  0.5470319       6.837899             79.87589
## comp 4  0.4340711       5.425889             85.30178
## comp 5  0.3687271       4.609089             89.91087
## comp 6  0.2928437       3.660546             93.57142
## comp 7  0.2798620       3.498275             97.06969
## comp 8  0.2344246       2.930308            100.00000
```

Ejemplo (solo válido para fines del ejercicio)

Índice = PC1 = combinación lineal de todas las variables originales.

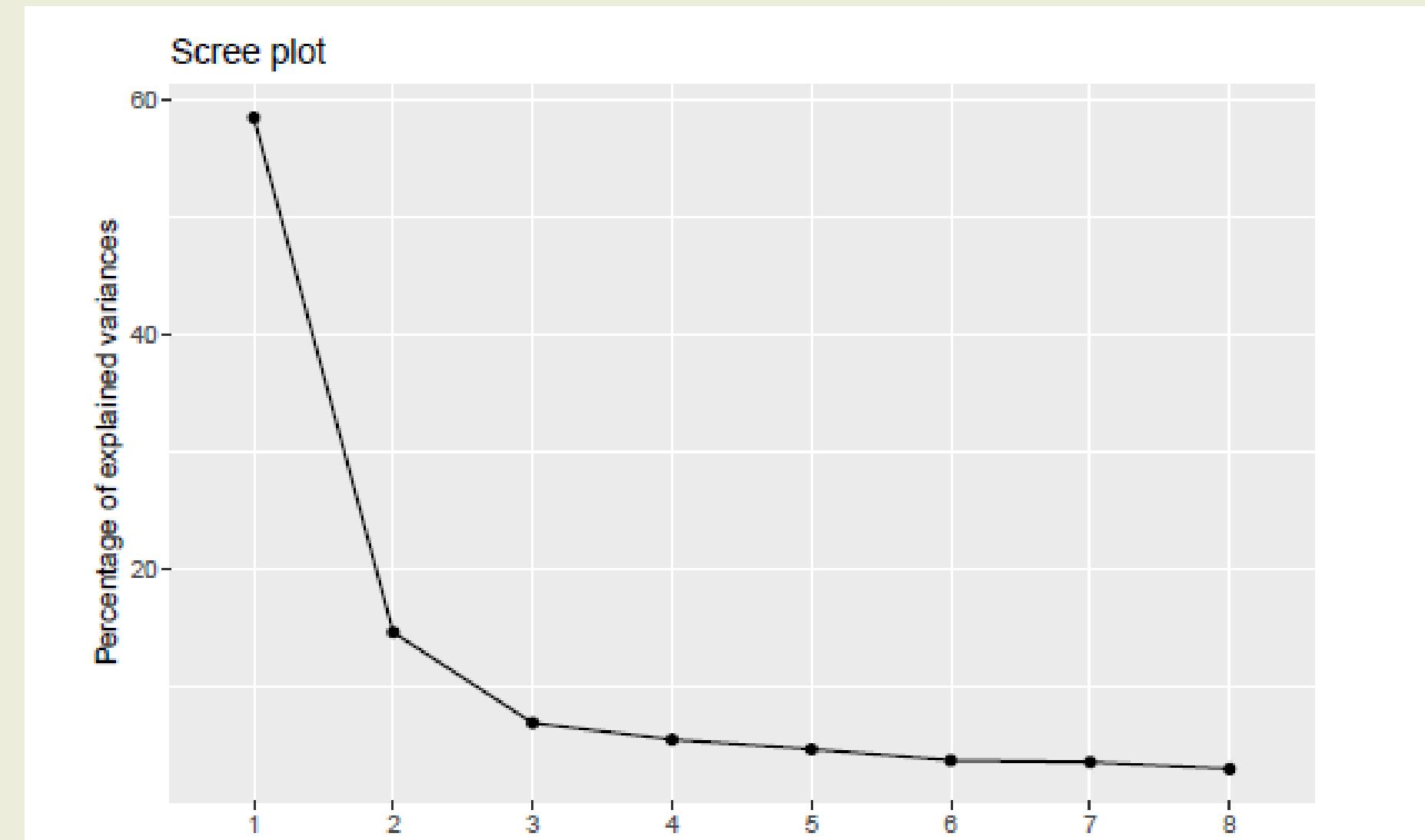
Calculamos los autovalores y la varianza explicada de cada componente principal

```
fit$eig
```

```
##          eigenvalue percentage of variance cumulative percentage of variance
## comp 1  4.6776837      58.471047             58.47105
## comp 2  1.1653557      14.566946             73.03799
## comp 3  0.5470319       6.837899             79.87589
## comp 4  0.4340711       5.425889             85.30178
## comp 5  0.3687271       4.609089             89.91087
## comp 6  0.2928437       3.660546             93.57142
## comp 7  0.2798620       3.498275             97.06969
## comp 8  0.2344246       2.930308            100.00000
```

Tanto el método de autovalor > 1 , como el método del codo parecieran estar sugiriendo usar dos componentes.

