



CIENCIA DE DATOS

ACP: Análisis de componentes principales

MATEMÁTICA PARA CIENCIA DE DATOS

PROF. ESTEBAN BALLESTERO

Algunos problemas iniciales

- ▶ ¿Cómo podemos visualizar datos con más de dos variables?
- ▶ ¿Por qué ACP es un método de reducción de la dimensión?
- ▶ ¿A qué se refiere el concepto de inercia?
- ▶ ¿Cuál sería el mejor ángulo para tomar la foto de manera que la mayoría de los individuos queden bien representados?

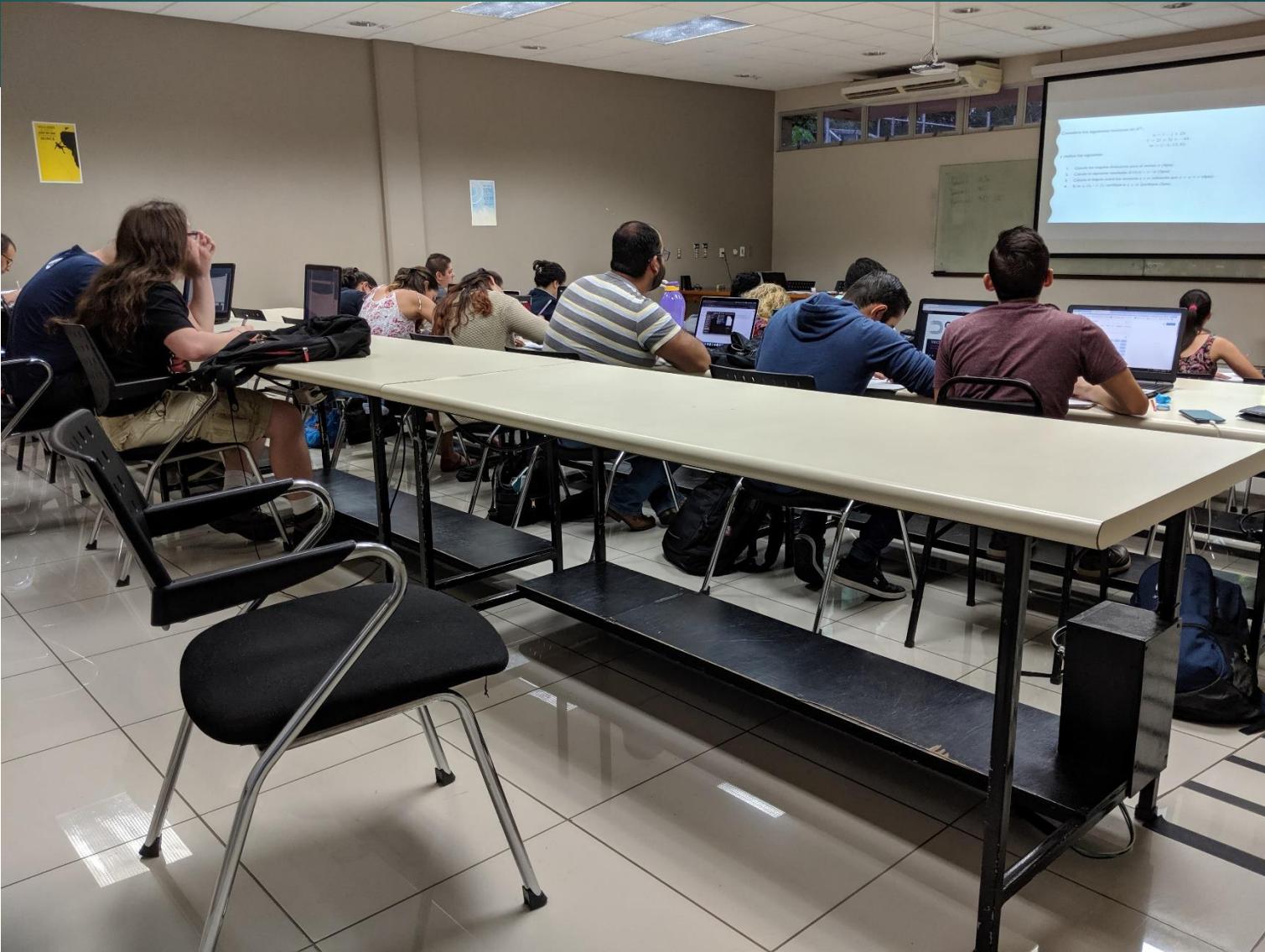
Foto01



Foto02



Foto03





Los datos

Se parte de una tabla de datos:

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{im} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{nm} \end{pmatrix} \leftarrow \text{individuo } i$$


Variable j

Ejemplo de tabla de datos:

	Matemáticas	Ciencias	Español	Historia	EdFísica
Lucía	7.0	6.5	9.2	8.6	8.0
Pedro	7.5	9.4	7.3	7.0	7.0
Inés	7.6	9.2	8.0	8.0	7.5
Luis	5.0	6.5	6.5	7.0	9.0
Andrés	6.0	6.0	7.8	8.9	7.3
Ana	7.8	9.6	7.7	8.0	6.5
Carlos	6.3	6.4	8.2	9.0	7.2
José	7.9	9.7	7.5	8.0	6.0
Sonia	6.0	6.0	6.5	5.5	8.7
María	6.8	7.2	8.7	9.0	7.0

Nube de puntos...

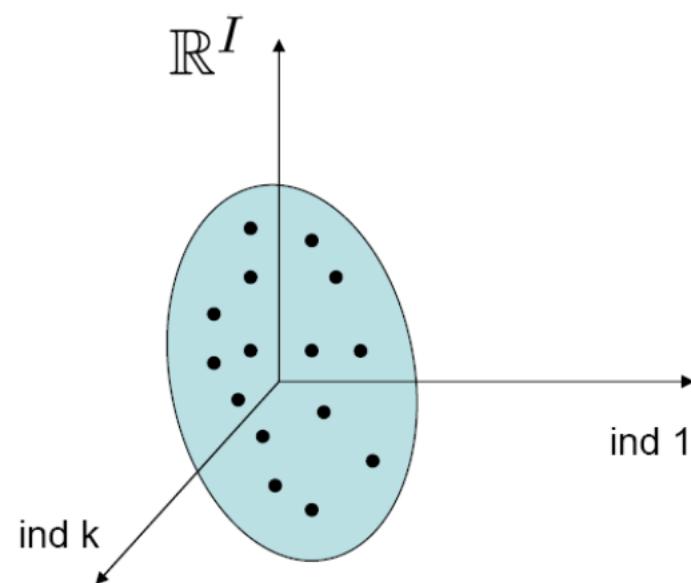
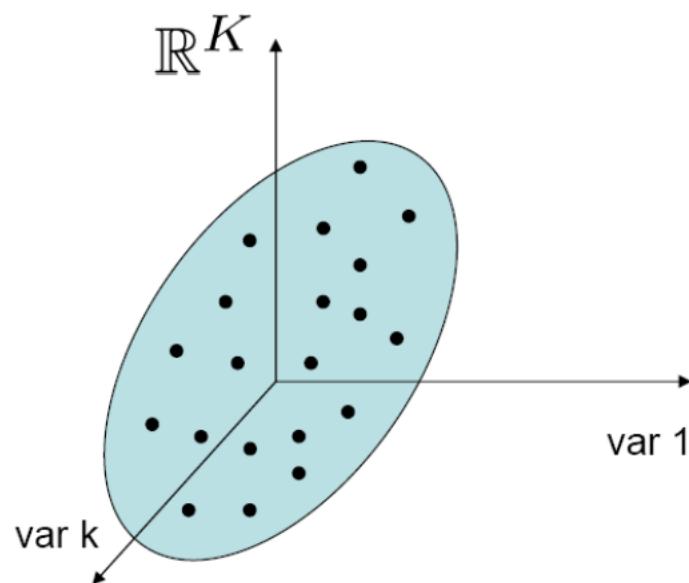
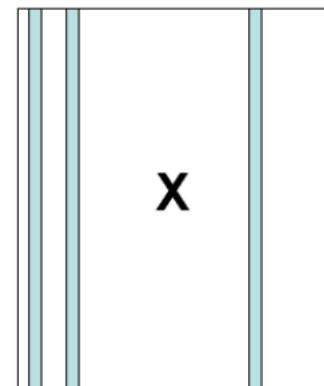
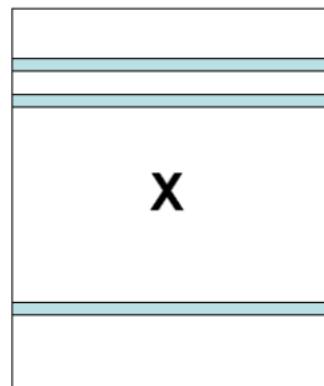
INDIVIDUOS - FILAS

$$\begin{matrix} \text{Luis} & 5.0 & 6.5 & 6.5 & 7.0 & 9.0 \end{matrix} \in \mathbb{R}^5$$

VARIABLES - COLUMNAS

Español
9.2
7.3
8.0
6.5
7.8
7.7
8.2
7.5
6.5
8.7

$$\in \mathbb{R}^{10}$$



Datos crudos vs componentes

Tabla de Datos

$$\begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}$$



Componentes

$$\begin{bmatrix} C_{11} & C_{12} & \cdots & C_{1p} \\ C_{21} & C_{22} & \cdots & C_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ C_{n1} & C_{n2} & \cdots & C_{np} \end{bmatrix}$$

100% de la información

80%

16%

.....

0.02%

Transformación de las variables originales en componentes

Datos crudos vs CP

DATOS

	Matemáticas	Ciencias	Español	Historia	EdFísica
Lucía	7.0	6.5	9.2	8.6	8.0
Pedro	7.5	9.4	7.3	7.0	7.0
Inés	7.6	9.2	8.0	8.0	7.5
Luis	5.0	6.5	6.5	7.0	9.0
Andrés	6.0	6.0	7.8	8.9	7.3
Ana	7.8	9.6	7.7	8.0	6.5
Carlos	6.3	6.4	8.2	9.0	7.2
José	7.9	9.7	7.5	8.0	6.0
Sonia	6.0	6.0	6.5	5.5	8.7
María	6.8	7.2	8.7	9.0	7.0

COMPONENTES

	Comp1	Comp2	Comp3	Comp4	Comp5
Lucia	0,3231	1,7725	1,1988	-0,055	0,0036
Pedro	0,6654	-1,6387	0,1455	-0,0231	-0,1234
Ines	1,0025	-0,5157	0,6289	0,5164	0,1429
Luis	-3,1721	-0,2628	-0,382	0,6778	-0,0625
Andres	-0,4889	1,3654	-0,8352	-0,1558	0,1234
Ana	1,7086	-1,0217	-0,1271	0,0668	0,0253
Carlos	0,0676	1,4623	-0,5062	-0,1179	0,0131
Jose	2,0119	-1,2759	-0,5422	-0,1978	0,0174
Sonia	-3,042	-1,2549	0,4488	-0,64	0,0379
Maria	0,9239	1,3694	-0,0293	-0,0715	-0,1777

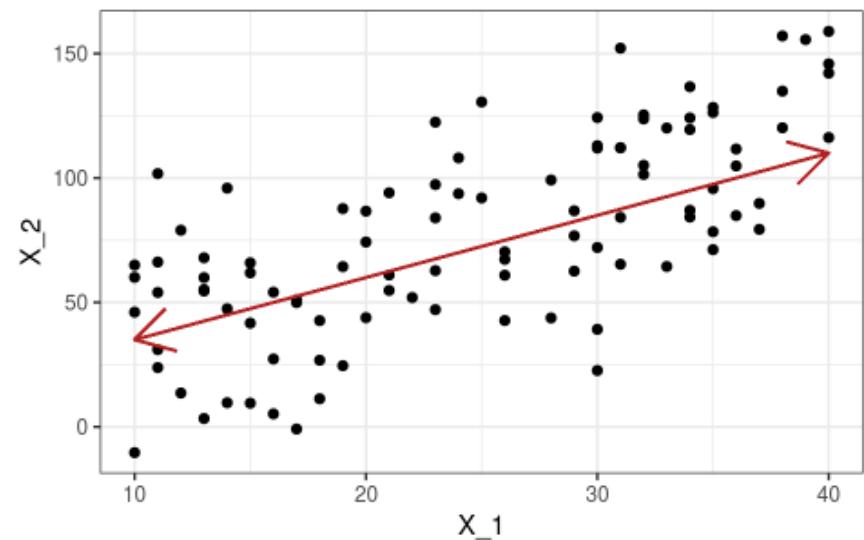
¿De donde sale la matriz de CP?

