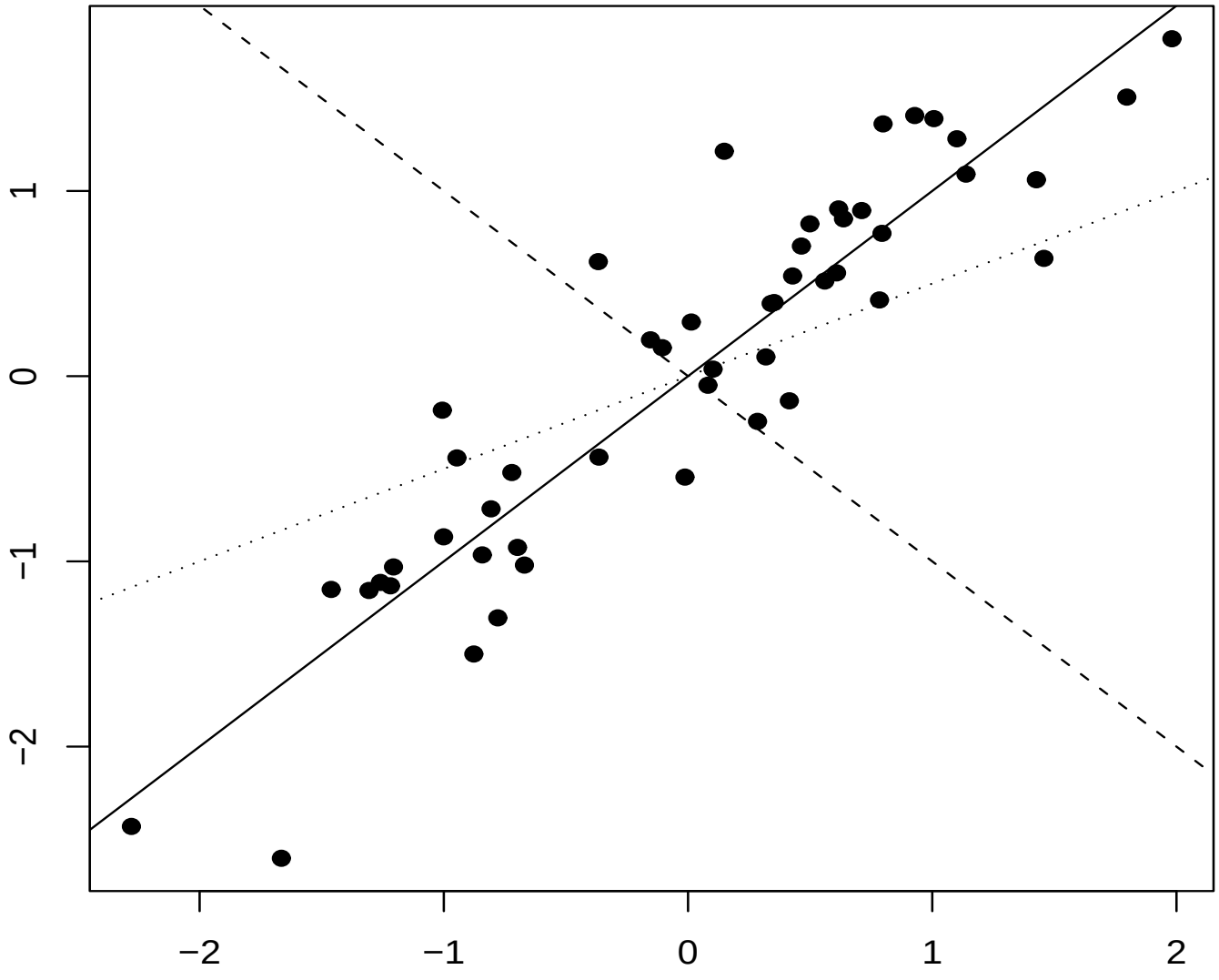
Tema 4
Técnicas de reducción de la dimensión

José R. Berrendero

Departamento de Matemáticas
Universidad Autónoma de Madrid

¿En qué dirección es conveniente proyectar?



Componentes principales de los datos anteriores

- Los coeficientes que definen las componentes principales son

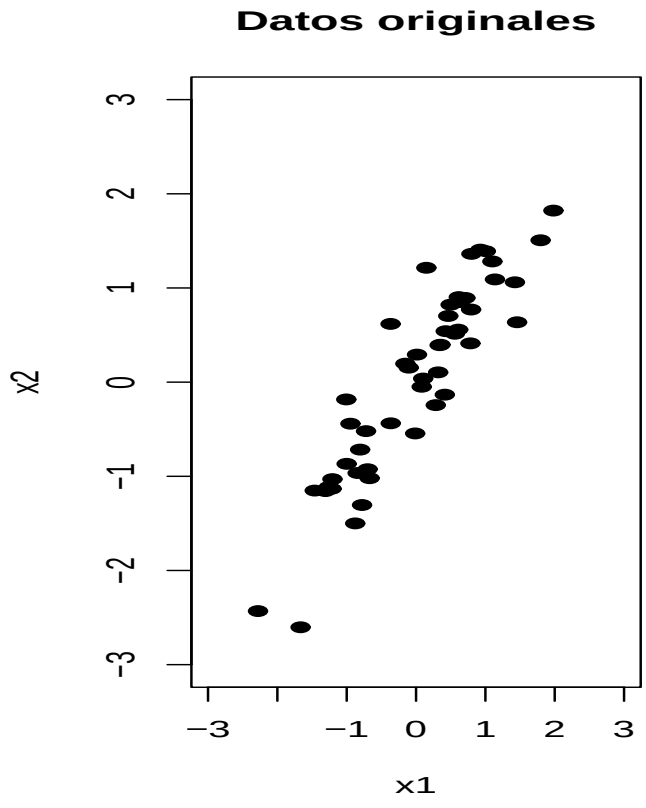
Primera componente	Segunda componente
-0.68	0.73
-0.73	-0.68

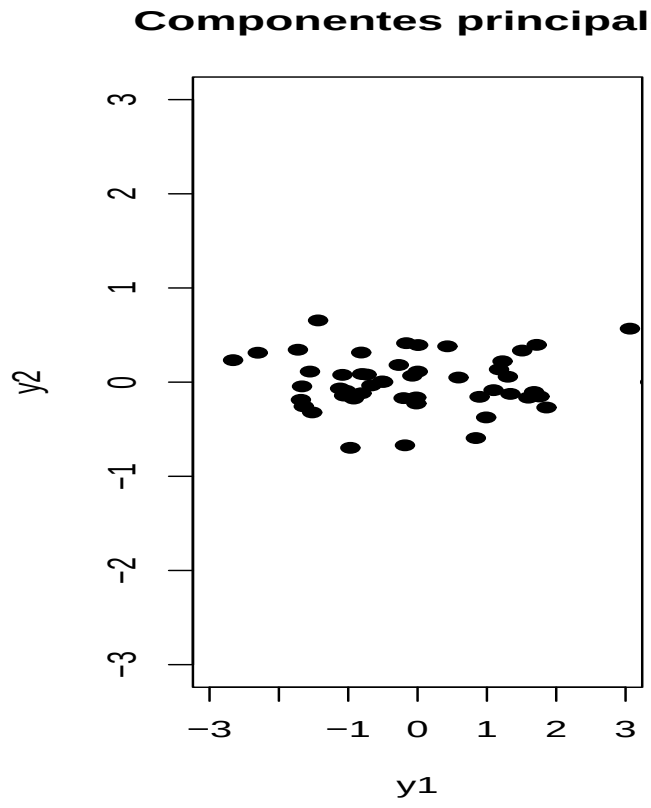
- El producto escalar es cero (son direcciones perpendiculares)
- Las dos componentes principales son

$$\begin{aligned} \mathbf{y}_{(1)} &= -0.68\,\text{VAR1} - 0.73\,\text{VAR2} \\ \mathbf{y}_{(2)} &= 0.73\,\text{VAR1} - 0.68\,\text{VAR2} \end{aligned}$$

