

---

# **Descomposición en valores singulares: introducción y aplicaciones**

---

## **Estudio de caso y guía de resolución en R**

PID\_00262391

Francesc Pozo Montero  
Jordi Ripoll Missé

**Francesc Pozo Montero**

Licenciado en Matemáticas por la Universidad de Barcelona (2000) y doctor en Matemática Aplicada por la Universidad Politécnica de Cataluña (2005). Ha sido profesor asociado de la Universidad Autónoma de Barcelona y profesor asociado, colaborador y actualmente profesor agregado en la Universidad Politécnica de Cataluña. Además, es cofundador del Grupo de Innovación Matemática E-learning (GIMEL), responsable de varios proyectos de innovación docente y autor de varias publicaciones. Como miembro del grupo de investigación consolidado CoDALab, centra su investigación en la teoría de control y las aplicaciones en ingeniería mecánica y civil, así como en el uso de la ciencia de datos para la monitorización de la integridad estructural y para la monitorización de la condición, sobre todo en turbinas eólicas.

**Jordi Ripoll Missé**

Licenciado en Matemáticas y doctor en Ciencias Matemáticas por la Universidad de Barcelona (2005). Profesor colaborador de la Universitat Oberta de Catalunya desde 2011 y profesor del Departamento de Informática, Matemática Aplicada y Estadística de la Universidad de Girona (UdG) desde 1996, donde actualmente es profesor agregado y desarrolla tareas de investigación en el ámbito de la biología matemática (modelos con ecuaciones en derivadas parciales y dinámica evolutiva). También ha sido profesor y tutor de la UNED en dos etapas, primero en el centro asociado de Terrassa y actualmente en el de Girona. Ha participado en numerosos proyectos de innovación docente, especialmente en cuanto al aprendizaje de las matemáticas en línea.

El encargo y la creación de este recurso de aprendizaje UOC han sido coordinados por la profesora: Cristina Cano Bastidas (2019)

Primera edición: febrero 2019

© Francesc Pozo Montero, Jordi Ripoll Missé

Todos los derechos reservados

© de esta edición, FUOC, 2019

Av. Tibidabo, 39-43, 08035 Barcelona

Diseño: Manel Andreu

Realización editorial: Oberta UOC Publishing, SL

*Ninguna parte de esta publicación, incluido el diseño general y la cubierta, puede ser copiada, reproducida, almacenada o transmitida de ninguna forma, ni por ningún medio, sea éste eléctrico, químico, mecánico, óptico, grabación, fotocopia, o cualquier otro, sin la previa autorización escrita de los titulares del copyright.*

## Índice

<b>1. Estudio de caso .....</b>	<b>5</b>
<b>Bibliografía .....</b>	<b>28</b>



## 1. Estudio de caso

Para este estudio de caso tendremos en cuenta el conjunto de datos Iris que ya hemos trabajado en el módulo “Descomposición en valores singulares: introducción y aplicaciones”. Estos datos proporcionan las medidas en centímetros de las variables de longitud y anchura del sépalo y de longitud y anchura del pétalo, respectivamente, para cincuenta flores de cada una de las especies *Iris setosa*, *Iris versicolor* e *Iris virginica*. En el módulo “Descomposición en valores singulares: introducción y aplicaciones” hemos considerado una muestra de quince flores, cinco de cada tipo. En este estudio de caso consideramos la totalidad de la muestra. Los valores de estas variables para las cinco primeras flores de cada tipo se encuentran en la tabla 1.

### Iris

Los datos fueron recogidos por Edgar Anderson en 1935 y publicados en el artículo “The irises of the Gaspé Peninsula”, *Bulletin of the American Iris Society* (59, págs. 2-5).