En el método PCA, cada una de las componentes se corresponde con un *eigenvector*, y el orden de componente se establece por orden decreciente de *eigenvalue*. Así pues, la primera componente es el *eigenvector* con el *eigenvalue* asociado más alto.

¿Cómo lo hace el ACP?

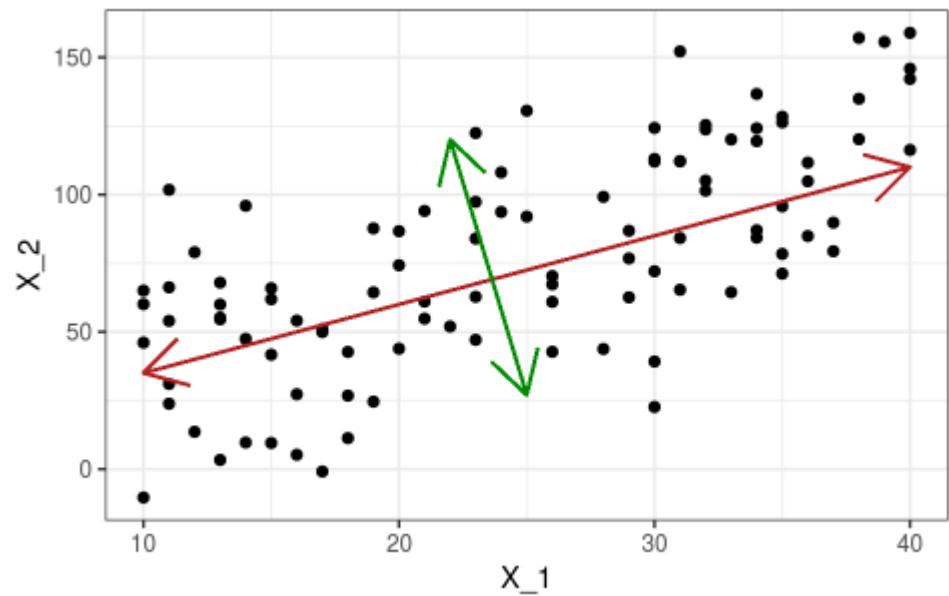
En el método PCA, cada una de las componentes se corresponde con un eigenvector, y el orden de componente se establece por orden decreciente

de eigenvalue. Así pues, la primera componente es el eigenvector con el eigenvalue asociado más alto.

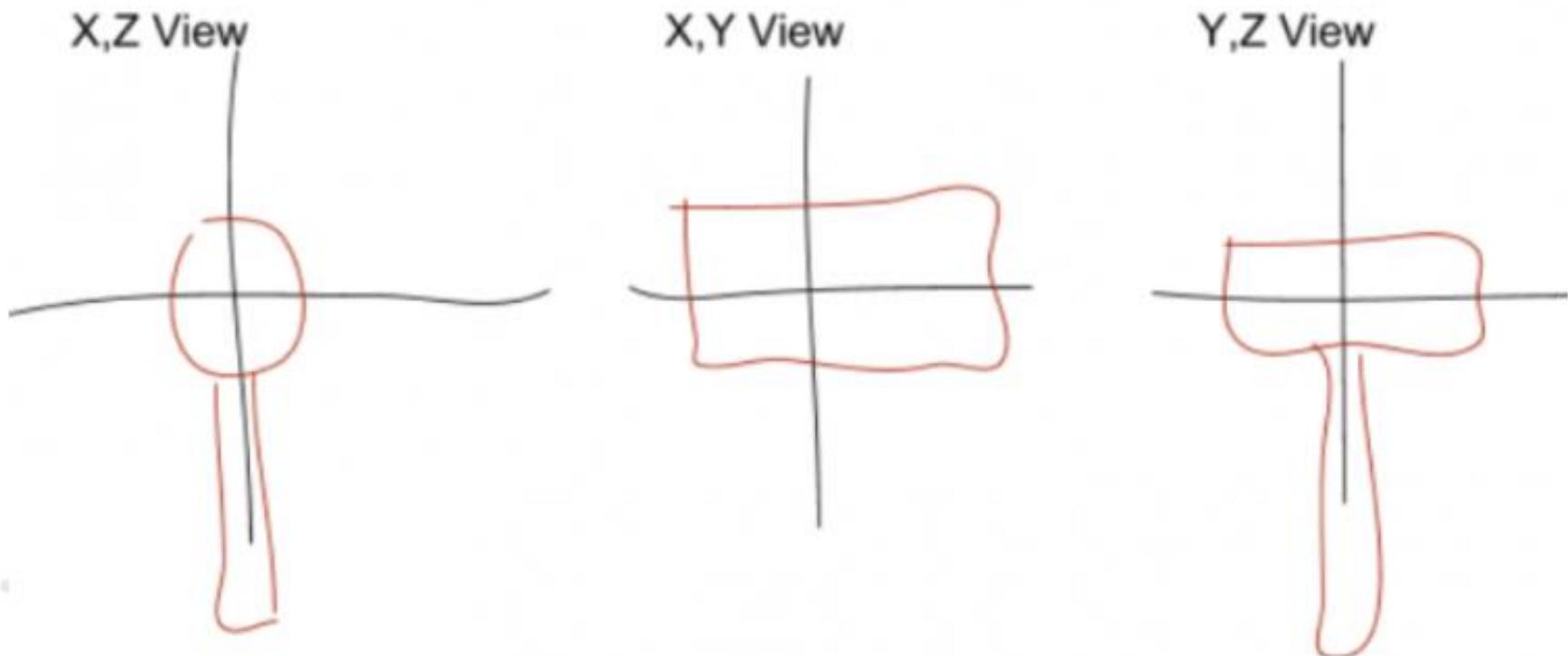


¿Cómo lo hace el ACP?

La segunda componente (Z_2) sigue la segunda dirección en la que los datos muestran mayor varianza y que no está correlacionada con la primera componente. La condición de no correlación entre componentes principales equivale a decir que sus direcciones son perpendiculares/ortogonales



ACP a partir de un mazo



Escogencia de las CP

DATOS

	Matemáticas	Ciencias	Español	Historia	EdFísica
Lucía	7.0	6.5	9.2	8.6	8.0
Pedro	7.5	9.4	7.3	7.0	7.0
Inés	7.6	9.2	8.0	8.0	7.5
Luis	5.0	6.5	6.5	7.0	9.0
Andrés	6.0	6.0	7.8	8.9	7.3
Ana	7.8	9.6	7.7	8.0	6.5
Carlos	6.3	6.4	8.2	9.0	7.2
José	7.9	9.7	7.5	8.0	6.0
Sonia	6.0	6.0	6.5	5.5	8.7
María	6.8	7.2	8.7	9.0	7.0

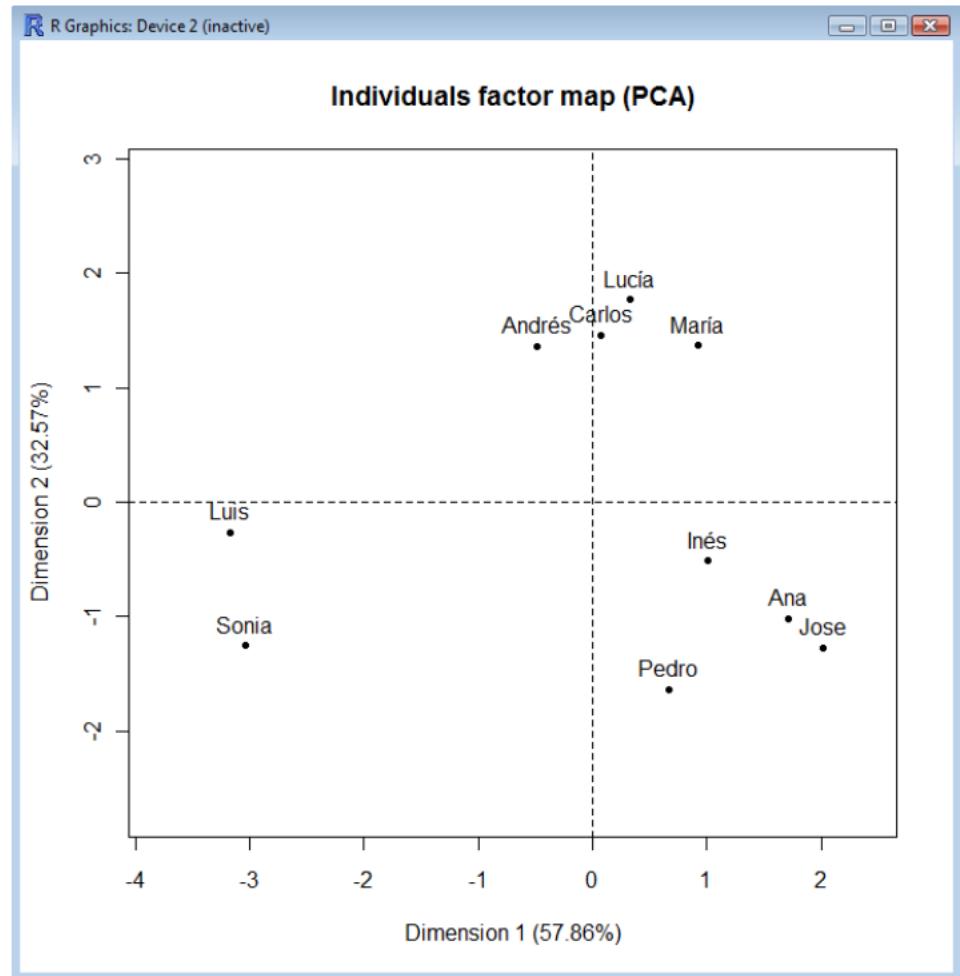
COMPONENTES

	Comp1	Comp2			
Lucia	0,3231	1,7725			
Pedro	0,6654	-1,6387			
Ines	1,0025	-0,5157			
Luis	-3,1721	-0,2628			
Andres	-0,4889	1,3654			
Ana	1,7086	-1,0217			
Carlos	0,0676	1,4623			
Jose	2,0119	-1,2759			
Sonia	-3,042	-1,2549			
Maria	0,9239	1,3694			