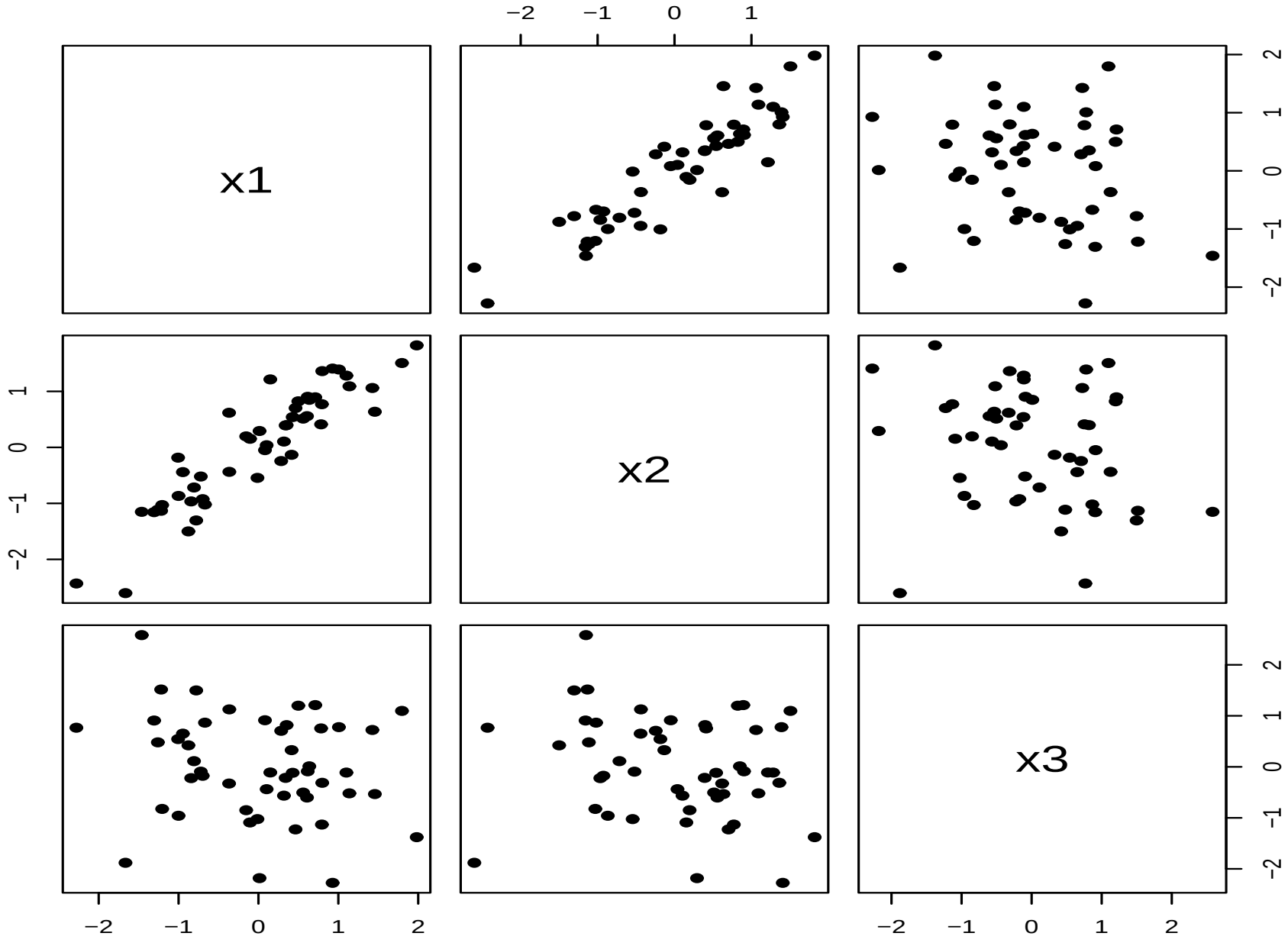
Variabilidad explicada por las componentes:

	CP1	CP2
Desviación típica	1.37	0.29
Porcentaje de varianza	95.8 %	4.2 %
Porcentaje acumulado	95.8 %	100 %