Tabla 1. Longitud y anchura del sépalo y longitud y anchura del pétalo (en centímetros)

flor	long. sépalo	anchura sépalo	long. pétalo	anchura pétalo	tipo
1	5.1	3.5	1.4	0.2	<i>setosa</i>
2	4.9	3.0	1.4	0.2	<i>setosa</i>
3	4.7	3.2	1.3	0.2	<i>setosa</i>
4	4.6	3.1	1.5	0.2	<i>setosa</i>
5	5.0	3.6	1.4	0.2	<i>setosa</i>
51	7.0	3.2	4.7	1.4	<i>versicolor</i>
52	6.4	3.2	4.5	1.5	<i>versicolor</i>
53	6.9	3.1	4.9	1.5	<i>versicolor</i>
54	5.5	2.3	4.0	1.3	<i>versicolor</i>
55	6.5	2.8	4.6	1.5	<i>versicolor</i>
101	6.3	3.3	6.0	2.5	<i>virginica</i>
102	5.8	2.7	5.1	1.9	<i>virginica</i>
103	7.1	3.0	5.9	2.1	<i>virginica</i>
104	6.3	2.9	5.6	1.8	<i>virginica</i>
105	6.5	3.0	5.8	2.2	<i>virginica</i>

Fuente: Edgar Anderson (1935). “The irises of the Gaspé Peninsula”

En este estudio de caso emplearemos el *software* libre R para llevar a cabo el análisis de las componentes principales. Lo haremos, además, de dos maneras: *a)* siguiendo los pasos de la técnica y *b)* aplicando directamente la instrucción `prcomp` y analizando e interpretando los resultados.