Sin embargo, la Dim 1 explica el 58,47 % de la varianza total, y la Dim 2 solo 14,57%. Además, la Dim.1 tiene un autovalor mucho mayor que la Dim.2 Y las correlaciones entre las variables originales y la Dim. 1 son altas, pero con la Dim.2 son bajas ($<0,5$)



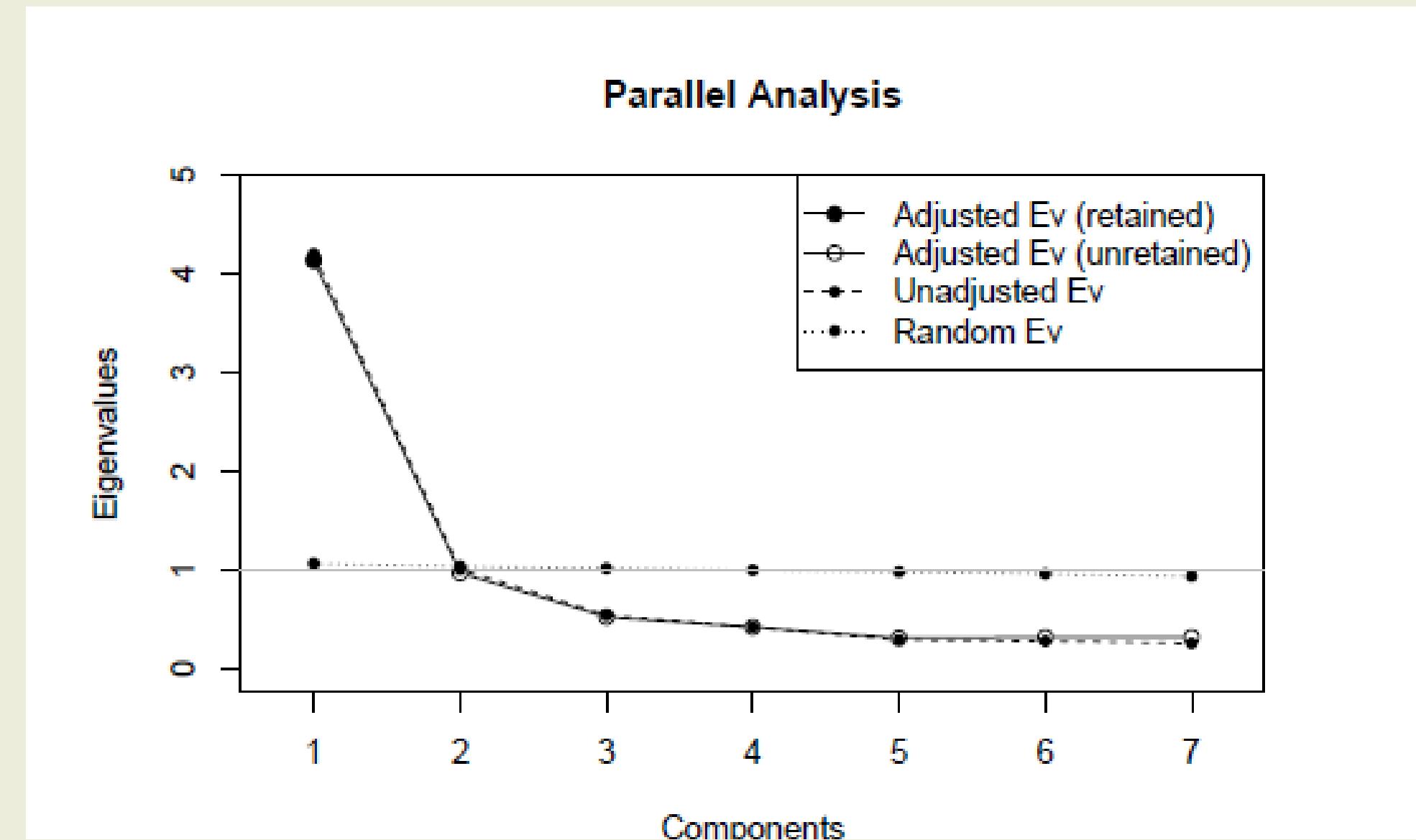
Ejemplo (solo válido para fines del ejercicio)

Índice = PC1 = combinación lineal de todas las variables originales.