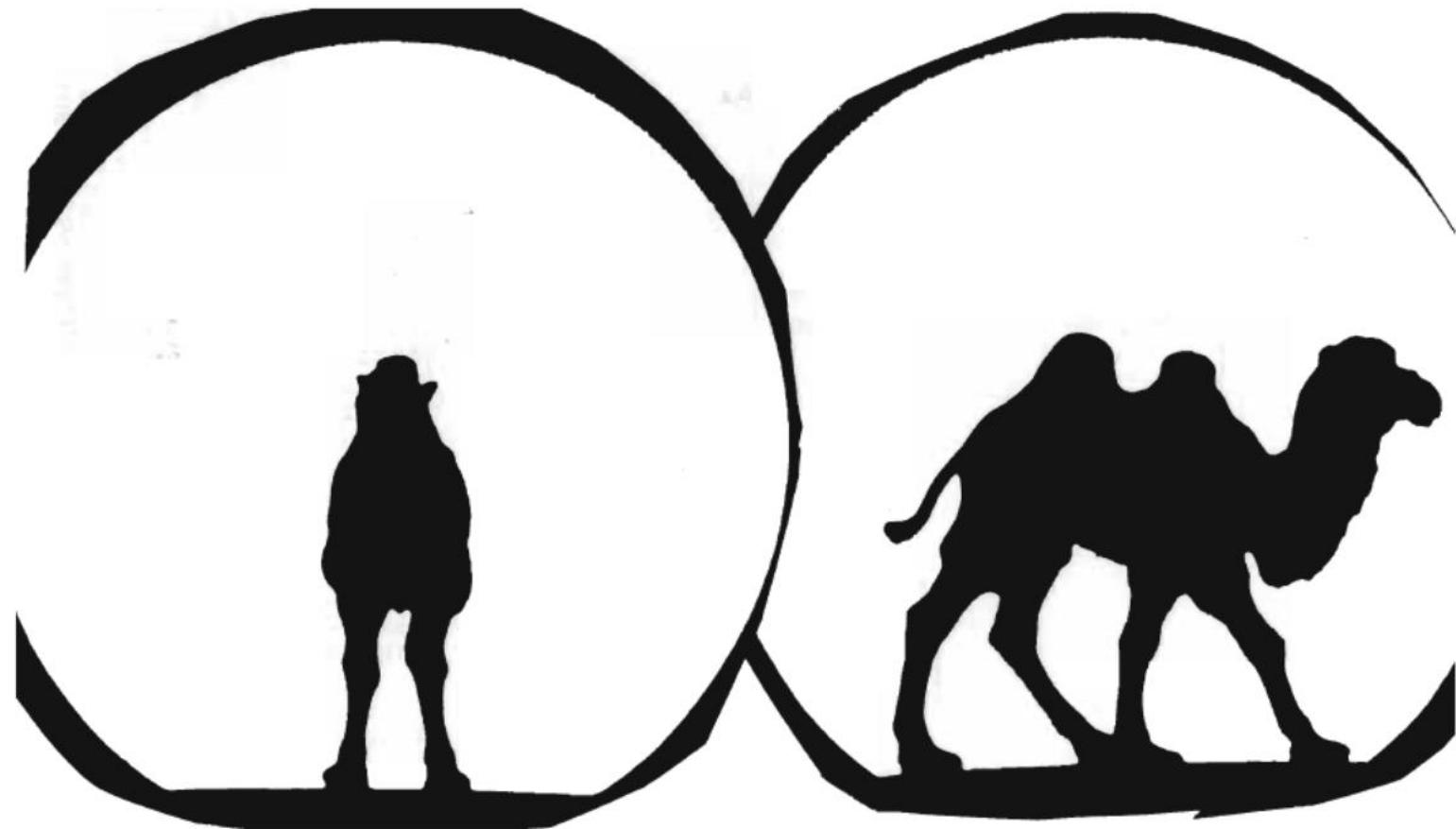
Plano Principal

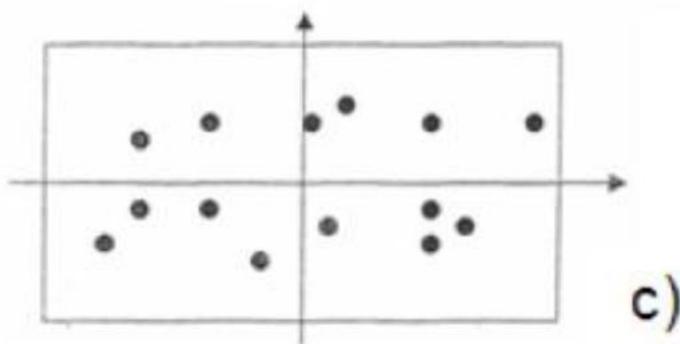
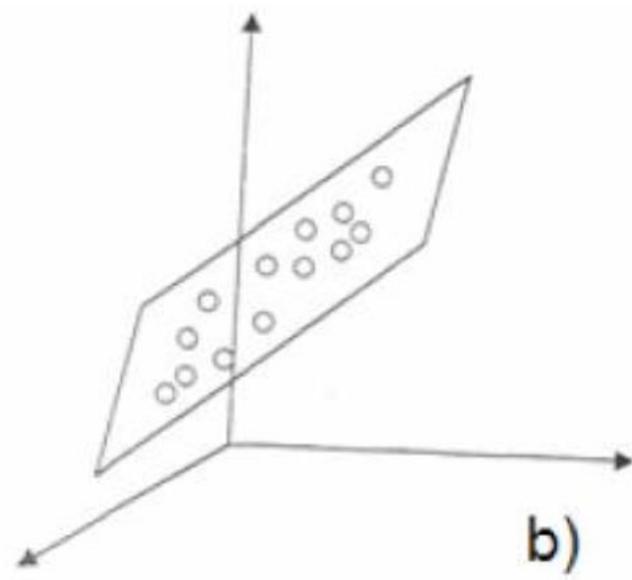
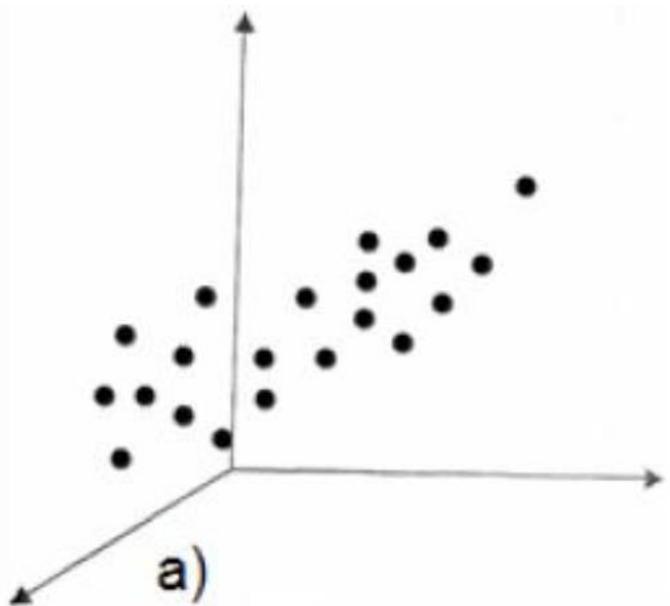
COMPONENTES

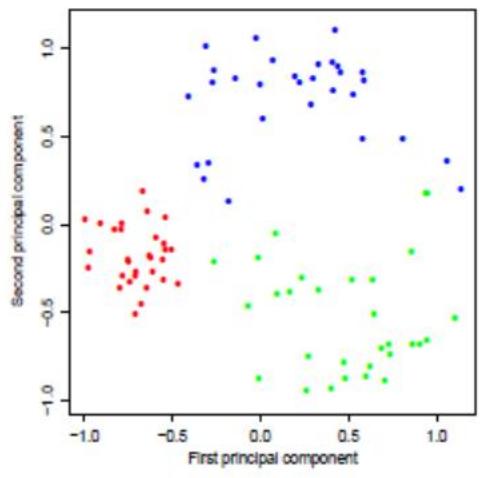
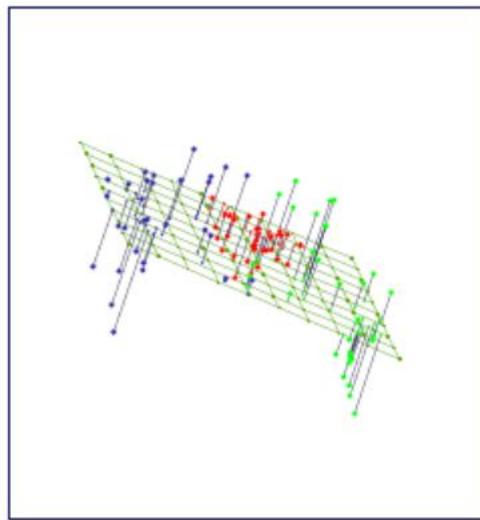
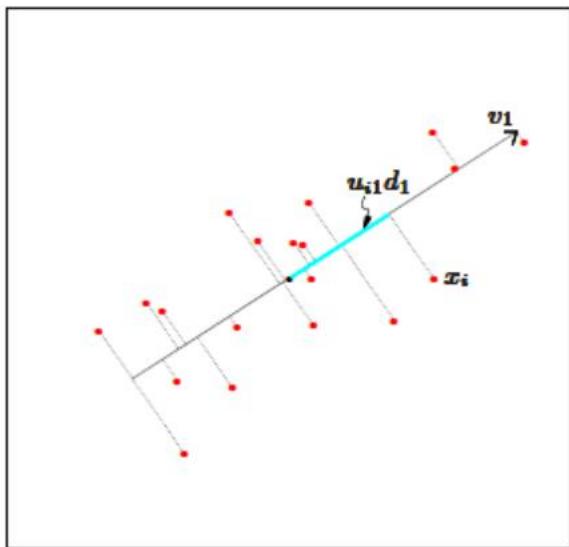
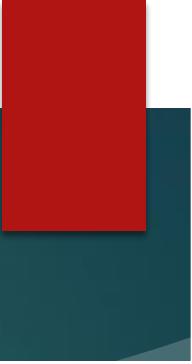
	Comp1	Comp2			
Lucia	0,3231	1,7725			
Pedro	0,6654	-1,6387			
Ines	1,0025	-0,5157			
Luis	-3,1721	-0,2628			
Andres	-0,4889	1,3654			
Ana	1,7086	-1,0217			
Carlos	0,0676	1,4623			
Jose	2,0119	-1,2759			
Sonia	-3,042	-1,2549			
Maria	0,9239	1,3694			



Objetivo: Encontrar el mejor plano (subespacio) para ver la nube de puntos.