En primer lugar, cargamos el conjunto de datos `iris` y mostramos el resultado:

```
> data(`iris`)
> iris
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa

2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa
11	5.4	3.7	1.5	0.2	setosa
12	4.8	3.4	1.6	0.2	setosa
13	4.8	3.0	1.4	0.1	setosa
14	4.3	3.0	1.1	0.1	setosa
15	5.8	4.0	1.2	0.2	setosa
16	5.7	4.4	1.5	0.4	setosa
17	5.4	3.9	1.3	0.4	setosa
18	5.1	3.5	1.4	0.3	setosa
19	5.7	3.8	1.7	0.3	setosa
20	5.1	3.8	1.5	0.3	setosa
21	5.4	3.4	1.7	0.2	setosa
22	5.1	3.7	1.5	0.4	setosa
23	4.6	3.6	1.0	0.2	setosa
24	5.1	3.3	1.7	0.5	setosa
25	4.8	3.4	1.9	0.2	setosa
26	5.0	3.0	1.6	0.2	setosa
27	5.0	3.4	1.6	0.4	setosa
28	5.2	3.5	1.5	0.2	setosa
29	5.2	3.4	1.4	0.2	setosa
30	4.7	3.2	1.6	0.2	setosa
31	4.8	3.1	1.6	0.2	setosa
32	5.4	3.4	1.5	0.4	setosa
33	5.2	4.1	1.5	0.1	setosa
34	5.5	4.2	1.4	0.2	setosa
35	4.9	3.1	1.5	0.2	setosa
36	5.0	3.2	1.2	0.2	setosa
37	5.5	3.5	1.3	0.2	setosa
38	4.9	3.6	1.4	0.1	setosa
39	4.4	3.0	1.3	0.2	setosa
40	5.1	3.4	1.5	0.2	setosa
41	5.0	3.5	1.3	0.3	setosa
42	4.5	2.3	1.3	0.3	setosa
43	4.4	3.2	1.3	0.2	setosa
44	5.0	3.5	1.6	0.6	setosa
45	5.1	3.8	1.9	0.4	setosa
46	4.8	3.0	1.4	0.3	setosa
47	5.1	3.8	1.6	0.2	setosa
48	4.6	3.2	1.4	0.2	setosa

49	5.3	3.7	1.5	0.2	setosa
50	5.0	3.3	1.4	0.2	setosa
51	7.0	3.2	4.7	1.4	versicolor
52	6.4	3.2	4.5	1.5	versicolor
53	6.9	3.1	4.9	1.5	versicolor
54	5.5	2.3	4.0	1.3	versicolor
55	6.5	2.8	4.6	1.5	versicolor
56	5.7	2.8	4.5	1.3	versicolor
57	6.3	3.3	4.7	1.6	versicolor
58	4.9	2.4	3.3	1.0	versicolor
59	6.6	2.9	4.6	1.3	versicolor
60	5.2	2.7	3.9	1.4	versicolor
61	5.0	2.0	3.5	1.0	versicolor
62	5.9	3.0	4.2	1.5	versicolor
63	6.0	2.2	4.0	1.0	versicolor
64	6.1	2.9	4.7	1.4	versicolor
65	5.6	2.9	3.6	1.3	versicolor
66	6.7	3.1	4.4	1.4	versicolor
67	5.6	3.0	4.5	1.5	versicolor
68	5.8	2.7	4.1	1.0	versicolor
69	6.2	2.2	4.5	1.5	versicolor
70	5.6	2.5	3.9	1.1	versicolor
71	5.9	3.2	4.8	1.8	versicolor
72	6.1	2.8	4.0	1.3	versicolor
73	6.3	2.5	4.9	1.5	versicolor
74	6.1	2.8	4.7	1.2	versicolor
75	6.4	2.9	4.3	1.3	versicolor
76	6.6	3.0	4.4	1.4	versicolor
77	6.8	2.8	4.8	1.4	versicolor
78	6.7	3.0	5.0	1.7	versicolor
79	6.0	2.9	4.5	1.5	versicolor
80	5.7	2.6	3.5	1.0	versicolor
81	5.5	2.4	3.8	1.1	versicolor
82	5.5	2.4	3.7	1.0	versicolor
83	5.8	2.7	3.9	1.2	versicolor
84	6.0	2.7	5.1	1.6	versicolor
85	5.4	3.0	4.5	1.5	versicolor
86	6.0	3.4	4.5	1.6	versicolor
87	6.7	3.1	4.7	1.5	versicolor
88	6.3	2.3	4.4	1.3	versicolor
89	5.6	3.0	4.1	1.3	versicolor