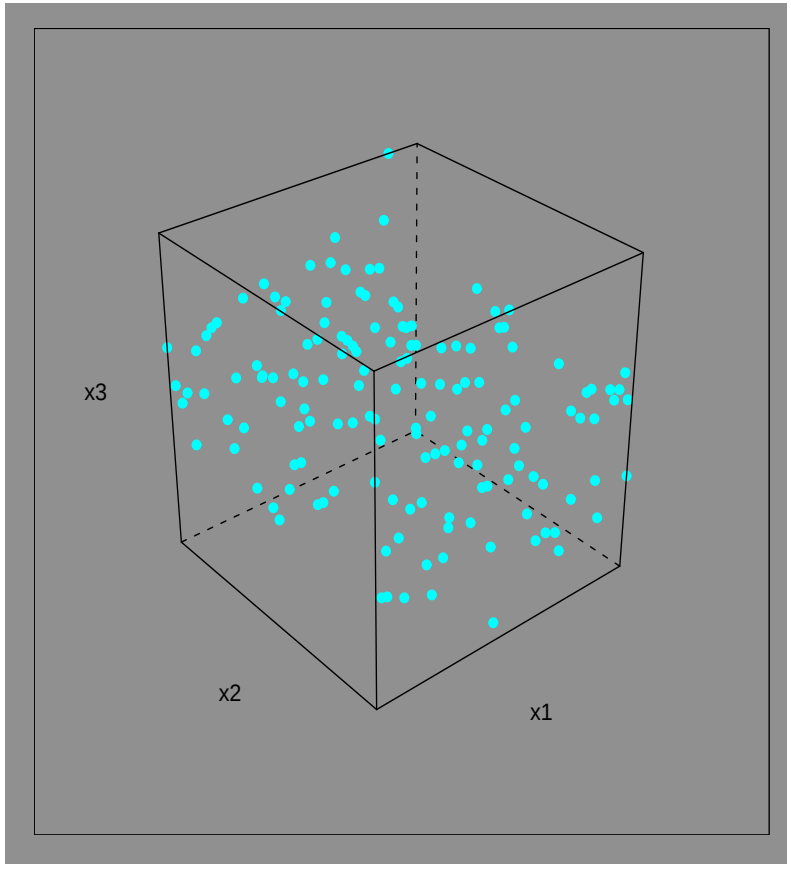
Un ejemplo con $p = 3$

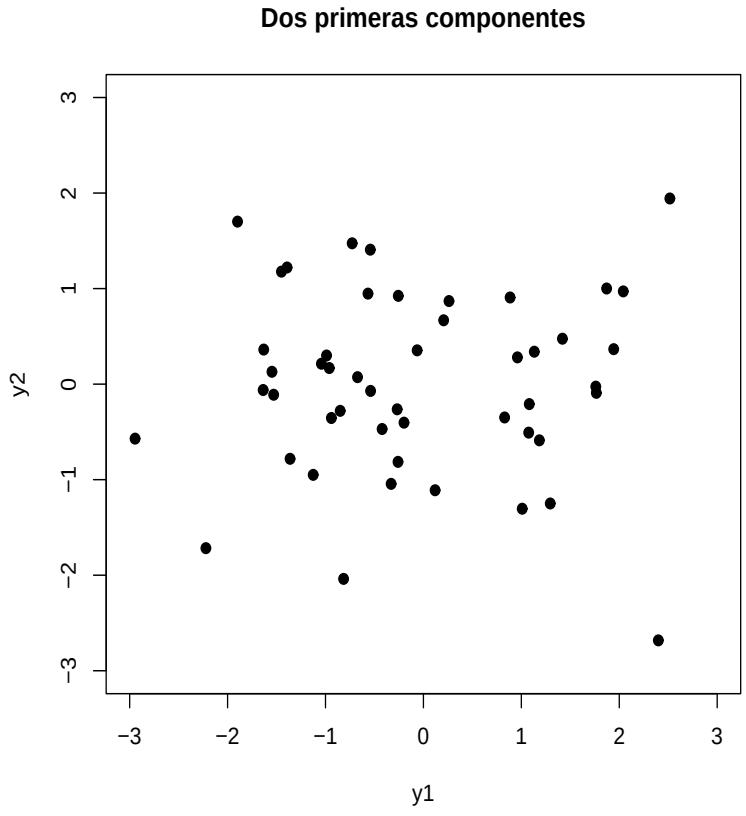


CP1	CP2	CP3
-0.65	0.19	0.73
-0.70	0.21	-0.68
0.28	0.96	0.00

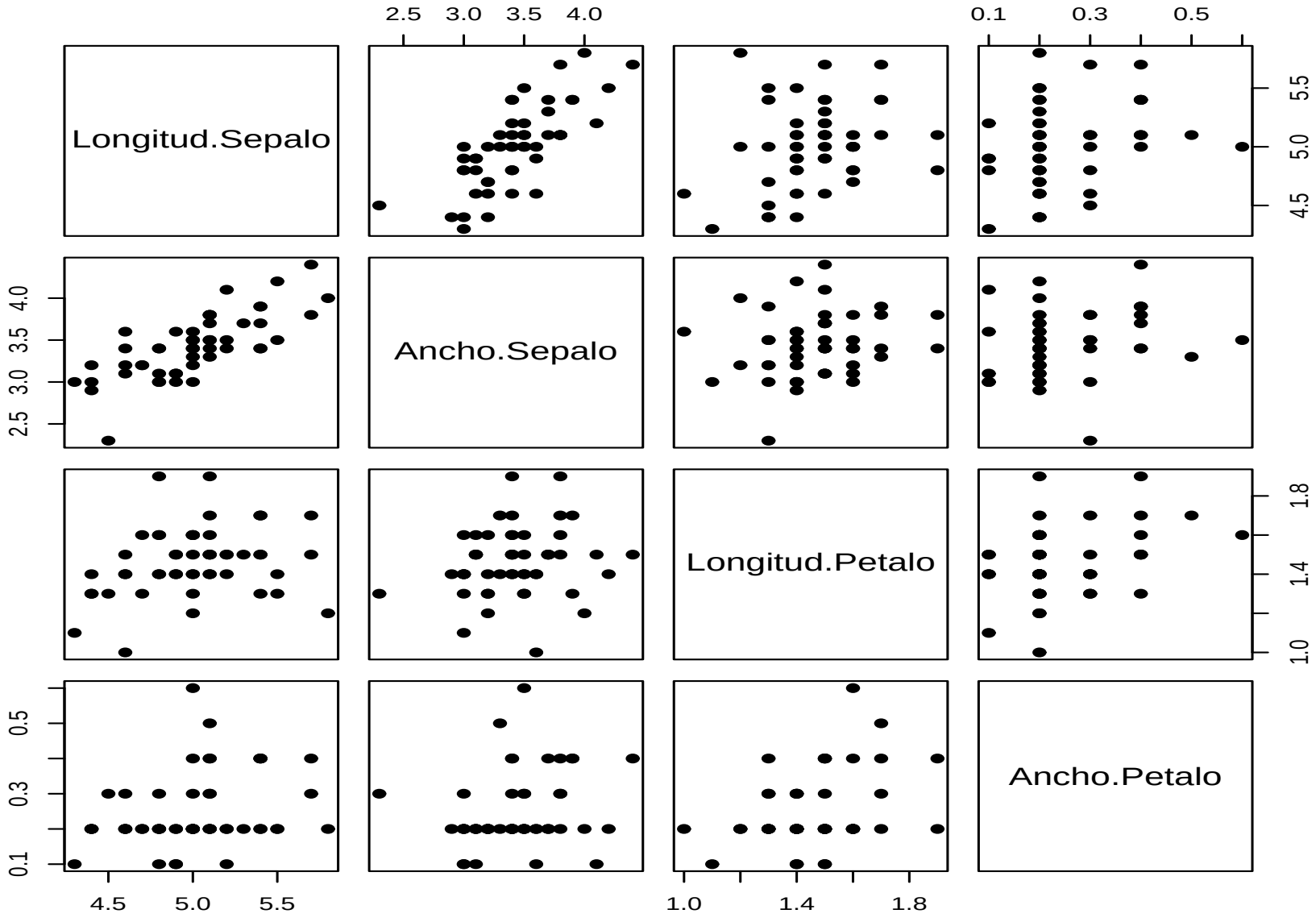
$$
\begin{aligned}
\mathbf{y}_{(1)} &= -0.65\,\mathrm{VAR1} - 0.70\,\mathrm{VAR2} + 0.28\,\mathrm{VAR3} \\
\mathbf{y}_{(2)} &= 0.19\,\mathrm{VAR1} + 0.21\,\mathrm{VAR2} + 0.96\ \mathrm{VAR3}, \\
\mathbf{y}_{(3)} &= 0.73\,\mathrm{VAR1} - 0.68\,\mathrm{VAR2}
\end{aligned}
$$

	CP1	CP2	CP3
Desviación típica	1.40	0.95	0.29
Porcentaje de varianza	66.5 %	30.7 %	2.8 %
Porcentaje acumulado	66.5 %	97.2 %	100 %





Lirios de la especie *setosa*



Resultados SPSS: componentes normadas

Salida SPSS para variables estandarizadas (matriz de correlaciones)

Componente	Autovalores iniciales			Sumas de las saturaciones al cuadrado de la extracción		
	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
1	2,075	51,870	51,870	2,075	51,870	51,870
2	,986	24,656	76,525	,986	24,656	76,525
3	,692	17,301	93,826	,692	17,301	93,826
4	,247	6,174	100,000	,247	6,174	100,000

Método de extracción: Análisis de Componentes principales.

a. Sólo aquellos casos para los que ESPECIE = 1, serán utilizados en la fase de análisis.

Matriz de componentes[a,b]

	Componente			
	1	2	3	4
SL	,865	-,339	,109	-,353
SW	,841	-,417	,003	,345
PL	,519	,667	,532	,050
PW	,592	,502	-,630	-,018

Método de extracción: Análisis de componentes principales.

a. 4 componentes extraídos

b. Sólo aquellos casos para los que ESPECIE = 1, serán utilizados en la fase de análisis.

Interpretación

- En la salida anterior se ha utilizado la matriz de correlaciones (lo que equivale a estandarizar previamente las variables)
- La primera tabla contiene los autovalores (ordenados de mayor a menor) es decir las varianzas correspondientes a cada componente.

- En la salida de SPSS la norma de los autovectores es $\sqrt{\lambda_i}$. Es decir, los autovectores unitarios $\mathbf{a}_i$ se obtienen dividiendo por $\sqrt{\lambda_i}$. Por ejemplo, para la primera componente principal, el autovector unitario es:

$$\frac{(0.865, 0.841, 0.519, 0.592)}{\sqrt{2.075}} = (0.6004, 0.5838, 0.3603, 0.4109)$$

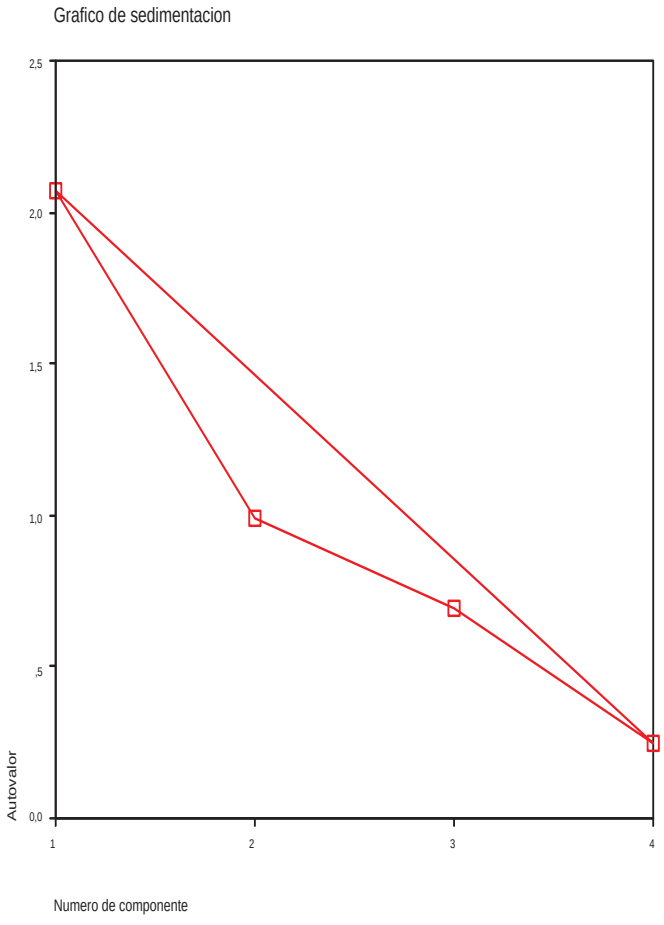
- Muchos programas calculan directamente los autovectores unitarios
- La ventaja de la salida de SPSS es que, si el análisis se hace con los datos estandarizados ($s_{jj} = 1$), entonces las coordenadas de $\sqrt{\lambda_i}\mathbf{a}_i$ son las correlaciones entre la componente principal i y cada una de las variables originales
- Por ejemplo, la correlación entre la primera componente principal y la anchura del sépalo es: 0.841
- ¿Cuál es la correlación entre la segunda componente y la longitud del pétalo?

Interpretación de las componentes:

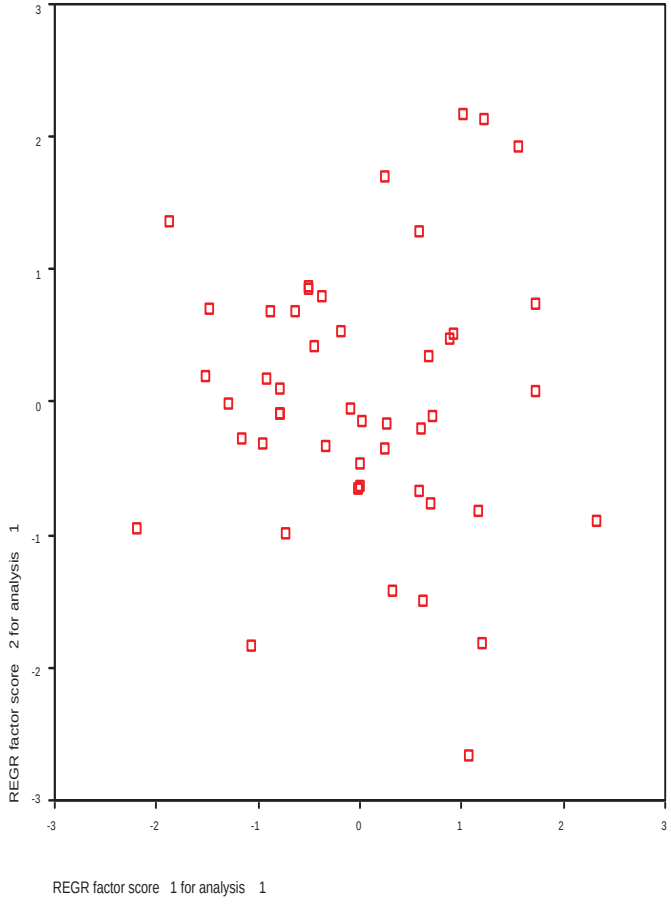
- Primera: Es esencialmente una media de las 4 variables (donde las dimensiones de los sépalos tienen más peso que la de los sépalos. Representa el *tamaño* de los lirios.
- Segunda: Esta componente mide la diferencia entre el pétalo y el sépalo.
- Tercera: Mide la diferencia entre la anchura y la longitud del pétalo.
- Cuarta: Mide la diferencia entre la anchura y la longitud del sépalo.