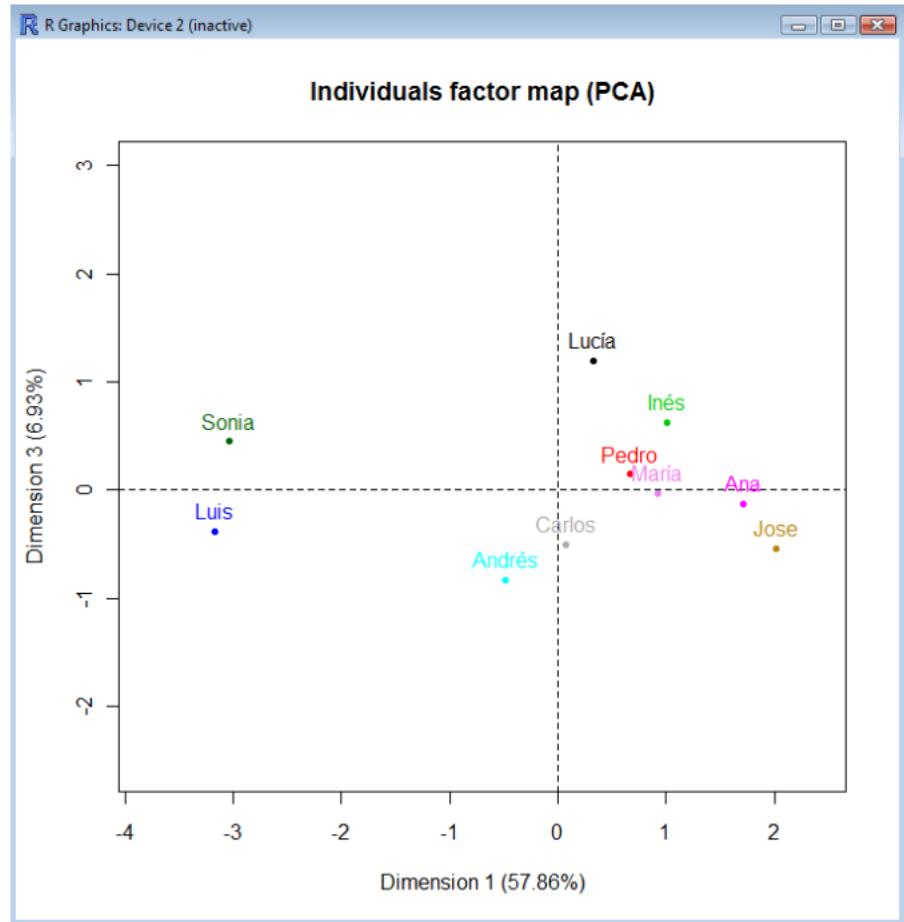
DATOS**COMPONENTES**

	Matemáticas	Ciencias	Español	Historia	EdFísica
Lucía	7.0	6.5	9.2	8.6	8.0
Pedro	7.5	9.4	7.3	7.0	7.0
Inés	7.6	9.2	8.0	8.0	7.5
Luis	5.0	6.5	6.5	7.0	9.0
Andrés	6.0	6.0	7.8	8.9	7.3
Ana	7.8	9.6	7.7	8.0	6.5
Carlos	6.3	6.4	8.2	9.0	7.2
José	7.9	9.7	7.5	8.0	6.0
Sonia	6.0	6.0	6.5	5.5	8.7
María	6.8	7.2	8.7	9.0	7.0

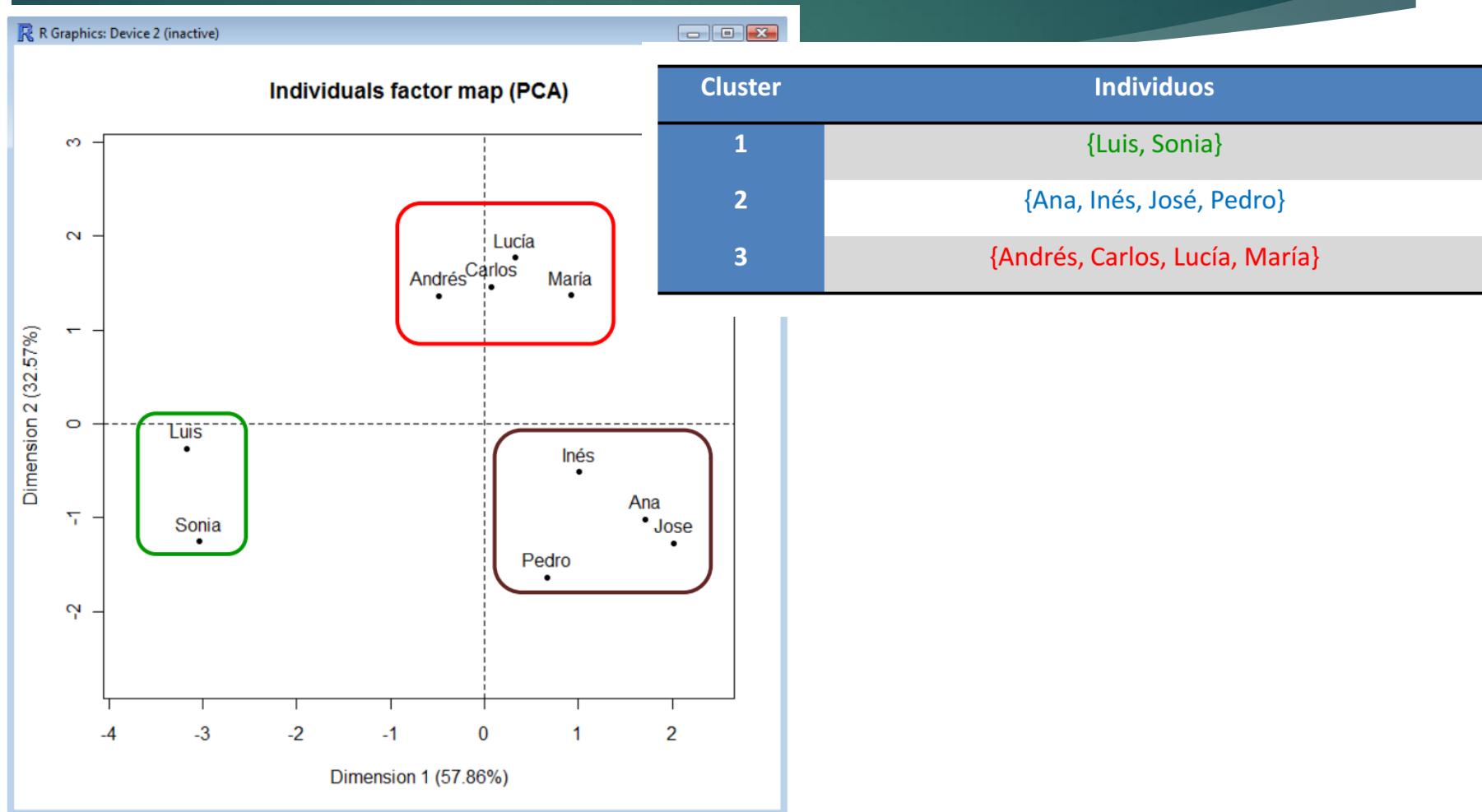
	Comp1	Comp3		
Lucia	0,3231	1,1988		
Pedro	0,6654	0,1455		
Ines	1,0025	0,6289		
Luis	-3,1721	-0,382		
Andres	-0,4889	-0,8352		
Ana	1,7086	-0,1271		
Carlos	0,0676	-0,5062		
Jose	2,0119	-0,5422		
Sonia	-3,042	0,4488		
Maria	0,9239	-0,0293		

COMPONENTES

	Comp1		Comp3		
Lucia	0,3231		1,1988		
Pedro	0,6654		0,1455		
Ines	1,0025		0,6289		
Luis	-3,1721		-0,382		
Andres	-0,4889		-0,8352		
Ana	1,7086		-0,1271		
Carlos	0,0676		-0,5062		
Jose	2,0119		-0,5422		
Sonia	-3,042		0,4488		
Maria	0,9239		-0,0293		