90	5.5	2.5	4.0	1.3	versicolor
91	5.5	2.6	4.4	1.2	versicolor
92	6.1	3.0	4.6	1.4	versicolor
93	5.8	2.6	4.0	1.2	versicolor
94	5.0	2.3	3.3	1.0	versicolor
95	5.6	2.7	4.2	1.3	versicolor

96	5.7	3.0	4.2	1.2 versicolor
97	5.7	2.9	4.2	1.3 versicolor
98	6.2	2.9	4.3	1.3 versicolor
99	5.1	2.5	3.0	1.1 versicolor
100	5.7	2.8	4.1	1.3 versicolor
101	6.3	3.3	6.0	2.5 virginica
102	5.8	2.7	5.1	1.9 virginica
103	7.1	3.0	5.9	2.1 virginica
104	6.3	2.9	5.6	1.8 virginica
105	6.5	3.0	5.8	2.2 virginica
106	7.6	3.0	6.6	2.1 virginica
107	4.9	2.5	4.5	1.7 virginica
108	7.3	2.9	6.3	1.8 virginica
109	6.7	2.5	5.8	1.8 virginica
110	7.2	3.6	6.1	2.5 virginica
111	6.5	3.2	5.1	2.0 virginica
112	6.4	2.7	5.3	1.9 virginica
113	6.8	3.0	5.5	2.1 virginica
114	5.7	2.5	5.0	2.0 virginica
115	5.8	2.8	5.1	2.4 virginica
116	6.4	3.2	5.3	2.3 virginica
117	6.5	3.0	5.5	1.8 virginica
118	7.7	3.8	6.7	2.2 virginica
119	7.7	2.6	6.9	2.3 virginica
120	6.0	2.2	5.0	1.5 virginica
121	6.9	3.2	5.7	2.3 virginica
122	5.6	2.8	4.9	2.0 virginica
123	7.7	2.8	6.7	2.0 virginica
124	6.3	2.7	4.9	1.8 virginica
125	6.7	3.3	5.7	2.1 virginica
126	7.2	3.2	6.0	1.8 virginica
127	6.2	2.8	4.8	1.8 virginica
128	6.1	3.0	4.9	1.8 virginica
129	6.4	2.8	5.6	2.1 virginica
130	7.2	3.0	5.8	1.6 virginica
131	7.4	2.8	6.1	1.9 virginica
132	7.9	3.8	6.4	2.0 virginica
133	6.4	2.8	5.6	2.2 virginica
134	6.3	2.8	5.1	1.5 virginica
135	6.1	2.6	5.6	1.4 virginica
136	7.7	3.0	6.1	2.3 virginica
137	6.3	3.4	5.6	2.4 virginica
138	6.4	3.1	5.5	1.8 virginica
139	6.0	3.0	4.8	1.8 virginica
140	6.9	3.1	5.4	2.1 virginica
141	6.7	3.1	5.6	2.4 virginica
142	6.9	3.1	5.1	2.3 virginica



143	5.8	2.7	5.1	1.9	virginica
144	6.8	3.2	5.9	2.3	virginica
145	6.7	3.3	5.7	2.5	virginica
146	6.7	3.0	5.2	2.3	virginica
147	6.3	2.5	5.0	1.9	virginica
148	6.5	3.0	5.2	2.0	virginica
149	6.2	3.4	5.4	2.3	virginica
150	5.9	3.0	5.1	1.8	virginica

Para las 150 flores que hemos considerado, podemos ver el valor de las variables:

- 1) longitud del sépalo
- 2) anchura del sépalo
- 3) longitud del pétalo
- 4) anchura del pétalo

También podemos ver una quinta columna, correspondiente al tipo de flor o especie (*setosa*, *versicolor* o *virginica*).

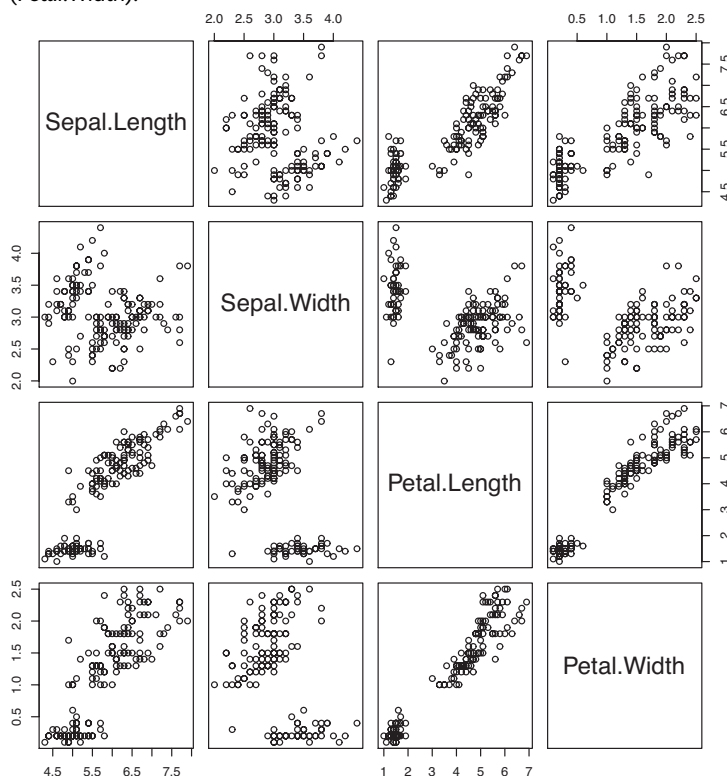
Una primera aproximación al estudio de este conjunto de datos podría ser elaborar los diagramas de dispersión para todos los pares de variables numéricas, es decir, eliminando la variable que nos indica el tipo de flor:

```
> plot(iris[,1:4])
```

#### Diagrama de dispersión

Un diagrama de dispersión (en inglés, *scatter plot*) muestra gráficamente la relación entre dos variables cuantitativas.

Figura 1. Diagrama de dispersión de las variables longitud del sépalo (Sepal.Length), anchura del sépalo (Sepal.Width), longitud del pétalo (Petal.Length) y anchura del pétalo (Petal.Width).



Fuente: elaboración propia

Notad que en la figura 1 hemos generado una matriz de dimensión  $4 \times 4$  con la instrucción `plot(iris[,1:4])`. En la diagonal principal aparece el nombre de las cuatro variables. En el resto de los elementos de la matriz de diagramas de dispersión están los diagramas de dispersión de cada par de variables. Por ejemplo, en la fila 1, columna 2 encontramos el diagrama de dispersión de la variable de longitud del sépalo *versus* la variable de anchura del sépalo. En la fila 2, columna 1 hallamos el mismo diagrama de dispersión pero cambiando los ejes; es decir, como variable horizontal está la anchura del sépalo y como variable vertical, la longitud del sépalo.

Una segunda observación de la figura 1 nos permite ver que:

- 1) Existe una fuerte correlación entre las variables de longitud del pétalo (`Petal.Length`) y de anchura del pétalo (`Petal.Width`). Esta correlación es observable porque la nube de puntos marca una cierta línea recta de pendiente positiva. Esto se traduce en que, cuando la longitud del pétalo crece, la anchura del pétalo también crece, y viceversa.
- 2) En contraposición con la observación anterior, la nube de puntos correspondiente a las variables de longitud del sépalo (`Sepal.Length`) y de anchura del sépalo (`Sepal.Width`) ya no marca una tendencia tan clara. De hecho, parece que los puntos estén distribuidos de forma aleatoria.

Salvo estas dos observaciones, no se pueden afirmar con certeza muchas más cosas. Por lo tanto, podremos extraer pocas conclusiones. Sin embargo, los cálculos que realizaremos después nos ayudarán a confirmar o a desestimar estas impresiones.

Recordad que en el procedimiento de análisis de las componentes principales intervienen las siguientes etapas:

- 1) escalado de datos,
- 2) cálculo de la matriz de covarianzas,
- 3) diagonalización de la matriz de covarianzas, que incluye el cálculo de los valores y vectores propios.

### Escalado de datos

Para el escalado de datos utilizaremos la instrucción `scale` de la forma siguiente:

```
> X<-iris[,1:4]
> X<-scale(X,center = TRUE,scale = TRUE)
```

Observad que en la variable `x` guardamos las variables numéricas del conjunto de datos. La opción `center = TRUE` significa que a todos los elementos de una misma columna les restaremos su media aritmética. La opción `scale = TRUE` significa que todos los elementos de una misma columna serán divididos por su desviación típica. El resultado es una matriz de 150 filas y 4 columnas:

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
[1,]	-0.89767388	1.01560199	-1.33575163	-1.3110521482
[2,]	-1.13920048	-0.13153881	-1.33575163	-1.3110521482
[3,]	-1.38072709	0.32731751	-1.39239929	-1.3110521482
[4,]	-1.50149039	0.09788935	-1.27910398	-1.3110521482
[5,]	-1.01843718	1.24503015	-1.33575163	-1.3110521482
[6,]	-0.53538397	1.93331463	-1.16580868	-1.0486667950
[7,]	-1.50149039	0.78617383	-1.33575163	-1.1798594716
[8,]	-1.01843718	0.78617383	-1.27910398	-1.3110521482
[9,]	-1.74301699	-0.36096697	-1.33575163	-1.3110521482
[10,]	-1.13920048	0.09788935	-1.27910398	-1.4422448248
[11,]	-0.53538397	1.47445831	-1.27910398	-1.3110521482
[12,]	-1.25996379	0.78617383	-1.22245633	-1.3110521482
[13,]	-1.25996379	-0.13153881	-1.33575163	-1.4422448248
[14,]	-1.86378030	-0.13153881	-1.50569459	-1.4422448248
[15,]	-0.05233076	2.16274279	-1.44904694	-1.3110521482
[16,]	-0.17309407	3.08045544	-1.27910398	-1.0486667950
[17,]	-0.53538397	1.93331463	-1.39239929	-1.0486667950
[18,]	-0.89767388	1.01560199	-1.33575163	-1.1798594716
[19,]	-0.17309407	1.70388647	-1.16580868	-1.1798594716
[20,]	-0.89767388	1.70388647	-1.27910398	-1.1798594716
[21,]	-0.53538397	0.78617383	-1.16580868	-1.3110521482
[22,]	-0.89767388	1.47445831	-1.27910398	-1.0486667950
[23,]	-1.50149039	1.24503015	-1.56234224	-1.3110521482
[24,]	-0.89767388	0.55674567	-1.16580868	-0.9174741184
[25,]	-1.25996379	0.78617383	-1.05251337	-1.3110521482
[26,]	-1.01843718	-0.13153881	-1.22245633	-1.3110521482
[27,]	-1.01843718	0.78617383	-1.22245633	-1.0486667950
[28,]	-0.77691058	1.01560199	-1.27910398	-1.3110521482
[29,]	-0.77691058	0.78617383	-1.33575163	-1.3110521482
[30,]	-1.38072709	0.32731751	-1.22245633	-1.3110521482
[31,]	-1.25996379	0.09788935	-1.22245633	-1.3110521482
[32,]	-0.53538397	0.78617383	-1.27910398	-1.0486667950
[33,]	-0.77691058	2.39217095	-1.27910398	-1.4422448248
[34,]	-0.41462067	2.62159911	-1.33575163	-1.3110521482
[35,]	-1.13920048	0.09788935	-1.27910398	-1.3110521482
[36,]	-1.01843718	0.32731751	-1.44904694	-1.3110521482
[37,]	-0.41462067	1.01560199	-1.39239929	-1.3110521482
[38,]	-1.13920048	1.24503015	-1.33575163	-1.4422448248

[39,]	-1.74301699	-0.13153881	-1.39239929	-1.3110521482
[40,]	-0.89767388	0.78617383	-1.27910398	-1.3110521482
[41,]	-1.01843718	1.01560199	-1.39239929	-1.1798594716
[42,]	-1.62225369	-1.73753594	-1.39239929	-1.1798594716
[43,]	-1.74301699	0.32731751	-1.39239929	-1.3110521482
[44,]	-1.01843718	1.01560199	-1.22245633	-0.7862814418
[45,]	-0.89767388	1.70388647	-1.05251337	-1.0486667950
[46,]	-1.25996379	-0.13153881	-1.33575163	-1.1798594716
[47,]	-0.89767388	1.70388647	-1.22245633	-1.3110521482
[48,]	-1.50149039	0.32731751	-1.33575163	-1.3110521482
[49,]	-0.65614727	1.47445831	-1.27910398	-1.3110521482
[50,]	-1.01843718	0.55674567	-1.33575163	-1.3110521482
[51,]	1.39682886	0.32731751	0.53362088	0.2632599711
[52,]	0.67224905	0.32731751	0.42032558	0.3944526477
[53,]	1.27606556	0.09788935	0.64691619	0.3944526477
[54,]	-0.41462067	-1.73753594	0.13708732	0.1320672944
[55,]	0.79301235	-0.59039513	0.47697323	0.3944526477
[56,]	-0.17309407	-0.59039513	0.42032558	0.1320672944
[57,]	0.55148575	0.55674567	0.53362088	0.5256453243
[58,]	-1.13920048	-1.50810778	-0.25944625	-0.2615107354
[59,]	0.91377565	-0.36096697	0.47697323	0.1320672944
[60,]	-0.77691058	-0.81982329	0.08043967	0.2632599711
[61,]	-1.01843718	-2.42582042	-0.14615094	-0.2615107354