Más información que proporciona SPSS

Gráfico de sedimentación:



Las dos primeras componentes estandarizadas:



Resultados SPSS: componentes no normadas

Salida de SPSS para variables no estandarizadas (matriz de covarianzas)

Matriz de componentes[a]

	Bruta				Reescalada			
	Componente				Componente			
	1	2	3	4	1	2	3	4
SL	,325	,115	-,071	-,001	,923	,327	-,202	-,002
SW	,360	-,118	,044	-,005	,944	-,310	,114	-,014
PL	,046	,094	,137	-,022	,267	,540	,788	-,128
PW	,034	,016	,034	,094	,321	,151	,320	,879

Método de extracción: Análisis de componentes principales.

a. 4 componentes extraídos

Varianza total explicada

		Sumas de las saturaciones al cuadrado de la extracción		
	Componente	Total	% de la varianza	% acumulado
Bruta	1	,239	76,687	76,687
	2	,036	11,658	88,345
	3	,027	8,632	96,978
	4	,009	3,022	100,000
Reescalada	1	1,917	47,924	47,924
	2	,517	12,924	60,847
	3	,777	19,433	80,281
	4	,789	19,719	100,000

Método de extracción: Análisis de Componentes principales.

Interpretación

Componente bruta: Si $\mathbf{a}_i$ es el autovector unitario correspondiente al autovalor λ_i, entonces la componente bruta i es

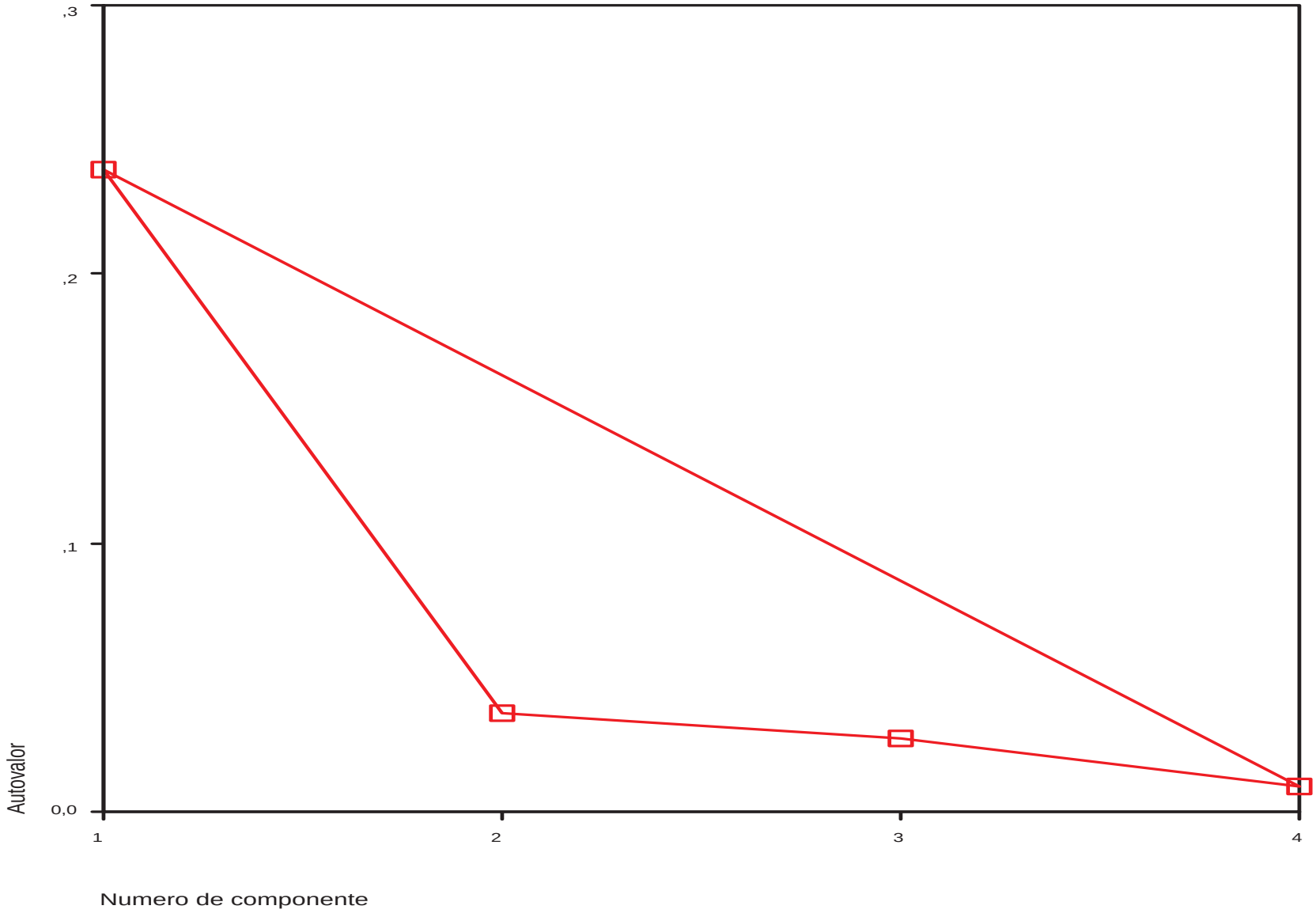
$$\mathbf{b}_i = \sqrt{\lambda_i}\mathbf{a}_i$$

Este resultado es el análogo al que se obtiene usando las correlaciones

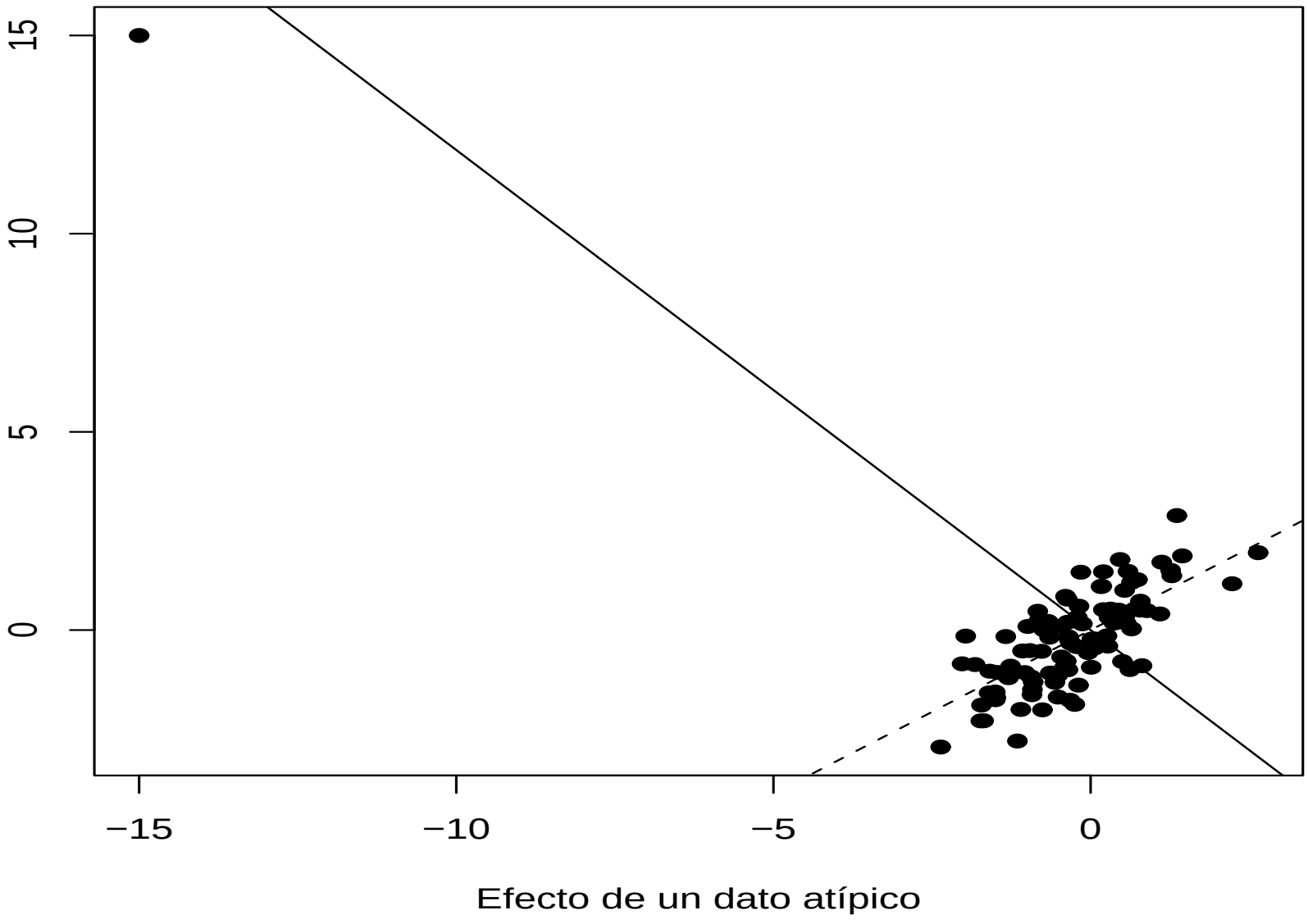
Componente reescalada: Si $\mathbf{a}_i$ es el autovector unitario correspondiente al autovalor λ_i, entonces la componente reescalada i es $\mathbf{c}_i = (c_{i1}, \ldots, c_{ip})'$, donde