Análisis de clústeres o conglomerados



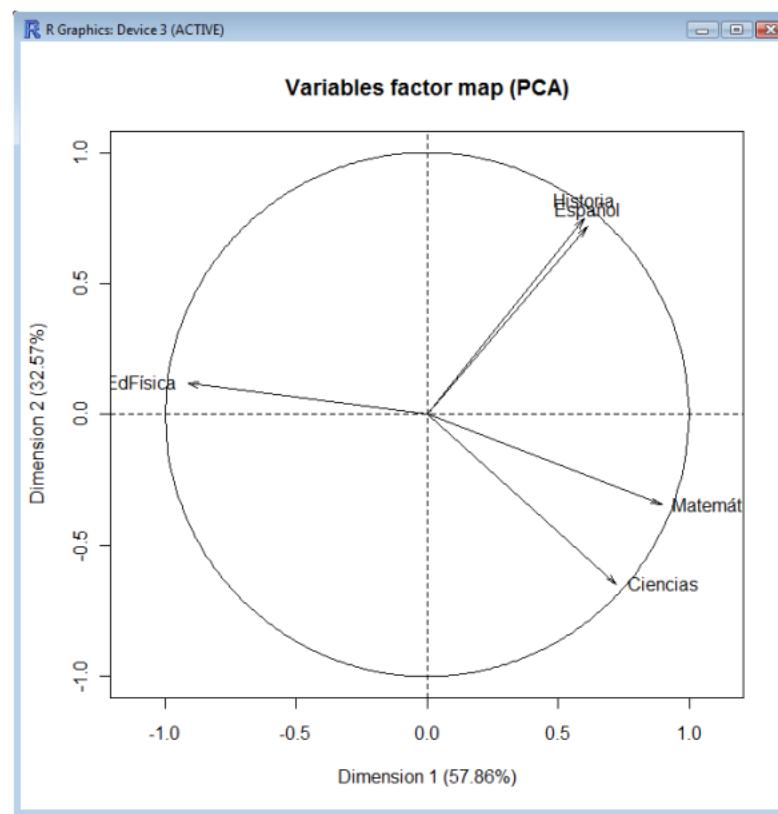
Análisis de las variables

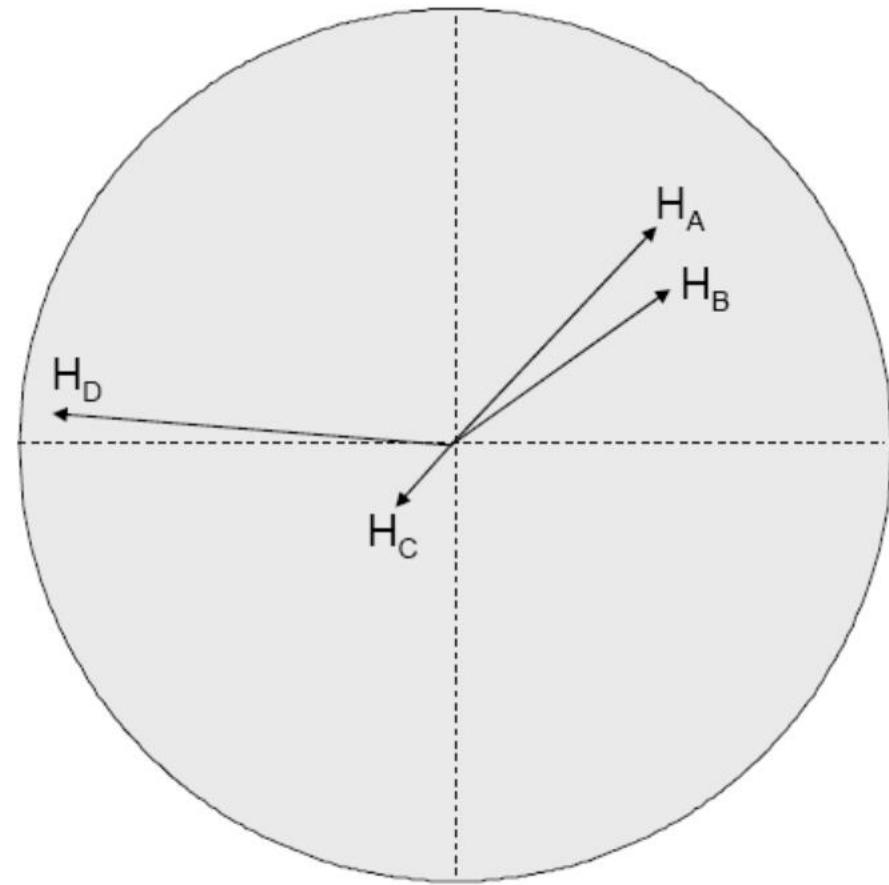
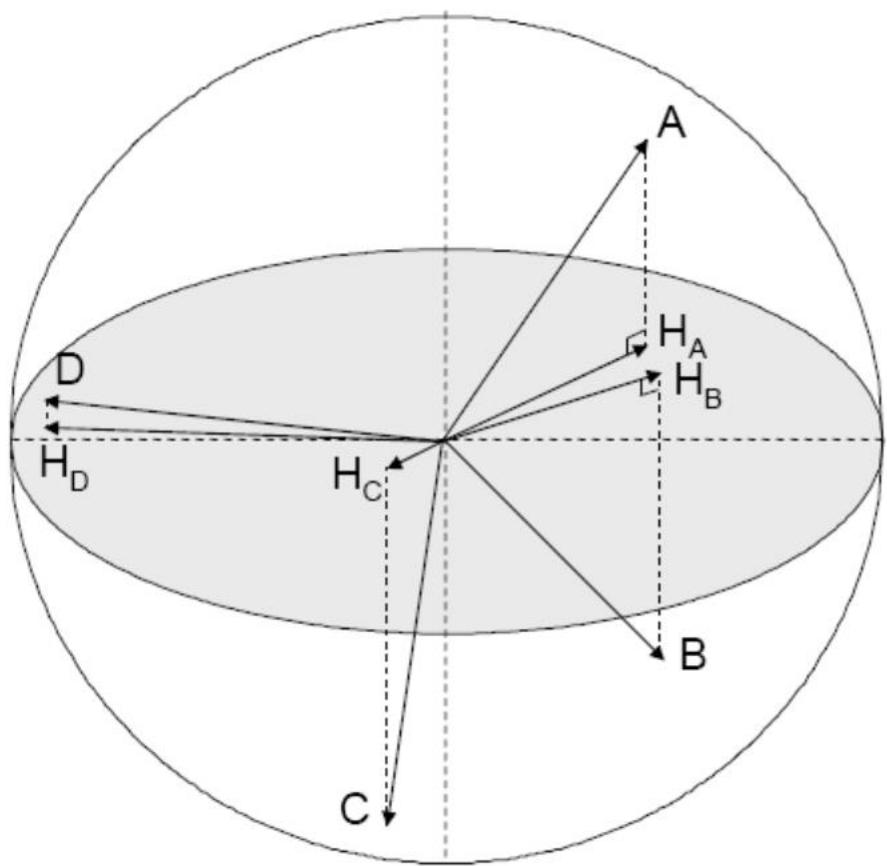
	Matemáticas	Ciencias	Español	Historia	EdFísica
Lucía	7.0	6.5	9.2	8.6	8.0
Pedro	7.5	9.4	7.3	7.0	7.0
Inés	7.6	9.2	8.0	8.0	7.5
Luis	5.0	6.5	6.5	7.0	9.0
Andrés	6.0	6.0	7.8	8.9	7.3
Ana	7.8	9.6	7.7	8.0	6.5
Carlos	6.3	6.4	8.2	9.0	7.2
José	7.9	9.7	7.5	8.0	6.0
Sonia	6.0	6.0	6.5	5.5	8.7
María	6.8	7.2	8.7	9.0	7.0

	Comp1	Comp2	Comp3	Comp4	Comp5
Matemáticas	0.8957980	-0.3452036	0.25797931	-0.09146818	0.05882803
Ciencias	0.7227976	-0.6483946	0.02384033	0.23587773	-0.03068234
Español	0.6108931	0.7173206	0.33102532	-0.02454152	-0.04561456
Historia	0.5999227	0.7484701	-0.23206345	0.15639747	0.03964443
EdFísica	-0.9139265	0.1196373	0.34065108	0.18315368	0.02892890

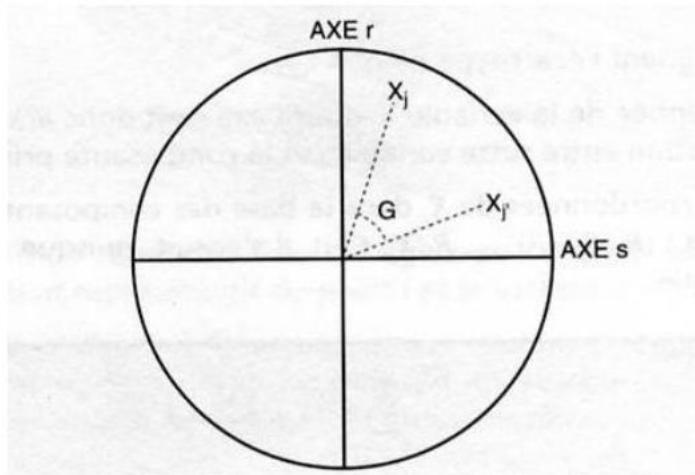
	Matemáticas	Ciencias	Español	Historia	EdFísica
Lucía	7.0	6.5	9.2	8.6	8.0
Pedro	7.5	9.4	7.3	7.0	7.0
Inés	7.6	9.2	8.0	8.0	7.5
Luis	5.0	6.5	6.5	7.0	9.0
Andrés	6.0	6.0	7.8	8.9	7.3
Ana	7.8	9.6	7.7	8.0	6.5
Carlos	6.3	6.4	8.2	9.0	7.2
José	7.9	9.7	7.5	8.0	6.0
Sonia	6.0	6.0	6.5	5.5	8.7
María	6.8	7.2	8.7	9.0	7.0

	Comp1	Comp2			
Matemáticas	0.8957980	-0.3452036			
Ciencias	0.7227976	-0.6483946			
Español	0.6108931	0.7173206			
Historia	0.5999227	0.7484701			
EdFísica	-0.9139265	0.1196373			





INTERPRETACIÓN

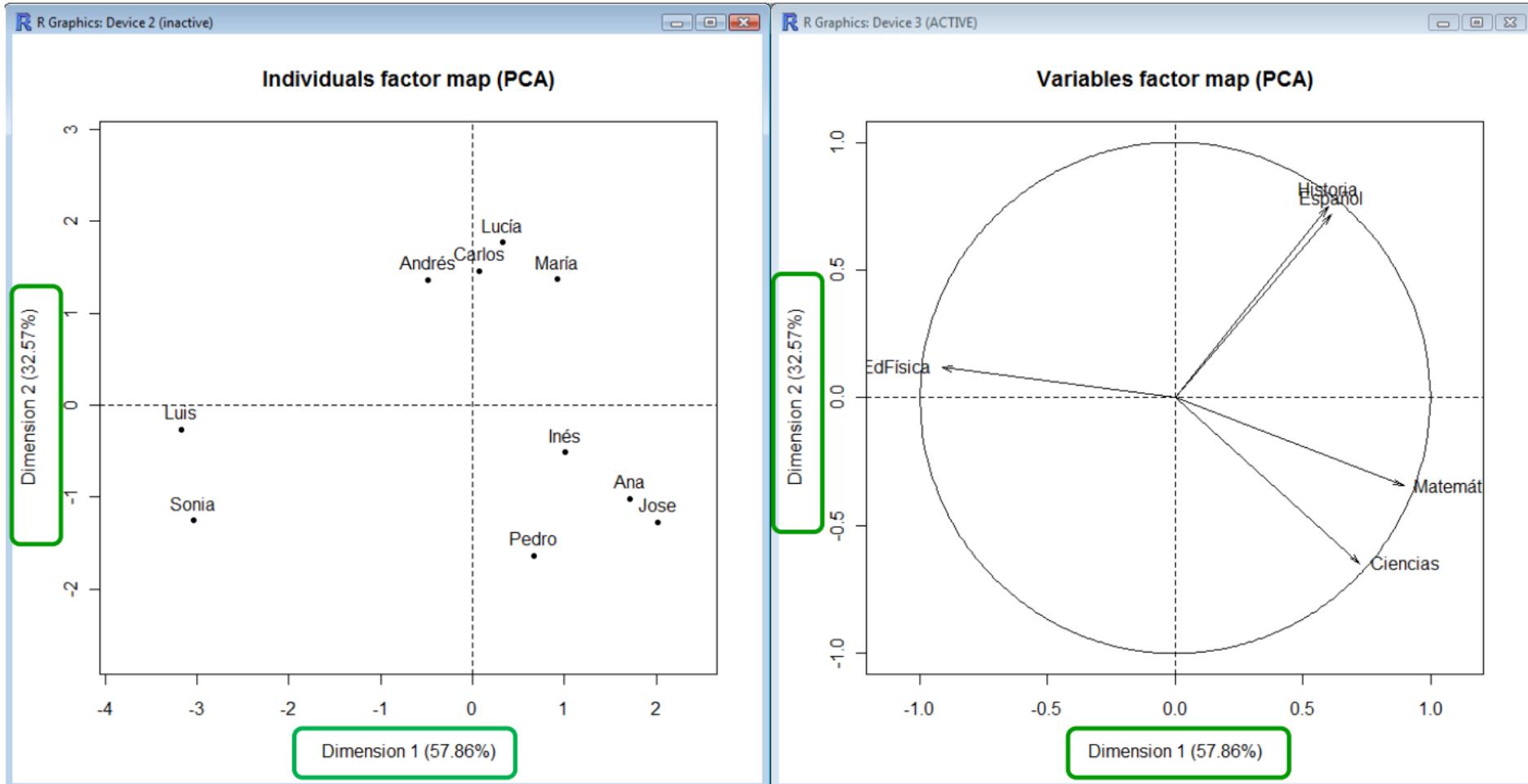


- Si X^j y $X^{j'}$ están cercanas entre si, entonces X^j y $X^{j'}$ son fuerte y positivamente correlacionadas.
- Si el ángulo entre X^j y $X^{j'}$ es cercano a los 90° entonces NO existe ninguna correlación entre ambas variables.
- Si X^j y $X^{j'}$ están opuestas al vértice (origen) entonces existe una correlación fuerte y negativa entre X^j y $X^{j'}$.