```
##  
##  
## Component Adjusted Unadjusted Estimated  
##          Eigenvalue Eigenvalue Bias  
##  
## 1 4.138217 4.202646 0.064429  
##
```

El método paran nos indica, por su parte, que solo debe tomar 1 componente principal.

Tomaremos este resultado y construimos el índice de cohesión social con la PC1



Ejemplo (solo válido para fines del ejercicio)

Índice = PC1 = combinación lineal de todas las variables originales.

El método paran nos indica, por su parte, que solo debe tomar 1 componente principal.

Tomaremos este resultado y construimos el índice de cohesión social con la PC1

Para ello debemos tomar las cargas factoriales de las variables originales en la PC1

$$PC_1 = w_1X_1 + w_2X_2 + w_3X_3 + \dots + w_pX_p$$

X1, X2,...,Xp = variables originales (estandarizadas si corresponde).

w1, w2,...,wp = cargas factoriales asociados a cada variable en ese componente

$$PC_1 = 0,77 \times barrio_ideal + 0,83 \times integrado + 0,83 \times identifico + 0,81 \times parte_de_mi + \\ - 0,72 \times amigos + 0,74 \times sociable + 0,76 \times cordialidad + 0,67 \times colaboracion$$

Calculamos las cargas factoriales de las variables

fit\$var\$coord

	Dim.1	Dim.2	Dim.3
# barrio_ideal	0.7444588	-0.4321671	0.07382953
# integrado	0.8256050	-0.3548389	-0.03191452
# identifico	0.8312220	-0.2823344	-0.02653115
# parte_de_mi	0.8059986	-0.3687355	-0.01307085
# amigos	0.7170664	-0.3630342	-0.45808825
# sociable	0.7432279	-0.4789087	-0.16513341
# cordialidad	0.7627293	-0.3487999	0.12388898
# colaboracion	0.6730152	-0.3926693	0.53593419
	Dim.7	Dim.8	
## barrio_ideal	-0.034476405	0.009909944	0.150340249
## integrado	-0.200178742	-0.212207990	-0.320011255
## identifico	-0.005189705	-0.168665084	0.301267080
## parte_de_mi	0.180652976	0.384555521	-0.099775719
## amigos	0.185143684	-0.088162291	-0.006587265
## sociable	-0.353363402	0.196734538	0.056720168
## cordialidad	0.244284347	-0.109167308	-0.073722928
## colaboracion	-0.010120143	-0.002475324	0.001601984

Ejemplo (solo válido para fines del ejercicio)

Índice de cohesión barrial

$$PC_1 = 0,77 \times \text{barrio_ideal} + 0,83 \times \text{integrado} + 0,83 \times \text{identifico} + 0,81 \times \text{parte_de_mi} + \\ - 0,72 \times \text{amigos} + 0,74 \times \text{sociable} + 0,76 \times \text{cordialidad} + 0,67 \times \text{colaboracion}$$

Rango: 6,13 a 24,52 (se calcula usando el valor mínimo en cada variable original (=1), y el valor máximo (=4))

A partir de esta variable podríamos tratar de conocer que tipo de personas tienen índices altos de cohesión social (según género, edad, nivel educacional, nivel educativo, tipo de barrio, etc.).
Ello mediante modelos de regresión o de correspondencia, por ejemplo.