$$c_{ij} = \frac{\sqrt{\lambda_i}\mathbf{a}_{ij}}{s_j}$$

Gráfico de sedimentación



Efecto de los datos atípicos



Resultado si no se incluye el dato atípico:

```
    CP1         CP2
0.6070823   -0.7946389
0.7946389    0.6070823
```

Resultado incluyendo el dato atípico:

```
    CP1         CP2
 0.6367243  -0.7710915
-0.7710915  -0.6367243
```

Las componentes se intercambian si se incluye el dato atípico

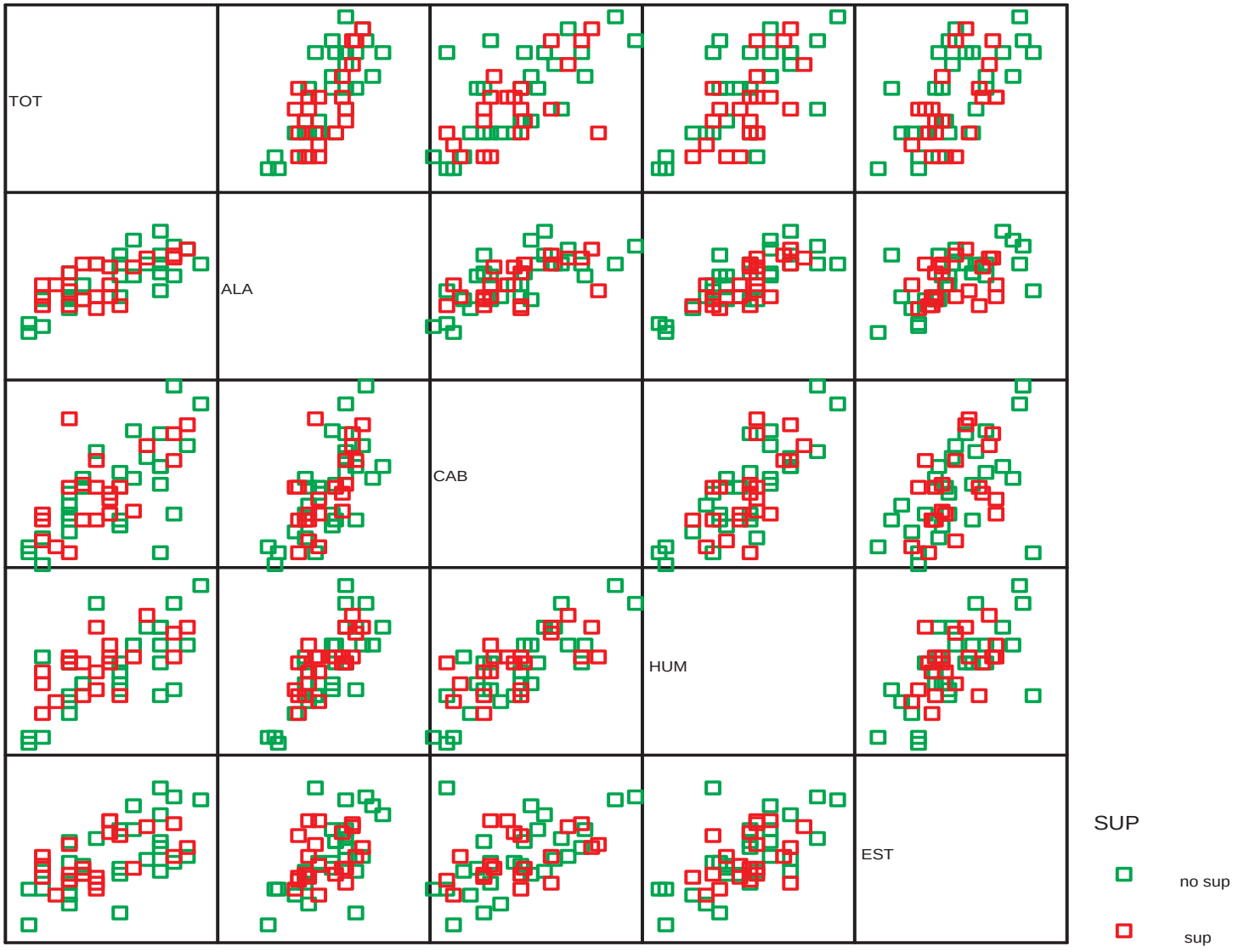
Datos de gorriones: variables

Nombre variable	Descripción
TOT	Longitud total
ALA	Extensión de las alas
CAB	Longitud del pico y la cabeza
HUM	Longitud del húmero
EST	Longitud del esternón

Observaciones:

- Todas las variables se miden en mm.
- El fichero contiene datos de 49 gorriones.
- Los 21 primeros gorriones fueron los supervivientes.

Matriz de diagramas de dispersión



Medidas descriptivas numéricas

Estadísticos descriptivos

	Media	Desviación típica	N
TOT	157,9796	3,65428	49
ALA	241,3265	5,06782	49
CAB	31,4592	,79475	49
HUM	18,4694	,56429	49
EST	20,8265	,99137	49

Correlaciones

		TOT	ALA	CAB	HUM	EST
TOT	Correlación de Pearson	1	,735	,662	,645	,605
	Sig. (bilateral)	.	,000	,000	,000	,000
	N	49	49	49	49	49
ALA	Correlación de Pearson	,735	1	,674	,769	,529
	Sig. (bilateral)	,000	.	,000	,000	,000
	N	49	49	49	49	49
CAB	Correlación de Pearson	,662	,674	1	,763	,526
	Sig. (bilateral)	,000	,000	.	,000	,000
	N	49	49	49	49	49
HUM	Correlación de Pearson	,645	,769	,763	1	,607
	Sig. (bilateral)	,000	,000	,000	.	,000
	N	49	49	49	49	49
EST	Correlación de Pearson	,605	,529	,526	,607	1
	Sig. (bilateral)	,000	,000	,000	,000	.
	N	49	49	49	49	49

Componentes principales normadas

Varianza total explicada

Componente	Autovalores iniciales			Sumas de las saturaciones al cuadrado de la extracción		
	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
1	3,616	72,320	72,320	3,616	72,320	72,320
2	,532	10,630	82,950	,532	10,630	82,950
3	,386	7,728	90,678	,386	7,728	90,678
4	,302	6,031	96,709	,302	6,031	96,709
5	,165	3,291	100,000	,165	3,291	100,000

Método de extracción: Análisis de Componentes principales.

Matriz de componentes[a]

	Componente				
	1	2	3	4	5
TOT	,859	,037	-,429	,231	-,152
ALA	,878	-,218	-,212	-,301	,215
CAB	,857	-,237	,283	,333	,139
HUM	,895	-,135	,255	-,213	-,264
EST	,756	,639	,111	-,038	,078

Método de extracción: Análisis de componentes principales.

a. 5 componentes extraídos

Interpretación de las dos primeras componentes

La primera componente principal es:

$$y_1 = 0.45\texttt{TOT} + 0.46\texttt{ALA} + 0.45\texttt{CAB} + 0.47\texttt{HUM} + 0.39\texttt{EST}$$

Puede interpretarse como un **índice del tamaño** de los gorriones.