Tabla de correlaciones presentes en el círculo

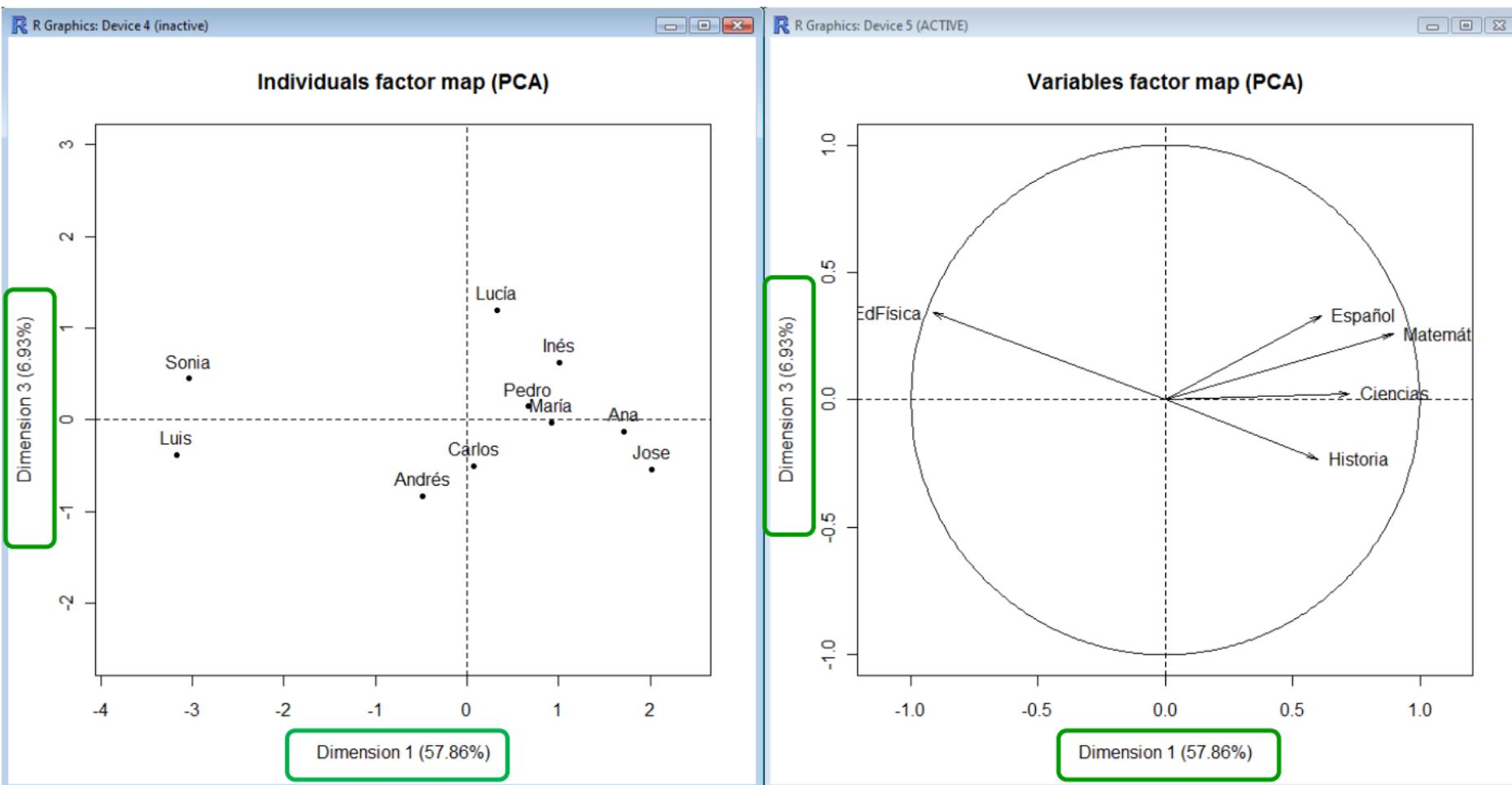
Variable	Matemática	Ciencias	Español	Historia	Ed. Física
Matemática	-	Positive	None	None	Negative
Ciencias		-	None	None	Negative
Español			-	Positive	None
Historia				-	None
Ed. Física					-

Calidad de los gráficos: inercia



$$\text{Inercia} = 57.86 + 32.57 = 90.43$$

Inercia Explicada = 64.79%



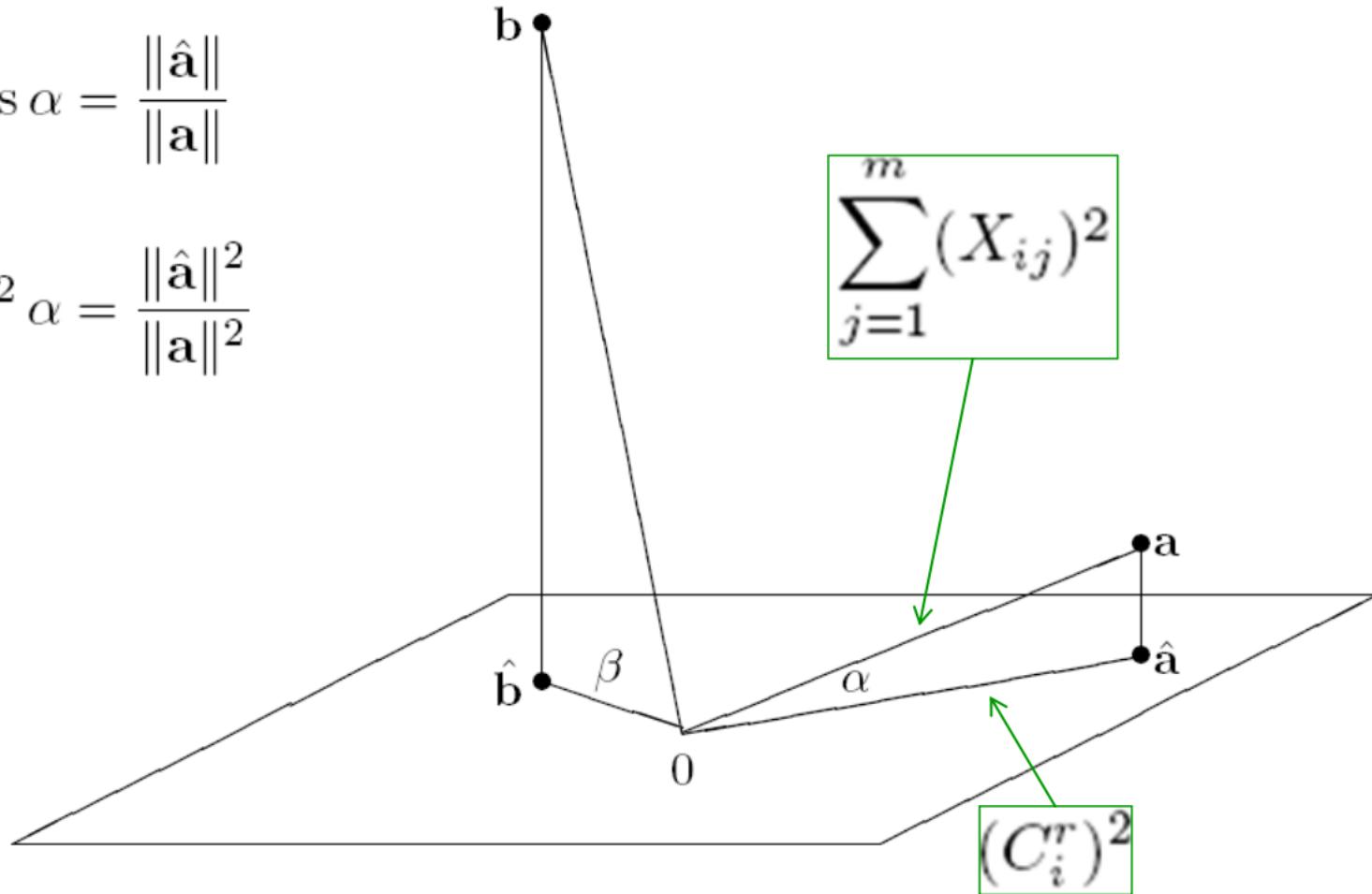
$$\text{Inercia} = 57.86 + 6.93 = 64.79$$

Calidad de la representación de los individuos y las variables

- ▶ Solamente los individuos y variables bien representados en el plano principal y en el círculo de correlaciones, respectivamente, pueden ser interpretados.
- ▶ ¿Cómo saber si un individuo o variable está bien representada?

$$\cos \alpha = \frac{\|\hat{\mathbf{a}}\|}{\|\mathbf{a}\|}$$

$$\cos^2 \alpha = \frac{\|\hat{\mathbf{a}}\|^2}{\|\mathbf{a}\|^2}$$



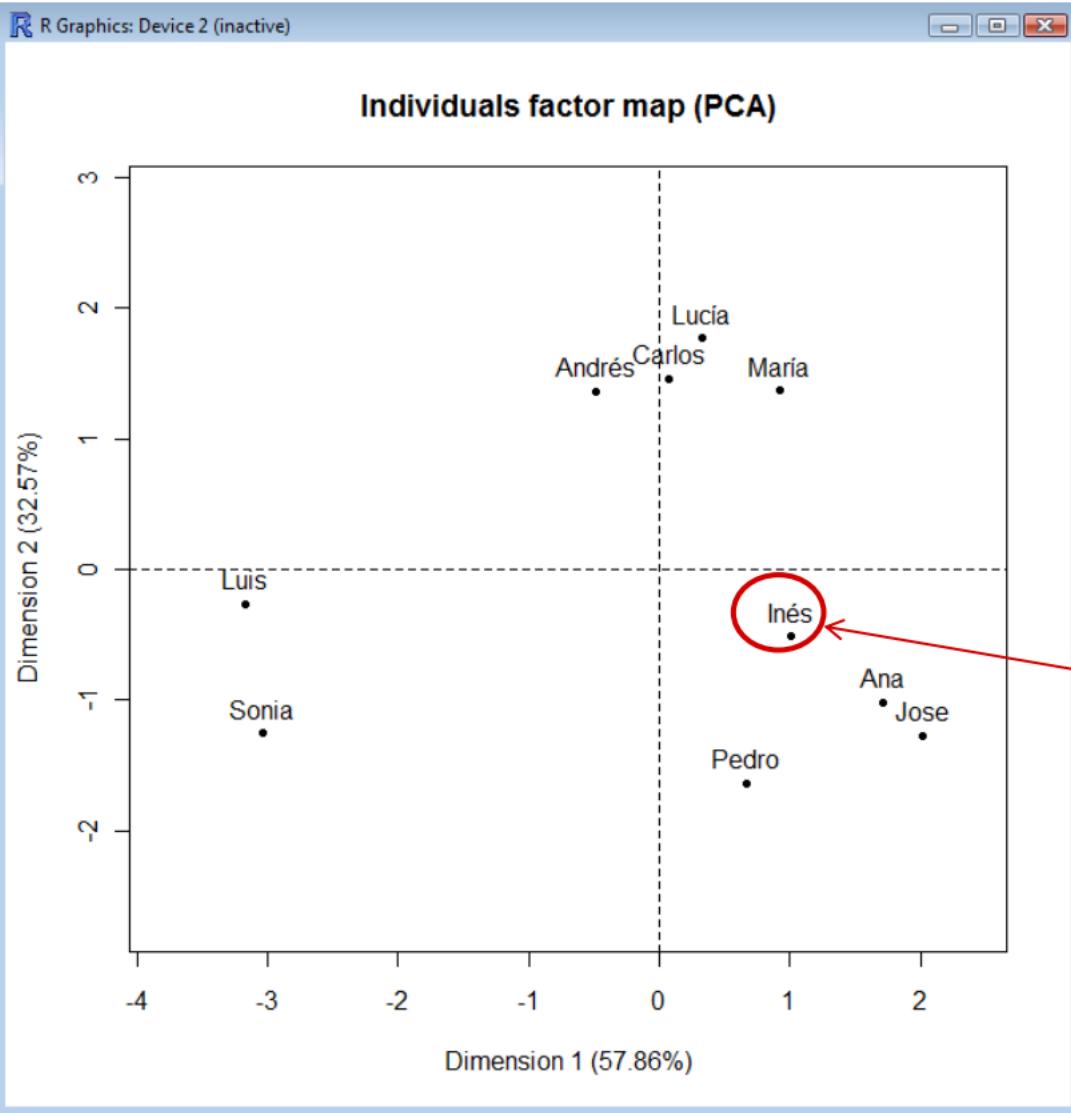
NOTA: Si el $\cos^2 \alpha$ es “cercano” a 1 la representación del individuo será muy buena o si multiplicado por 100 es mayor a 60 la representación es buena.