[62,]	0.06843254	-0.13153881	0.25038262	0.3944526477
[63,]	0.18919584	-1.96696410	0.13708732	-0.2615107354
[64,]	0.30995914	-0.36096697	0.53362088	0.2632599711
[65,]	-0.29385737	-0.36096697	-0.08950329	0.1320672944
[66,]	1.03453895	0.09788935	0.36367793	0.2632599711
[67,]	-0.29385737	-0.13153881	0.42032558	0.3944526477
[68,]	-0.05233076	-0.81982329	0.19373497	-0.2615107354
[69,]	0.43072244	-1.96696410	0.42032558	0.3944526477
[70,]	-0.29385737	-1.27867961	0.08043967	-0.1303180588
[71,]	0.06843254	0.32731751	0.59026853	0.7880306775
[72,]	0.30995914	-0.59039513	0.13708732	0.1320672944
[73,]	0.55148575	-1.27867961	0.64691619	0.3944526477
[74,]	0.30995914	-0.59039513	0.53362088	0.0008746178
[75,]	0.67224905	-0.36096697	0.30703027	0.1320672944
[76,]	0.91377565	-0.13153881	0.36367793	0.2632599711
[77,]	1.15530226	-0.59039513	0.59026853	0.2632599711
[78,]	1.03453895	-0.13153881	0.70356384	0.6568380009
[79,]	0.18919584	-0.36096697	0.42032558	0.3944526477
[80,]	-0.17309407	-1.04925145	-0.14615094	-0.2615107354
[81,]	-0.41462067	-1.50810778	0.02379201	-0.1303180588
[82,]	-0.41462067	-1.50810778	-0.03285564	-0.2615107354
[83,]	-0.05233076	-0.81982329	0.08043967	0.0008746178
[84,]	0.18919584	-0.81982329	0.76021149	0.5256453243
[85,]	-0.53538397	-0.13153881	0.42032558	0.3944526477

[86,]	0.18919584	0.78617383	0.42032558	0.5256453243
[87,]	1.03453895	0.09788935	0.53362088	0.3944526477
[88,]	0.55148575	-1.73753594	0.36367793	0.1320672944
[89,]	-0.29385737	-0.13153881	0.19373497	0.1320672944
[90,]	-0.41462067	-1.27867961	0.13708732	0.1320672944
[91,]	-0.41462067	-1.04925145	0.36367793	0.0008746178
[92,]	0.30995914	-0.13153881	0.47697323	0.2632599711
[93,]	-0.05233076	-1.04925145	0.13708732	0.0008746178
[94,]	-1.01843718	-1.73753594	-0.25944625	-0.2615107354
[95,]	-0.29385737	-0.81982329	0.25038262	0.1320672944
[96,]	-0.17309407	-0.13153881	0.25038262	0.0008746178
[97,]	-0.17309407	-0.36096697	0.25038262	0.1320672944
[98,]	0.43072244	-0.36096697	0.30703027	0.1320672944
[99,]	-0.89767388	-1.27867961	-0.42938920	-0.1303180588
[100,]	-0.17309407	-0.59039513	0.19373497	0.1320672944
[101,]	0.55148575	0.55674567	1.27004036	1.7063794137
[102,]	-0.05233076	-0.81982329	0.76021149	0.9192233541
[103,]	1.51759216	-0.13153881	1.21339271	1.1816087073
[104,]	0.55148575	-0.36096697	1.04344975	0.7880306775
[105,]	0.79301235	-0.13153881	1.15674505	1.3128013839
[106,]	2.12140867	-0.13153881	1.60992627	1.1816087073
[107,]	-1.13920048	-1.27867961	0.42032558	0.6568380009
[108,]	1.75911877	-0.36096697	1.43998331	0.7880306775
[109,]	1.03453895	-1.27867961	1.15674505	0.7880306775
[110,]	1.63835547	1.24503015	1.32668801	1.7063794137
[111,]	0.79301235	0.32731751	0.76021149	1.0504160307
[112,]	0.67224905	-0.81982329	0.87350679	0.9192233541
[113,]	1.15530226	-0.13153881	0.98680210	1.1816087073
[114,]	-0.17309407	-1.27867961	0.70356384	1.0504160307
[115,]	-0.05233076	-0.59039513	0.76021149	1.5751867371
[116,]	0.67224905	0.32731751	0.87350679	1.4439940605
[117,]	0.79301235	-0.13153881	0.98680210	0.7880306775
[118,]	2.24217198	1.70388647	1.66657392	1.3128013839
[119,]	2.24217198	-1.04925145	1.77986923	1.4439940605
[120,]	0.18919584	-1.96696410	0.70356384	0.3944526477
[121,]	1.27606556	0.32731751	1.10009740	1.4439940605
[122,]	-0.29385737	-0.59039513	0.64691619	1.0504160307
[123,]	2.24217198	-0.59039513	1.66657392	1.0504160307
[124,]	0.55148575	-0.81982329	0.64691619	0.7880306775
[125,]	1.03453895	0.55674567	1.10009740	1.1816087073
[126,]	1.63835547	0.32731751	1.27004036	0.7880306775
[127,]	0.43072244	-0.59039513	0.59026853	0.7880306775
[128,]	0.30995914	-0.13153881	0.64691619	0.7880306775
[129,]	0.67224905	-0.59039513	1.04344975	1.1816087073
[130,]	1.63835547	-0.13153881	1.15674505	0.5256453243
[131,]	1.87988207	-0.59039513	1.32668801	0.9192233541
[132,]	2.48369858	1.70388647	1.49663097	1.0504160307