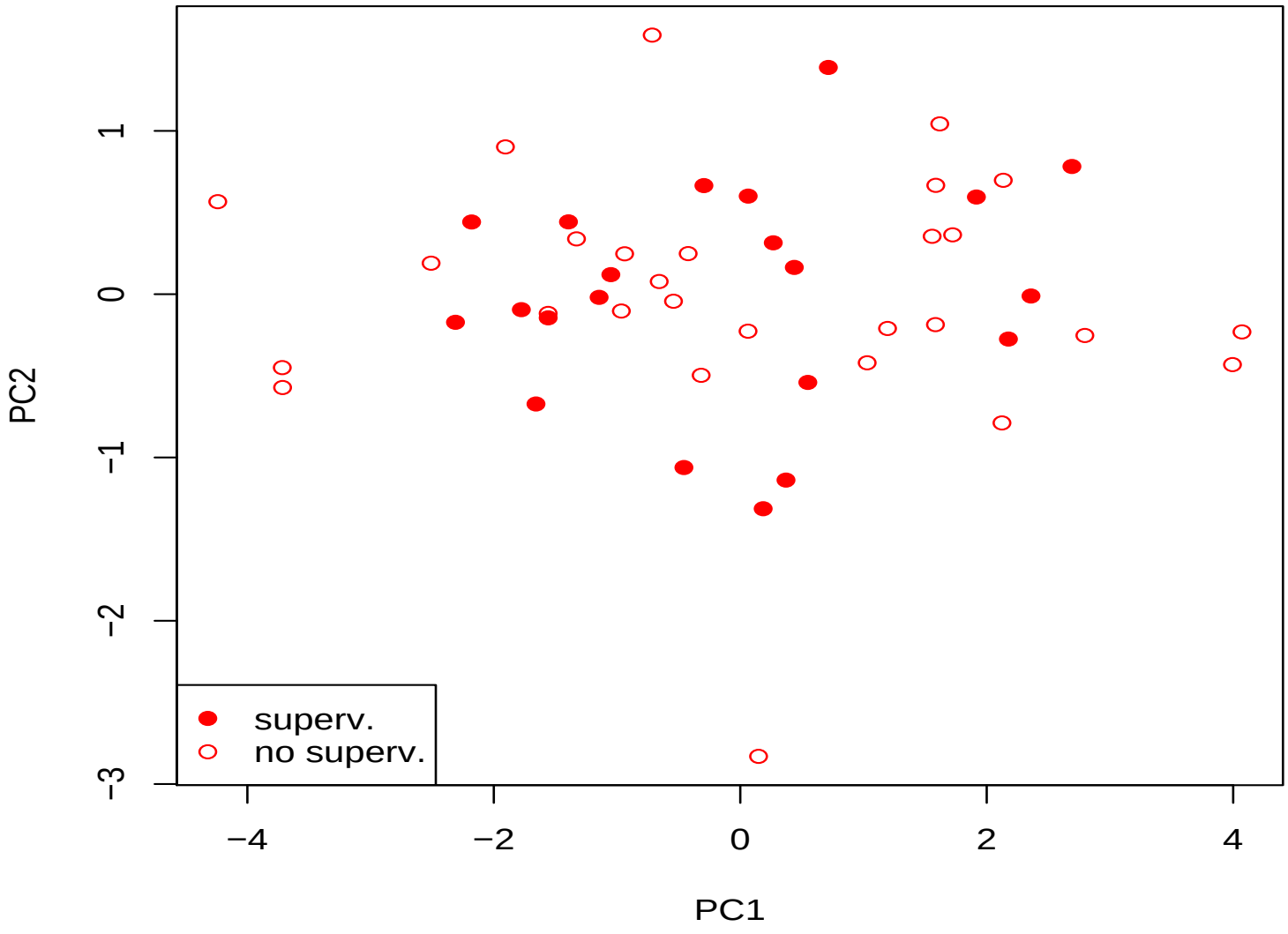
La segunda componente principal se calcula:

$$y_2 = 0.05\texttt{TOT} - 0.29\texttt{ALA} - 0.32\texttt{CAB} - 0.18\texttt{HUM} + 0.87\texttt{EST}$$

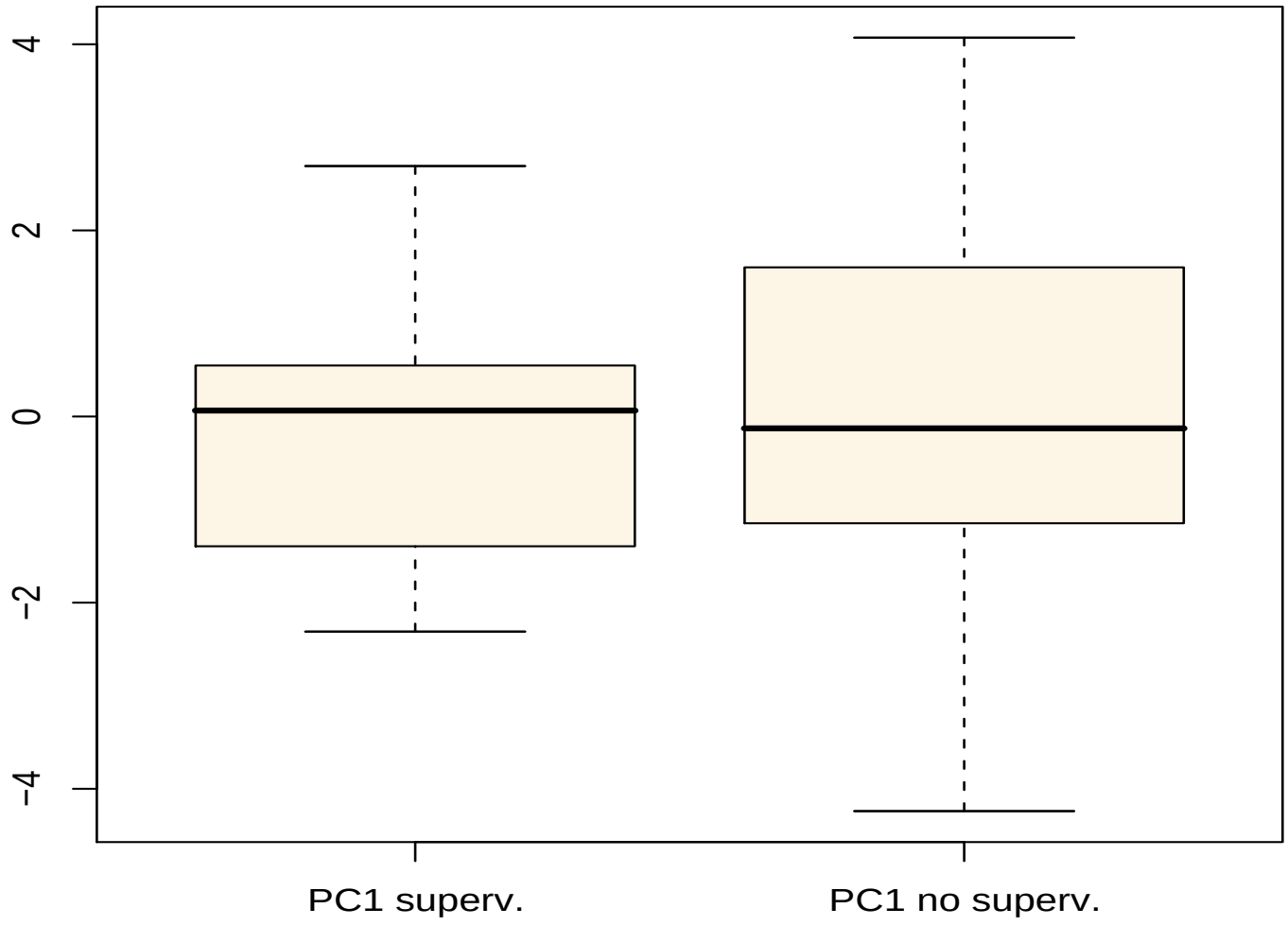
Puede interpretarse como un **índice de forma**. Opone las medidas de las alas, cabeza y húmero a la medida del esternón. La medida total apenas interviene.

Ambas componentes contienen conjuntamente un 82.95% de la varianza total.