Ejemplo $\text{Cos}^2(\alpha)$: Estudiantes

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
Luía	0.022270827	0.670420670	0.3066598387	0.0006458478	2,82E+00
Pedro	0.139905502	0.848430539	0.0066865270	0.0001680781	4,81E+03
Inés	0.514468899	0.136122895	0.2024397137	0.1365196756	1,04E+04
Luis	0.936851990	0.006429392	0.0135836048	0.0427712757	3,64E+02
Andrés	0.084139511	0.656353715	0.2456037030	0.0085448999	5,36E+03
Ana	0.732686110	0.261979570	0.0040527949	0.0011209894	1,61E+02
Carlos	0.001892733	0.886081139	0.1061921889	0.0057625700	7,14E+01
Jose	0.673612108	0.270910359	0.0489165039	0.0065104446	5,06E+01
Sonia	0.808829929	0.137636943	0.0176072367	0.0358004434	1,25E+02
María	0.308554271	0.677869212	0.0003109770	0.0018464085	1,14E+04

Ejemplo $\cos^2(\alpha)$: Estudiantes

	Dim.1	Dim.2	SUMA*100
Luía	0,022270827	0,67042067	69,27
Pedro	0,139905502	0,848430539	98,83
Inés	0,514468899	0,136122895	65,06
Luis	0,93685199	0,006429392	94,33
Andrés	0,084139511	0,656353715	74,05
Ana	0,73268611	0,26197957	99,47
Carlos	0,001892733	0,886081139	88,80
Jose	0,673612108	0,270910359	94,45
Sonia	0,808829929	0,137636943	94,65
María	0,308554271	0,677869212	98,64



Todos los individuos están bien representados en este plano.

Inés es la persona que tiene la menos buena representación en este plano.

Calidad de representación de cada variable

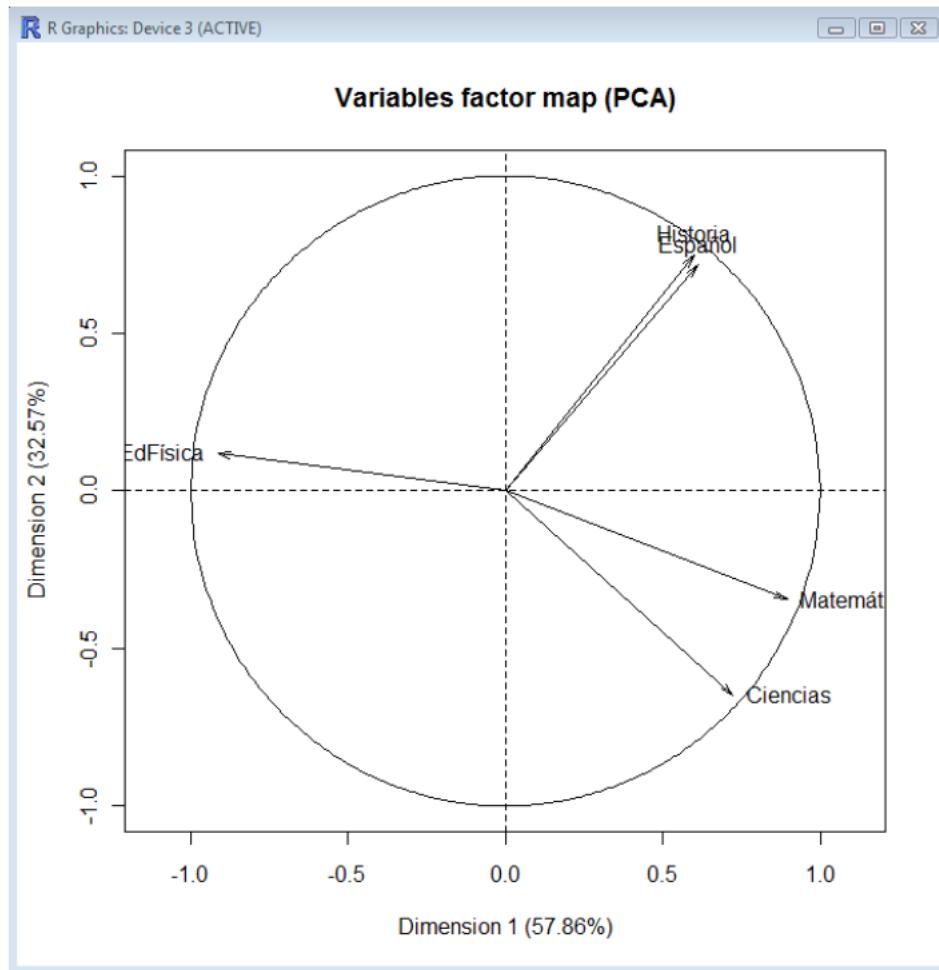
- ✓ La calidad de la representación de una variable sobre el círculo de correlaciones, será también medida con el **coseno al cuadrado** del ángulo la variable y su proyección. Ahora bien, recuérdese que entre variables, el coseno es igual a una correlación, por lo que serán las **correlaciones cercanas a 1** las que impliquen la calidad de la representación de las variables.
- ✓ Es decir estarán bien representadas aquellas variables que queden ubicadas cerca de la frontera o borde del círculo de correlaciones

Ejemplo $\text{Cos}^2(\alpha)$: Estudiantes

	Dim1	Dim2	Dim3	Dim4	Dim5
Matemáticas	0,802454	0,1191655	0,066553324	0,008366429	0,003460737
Ciencias	0,5224364	0,42041555	0,000568361	0,055638305	0,000941406
Español	0,3731904	0,51454884	0,109577763	0,000602286	0,002080688
Historia	0,3599073	0,56020745	0,053853443	0,02446017	0,001571681
EdFísica	0,8352616	0,01431309	0,116043157	0,03354527	0,000836881

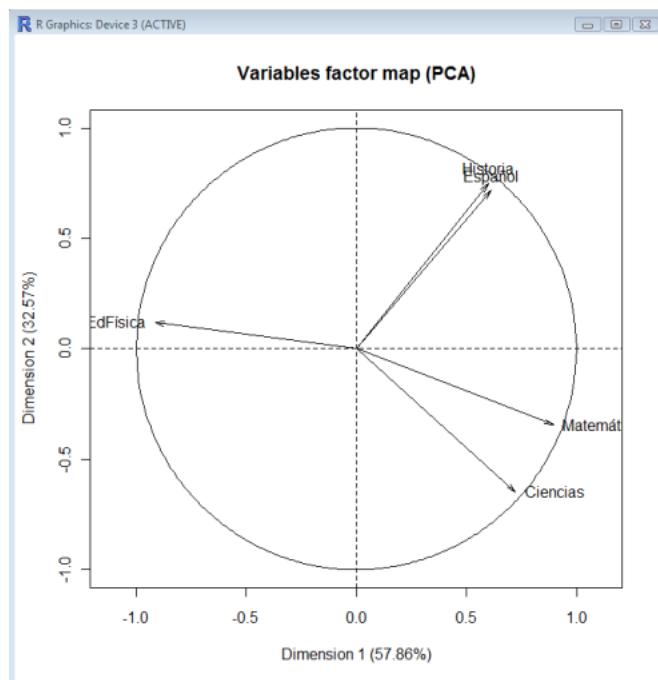
	Dim1	Dim2	Suma
Matemáticas	0,802454	0,1191655	0,9216195
Ciencias	0,5224364	0,42041555	0,94285195
Español	0,3731904	0,51454884	0,88773924
Historia	0,3599073	0,56020745	0,92011475
EdFísica	0,8352616	0,01431309	0,84957469

Ejemplo: Estudiantes



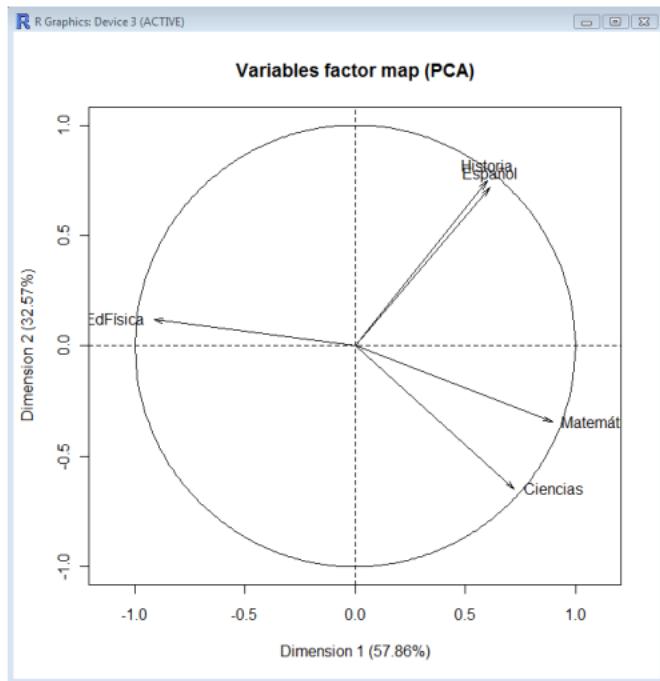
Todas las variables están bien representadas en este círculo de correlaciones