Un gráfico de las dos primeras componentes



La primera componente principal según supervivencia

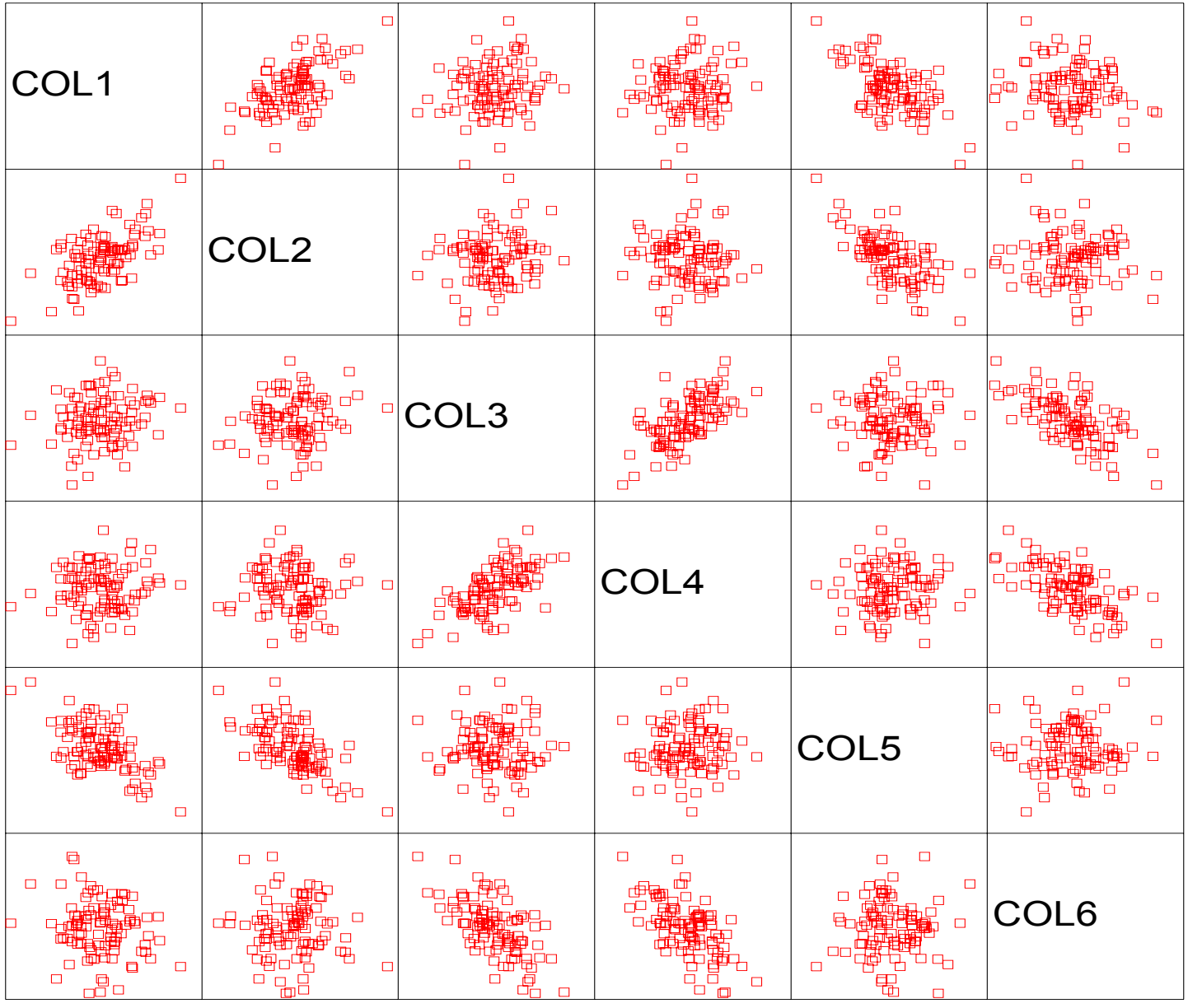


Modelo factorial

- En un estudio sobre el efecto de un analgésico, cada sujeto debe puntuar el medicamento en cada uno de los seis aspectos siguientes: no daña al estómago, no se detectan efectos secundarios, elimina el dolor, actúa con rapidez, no provoca somnolencia, produce un alivio limitado.
- El fichero contiene las respuestas de 100 individuos a las seis preguntas de la encuesta.
- Se ajusta a estos datos un modelo con dos factores:

$$
\begin{aligned}
x_1 &= a_{11} f_1 + a_{12} f_2 + u_1 \\
x_2 &= a_{21} f_1 + a_{22} f_2 + u_2 \\
x_3 &= a_{31} f_1 + a_{32} f_2 + u_3 \\
x_4 &= a_{41} f_1 + a_{42} f_2 + u_4 \\
x_5 &= a_{51} f_1 + a_{52} f_2 + u_5 \\
x_6 &= a_{61} f_1 + a_{62} f_2 + u_6
\end{aligned}
$$

Matriz de diagramas de dispersión



Matriz de correlaciones

Correlaciones

		COL1	COL2	COL3	COL4	COL5	COL6
COL1	Correlación de Pearson	1	,596**	,162	,089	-,592**	-,170
	Sig. (bilateral)	.	,000	,107	,381	,000	,091
	N	100	100	100	100	100	100
COL2	Correlación de Pearson	,596**	1	,131	-,067	-,641**	-,065
	Sig. (bilateral)	,000	.	,194	,506	,000	,522
	N	100	100	100	100	100	100
COL3	Correlación de Pearson	,162	,131	1	,632**	-,027	-,632**
	Sig. (bilateral)	,107	,194	.	,000	,788	,000
	N	100	100	100	100	100	100
COL4	Correlación de Pearson	,089	-,067	,632**	1	,057	-,611**
	Sig. (bilateral)	,381	,506	,000	.	,574	,000
	N	100	100	100	100	100	100
COL5	Correlación de Pearson	-,592**	-,641**	-,027	,057	1	,029
	Sig. (bilateral)	,000	,000	,788	,574	.	,775
	N	100	100	100	100	100	100
COL6	Correlación de Pearson	-,170	-,065	-,632**	-,611**	,029	1
	Sig. (bilateral)	,091	,522	,000	,000	,775	.
	N	100	100	100	100	100	100

**. La correlación es significativa al nivel 0,01 (bilateral).

Salida SPSS

La siguiente salida se ha obtenido utilizando:

- Matriz de correlaciones
- Estimación de parámetros por máxima verosimilitud
- Extracción de 2 factores
- Rotación por el criterio *varimax*

Comunalidades

	Inicial	Extracción
COL1	,453	,567
COL2	,509	,656
COL3	,517	,654
COL4	,499	,635
COL5	,487	,635
COL6	,479	,608

Método de extracción: Máxima verosimilitud.

Varianza total explicada

Factor	Autovalores iniciales			Sumas de las saturaciones al cuadrado de la extracción		
	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
1	2,431	40,512	40,512	2,054	34,231	34,231
2	2,070	34,498	75,010	1,701	28,344	62,575
3	,439	7,319	82,329			
4	,387	6,450	88,778			
5	,380	6,340	95,118			
6	,293	4,882	100,000			

Método de extracción: Máxima verosimilitud.

Varianza total explicada

Factor	Suma de las saturaciones al cuadrado de la rotación		
	Total	% de la varianza	% acumulado
1	1,899	31,656	31,656
2	1,855	30,919	62,575
3			
4			
5			
6			

Método de extracción: Máxima verosimilitud.

Matriz factorial[a]

	Factor	
	1	2
COL1	,566	-,496
COL2	,511	-,628
COL3	,694	,415
COL4	,591	,535
COL5	-,459	,651
COL6	-,657	-,421

Método de extracción: Máxima verosimilitud.

a. 2 factores extraídos. Requeridas 3 iteraciones.

El modelo estimado es:

$$
\begin{aligned}
x_1 &= 0.566 f_1 - 0.496 f_2 + u_1 \\
x_2 &= 0.511 f_1 - 0.628 f_2 + u_2 \\
x_3 &= 0.694 f_1 + 0.415 f_2 + u_3 \\
x_4 &= 0.591 f_1 + 0.535 f_2 + u_4 \\
x_5 &= -0.459 f_1 + 0.651 f_2 + u_5 \\
x_6 &= -0.657 f_1 - 0.421 f_2 + u_6
\end{aligned}
$$

Observaciones

- Las correlaciones de los dos factores con las seis variables es bastante similar. Esto hace difícil asociar factores con subconjuntos de variables y, por lo tanto, interpretar el significado de los factores.

- Las comunalidades se pueden obtener a partir del modelo estimado. Por ejemplo para x_1:

$$\text{Comunalidad}_1 = a_{11}^2 + a_{12}^2 = 0.566^2 + 0.496^2 = 0.567.$$

  Análogamente para el resto de las variables.

- El porcentaje de la variabilidad total correspondiente al primer factor es:

$$\frac{0.566^2 + 0.511^2 + \cdots 0.657^2}{6} \approx \frac{2.055}{6} \approx 0.3425$$

  Análogamente para el segundo factor.

- A partir del modelo, también se pueden estimar las correlaciones (correlaciones reproducidas). Por ejemplo:

$$\widehat{\text{Corr}(x_1, x_2)} = 0.566 \times 0.511 + 0.496 \times 0.628 \approx 0.600$$

Prueba de la bondad de ajuste

Chi-cuadrado	gl	Sig.
3,289	4	,511

Correlaciones reproducidas

		COL1	COL2	COL3	COL4	COL5	COL6
Correlación reproducida	COL1	,566[b]	,601	,187	,069	-,583	-,163
	COL2	,601	,656[b]	,094	-,034	-,644	-,071
	COL3	,187	,094	,654[b]	,632	-,048	-,630
	COL4	,069	-,034	,632	,635[b]	,077	-,613
	COL5	-,583	-,644	-,048	,077	,635[b]	,028
	COL6	-,163	-,071	-,630	-,613	,028	,608[b]
Residual[a]	COL1		-,005	-,024	,020	-,009	-,007
	COL2	-,005		,037	-,033	,003	,007
	COL3	-,024	,037		,001	,021	-,001
	COL4	,020	-,033	,001		-,020	,002
	COL5	-,009	,003	,021	-,020		,001
	COL6	-,007	,007	-,001	,002	,001	

Método de extracción: Máxima verosimilitud.

a. Los residuos se calculan entre las correlaciones observadas y reproducidas. Hay 0 (,0%) residuales no redundantes con valores absolutos mayores que 0,05.