Ejemplo: Estudiantes



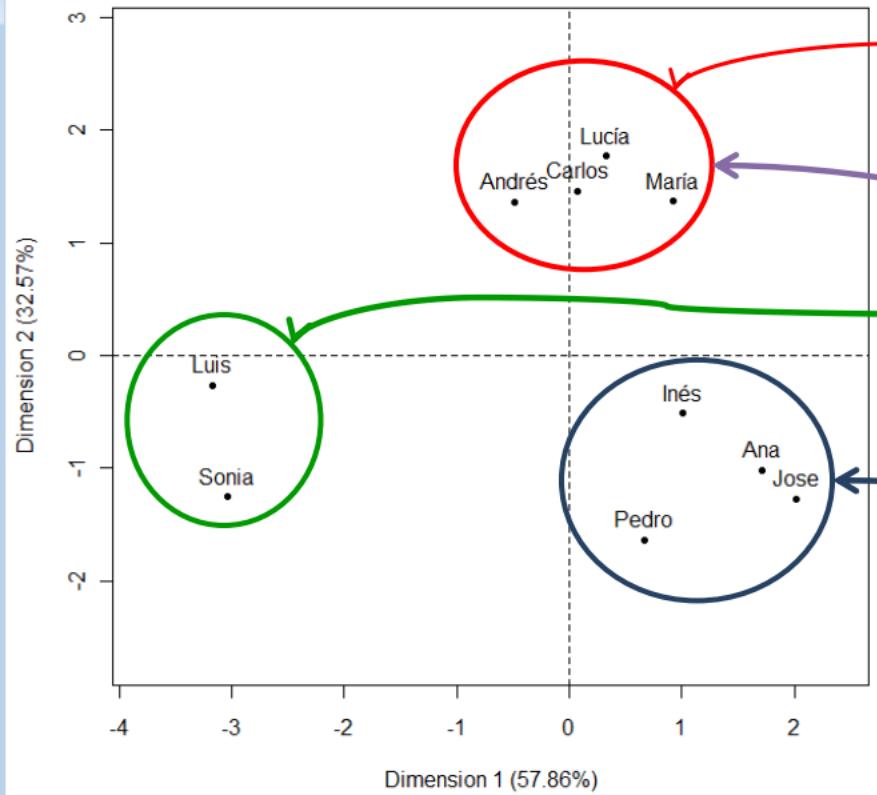
Un ángulo de apenas $1,71^\circ$ representa una clara correlación **positiva** entre las variables Español e Historia, es decir, que las notas en estas materias se comportan de manera similar en los estudiantes.

Ejemplo: Estudiantes

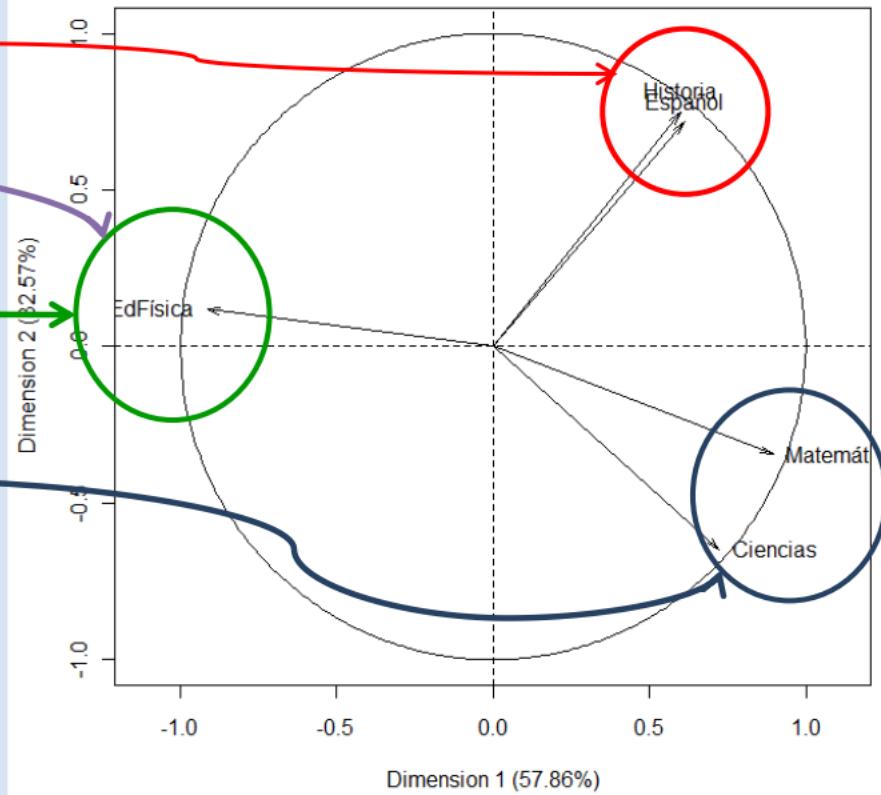


Ángulos cercanos a 180° , como este caso de $166,38^\circ$ entre Matemática y Ed. Física denotan correlaciones **negativas**. Entonces al crecer variable Matemáticas, la variable Educación Física va a disminuir y viceversa.

Individuals factor map (PCA)



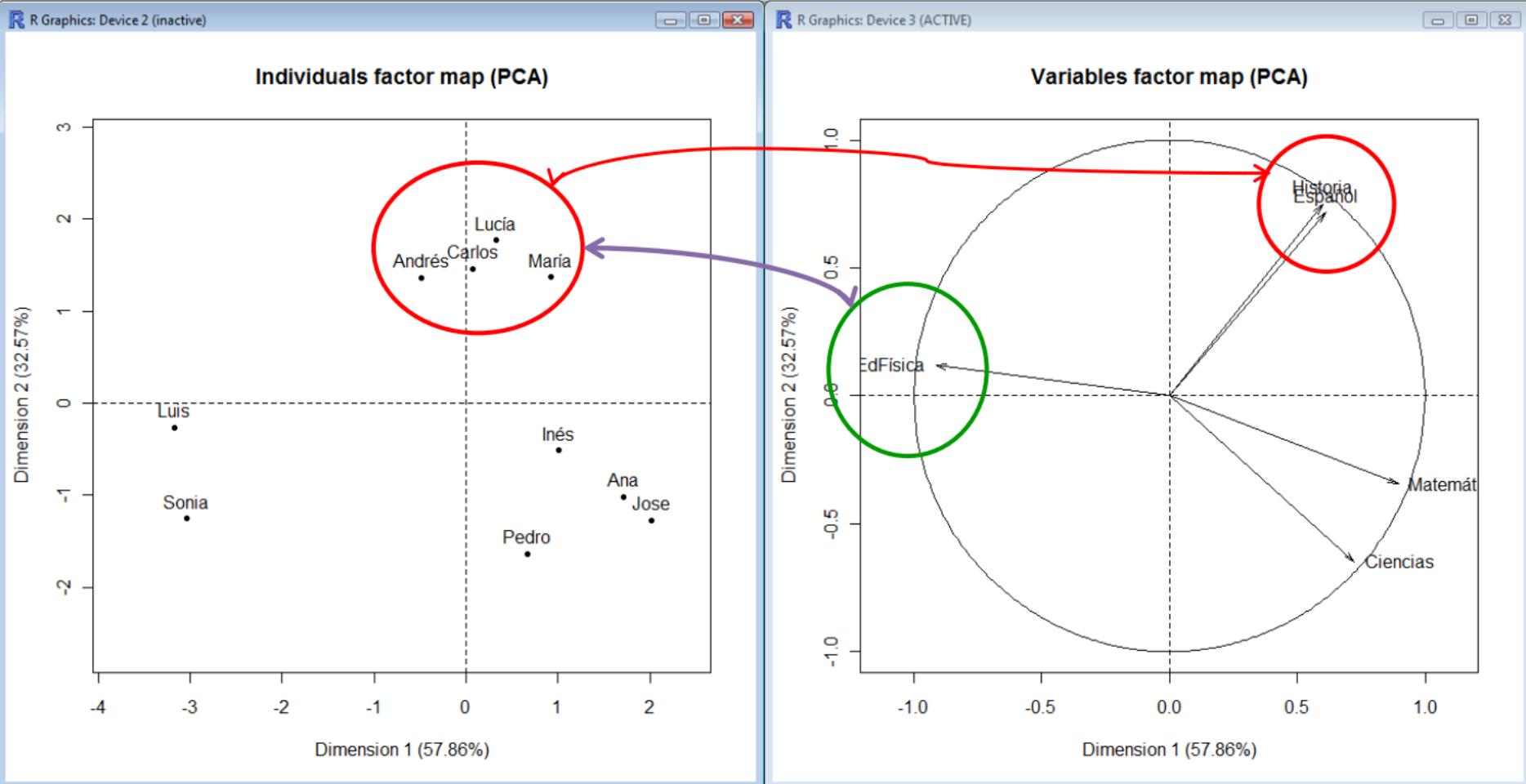
Variables factor map (PCA)



Sobreposición de Gráficos

- El clúster 1 (Luis y Sonia), se ve fuertemente impactado –de manera positiva– por las notas de educación física, es decir, son buenos deportistas. Recordando las correlaciones entre variables, podemos afirmar que son malos en ciencias y matemáticas, pues son variables inversamente correlacionadas con respecto a educación física. También se puede apreciar que son malos en español e historia, pues se oponen negativamente a estas variables en el plano.
- El clúster 2 (Ana, Inés, José y Pedro) parece ser el opuesto del clúster 1. Sus integrantes se destacan en las ciencias y las matemáticas, sin embargo son malos en deportes. Con respecto a las materias de español e historia, por estar de forma perpendicular o correlación nula, no es correcto asumir nada sobre su comportamiento.
- El clúster 3 (Andrés, Carlos, Lucía y María) se caracteriza por agrupar a los estudiantes destacados en el área de español e historia. Se puede inferir también que si bien no son excelentes en deportes, al menos no son los peores, pues se encuentran en la mitad del eje X.

Influencia de dos o más variables



Generalidades del Método

- ▶ La técnica que reviso es **PCA** y forma parte la teoría de “**Análisis de Variables Latentes**” o “**Análisis Factorial**”. Este tema requiere una entrada que explique los detalles teóricos, lo cual pondré posteriormente en la categoría [“Sobre Machine Learning”](#).
- ▶ La idea de **PCA** es reducir o proyectar nuestra información a un espacio de dimensión menor, pero también puede servir para construir un **índicador**

Generalidades del Método

- ▶ Una observación fundamental es que PCA es efectivo cuando la correlación entre variables es alta, es decir; la linealidad entre nuestras variables no es cero.
- ▶ El modelo parte de tener una matriz de datos de entrada X (filas y columnas) y se busca un modelo $Y=XT$ tal que los vectores de salida Y no estén correlacionados. Esto en álgebra lineal es hacer análisis del espectro de la matriz, que es el estudio de los vectores y valores propios

Generalidades del Método

- ▶ Una observación fundamental es que PCA es efectivo cuando la correlación entre variables es alta, es decir; la linealidad entre nuestras variables no es cero.
- ▶ El modelo parte de tener una matriz de datos de entrada X (filas y columnas) y se busca un modelo $Y=XT$ tal que los vectores de salida Y no estén correlacionados. Esto en álgebra lineal es hacer análisis del espectro de la matriz, que es el estudio de los vectores y valores propios