b. Comunalidades reproducidas

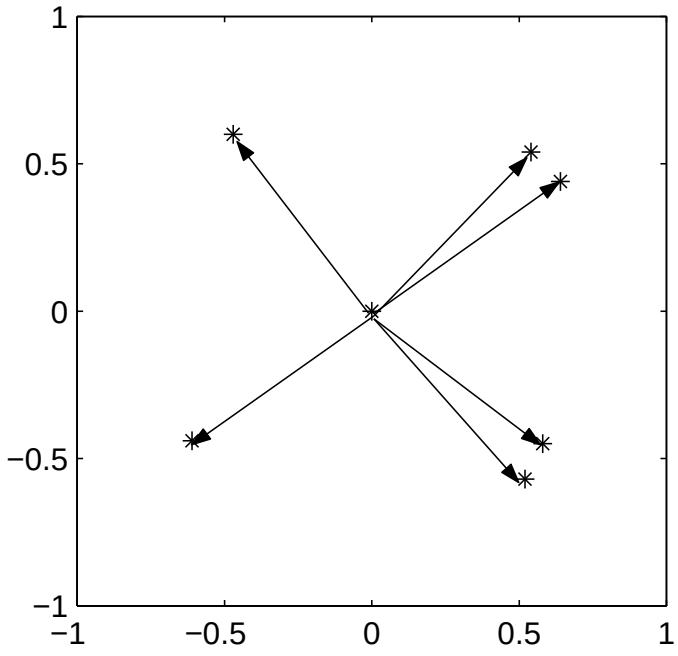
Matriz de factores rotados[a]

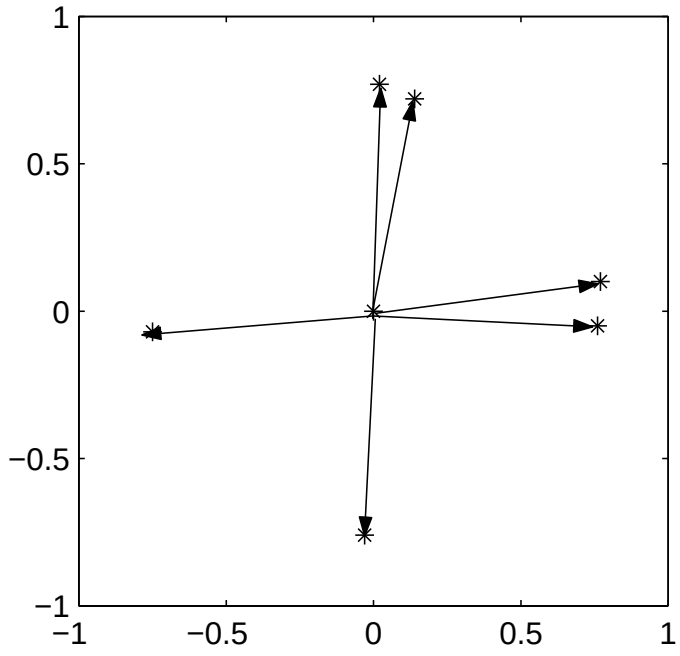
	Factor	
	1	2
COL1	,140	,740
COL2	,015	,810
COL3	,802	,101
COL4	,795	-,057
COL5	,039	-,796
COL6	-,777	-,073

Método de extracción: Máxima verosimilitud.
Método de rotación: Normalización Varimax con Kaiser.

a. La rotación ha convergido en 3 iteraciones.

Rotación



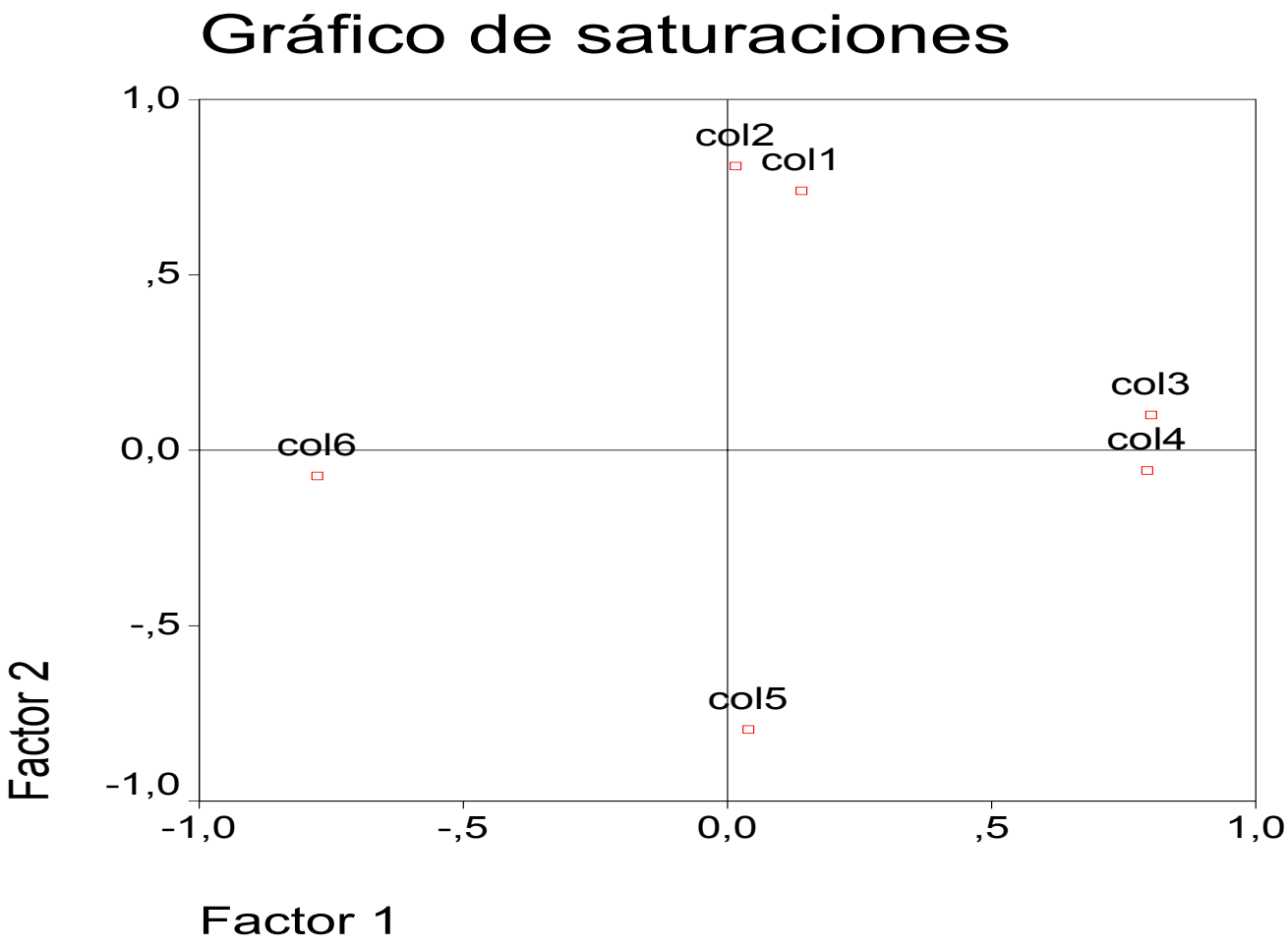


- Las diferencias entre las correlaciones reproducidas y la matriz de correlaciones aparecen con el nombre de *residual*. Estas diferencias son la base del contraste de bondad de ajuste.
- El p-valor del contraste de bondad de ajuste es 0.511, por lo tanto no podemos rechazar $H_0 : \Sigma = AA' + B$. Parece que el modelo con dos factores es adecuado.
- Los factores rotados son mucho más fáciles de interpretar. Cada factor está asociado con las siguientes variables:
  - **Factor 1:** elimina el dolor (+), actúa con rapidez (+), produce un alivio limitado (-).
  - **Factor 2:** no daña al estómago (+), no produce efectos secundarios (+), no provoca somnolencia (-).
- **Factor 1 = Eficacia**
- **Factor 2 = Efectos secundarios**

Matriz de transformación de los factores

Factor	1	2
1	,788	,616
2	,616	-,788

Método de extracción: Máxima verosimilitud.
Método de rotación: Normalización Varimax con Kaiser.



- La matriz de transformación de los factores C nos da las operaciones que hay que hacer para llevar a cabo la rotación. Si A es la matriz de cargas original y A^* es la matriz de cargas rotada, entonces $A^* = AC$.
- Por ejemplo, la primera fila de A^* se calcula de la siguiente forma:

$$(0.566, -0.496) \begin{pmatrix} 0.788 & 0.616 \\ 0.616 & -0.788 \end{pmatrix} \approx (0.1405, 0.7395).$$

Análogamente para el resto de filas.

Cuestiones

- ¿Cuánto vale la comunalidad para la variable *Elimina el dolor*?
- ¿Qué parte de la variabilidad de *No se detectan efectos secundarios* no corresponde a los dos factores comunes?
- ¿Cuánto vale la correlación estimada entre las variables *Elimina el dolor* y *No provoca somnolencia*?
- ¿Cuánto vale la correlación estimada entre la variable *No provoca somnolencia* y el segundo factor común (no rotado)?
- ¿Cuánto vale la correlación entre los dos factores comunes?