```

[133,] 0.67224905 -0.59039513 1.04344975 1.3128013839
[134,] 0.55148575 -0.59039513 0.76021149 0.3944526477
[135,] 0.30995914 -1.04925145 1.04344975 0.2632599711
[136,] 2.24217198 -0.13153881 1.32668801 1.4439940605
[137,] 0.55148575 0.78617383 1.04344975 1.5751867371
[138,] 0.67224905 0.09788935 0.98680210 0.7880306775
[139,] 0.18919584 -0.13153881 0.59026853 0.7880306775
[140,] 1.27606556 0.09788935 0.93015445 1.1816087073
[141,] 1.03453895 0.09788935 1.04344975 1.5751867371
[142,] 1.27606556 0.09788935 0.76021149 1.4439940605
[143,] -0.05233076 -0.81982329 0.76021149 0.9192233541
[144,] 1.15530226 0.32731751 1.21339271 1.4439940605
[145,] 1.03453895 0.55674567 1.10009740 1.7063794137
[146,] 1.03453895 -0.13153881 0.81685914 1.4439940605
[147,] 0.55148575 -1.27867961 0.70356384 0.9192233541
[148,] 0.79301235 -0.13153881 0.81685914 1.0504160307
[149,] 0.43072244 0.78617383 0.93015445 1.4439940605
[150,] 0.06843254 -0.13153881 0.76021149 0.7880306775
attr(,"scaled:center")
Sepal.Length Sepal.Width Petal.Length Petal.Width
      5.843333      3.057333      3.758000      1.199333
attr(,"scaled:scale")
Sepal.Length Sepal.Width Petal.Length Petal.Width
      0.8280661      0.4358663      1.7652982      0.7622377

```

Al mostrar por pantalla el contenido de la matriz escalada  $X$ , se pueden ver los valores de las medias aritméticas por columnas, y las desviaciones típicas, también por columnas.

### Cálculo de la matriz de covarianzas

La matriz de covarianzas también se puede calcular de dos maneras, es decir, mediante la definición o mediante la instrucción `cov`:

```
> CX<-t(X)%*%X/149
```

o bien

```
> CX<-cov(X).
```

El resultado es el mismo en ambos casos:

```

      Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length  1.0000000 -0.1175698  0.8717538  0.8179411
Sepal.Width   -0.1175698  1.0000000 -0.4284401 -0.3661259
Petal.Length   0.8717538 -0.4284401  1.0000000  0.9628654
Petal.Width    0.8179411 -0.3661259  0.9628654  1.0000000

```

Además, podemos observar que:

- 1) Los elementos de la diagonal principal son todos igual a 1. Esto es así porque nuestros datos han sido escalados o normalizados y, por lo tanto, la varianza de todos es 1.
- 2) La matriz de covarianzas es una matriz simétrica. Esto es así porque la covarianza es simétrica, es decir:

$$\sigma_{jk}^2 = \sigma_{kj}^2.$$

- 3) Las variables 3 y 4, correspondientes a la longitud y a la anchura del pétalo, están muy relacionadas, puesto que su covarianza es 0.9628654 (muy próxima a 1). Las variables 1 y 3 —longitud del sépalo y del pétalo, respectivamente— también están significativamente relacionadas, aunque con menor medida, puesto que su covarianza es 0.8717538.
- 4) Contrariamente, las variables 1 y 2 —longitud y anchura del sépalo, respectivamente— no están muy relacionadas. En efecto, su covarianza es -0.1175698.

Esta alta correlación se puede observar en las figuras 2 y 3. En particular, con relación a la figura 2, el código de R siguiente:

```
> plot(X[,3], X[,4])
```

es el encargado de generar el diagrama de dispersión de las variables escaladas longitud del pétalo y anchura del pétalo. La nube de puntos correspondiente sigue una línea recta con pendiente positiva. Dicho de otro modo: la nube de puntos se distribuye de forma lineal. Como hemos afirmado antes, esto implica que un aumento en la variable que hemos representado horizontalmente (longitud del pétalo) comporta un aumento en la misma proporción de la variable que hemos representado verticalmente (anchura del pétalo). Ahora bien, este diagrama de dispersión no establece una relación causal. Simplemente, ambas variables tienen una alta correlación.

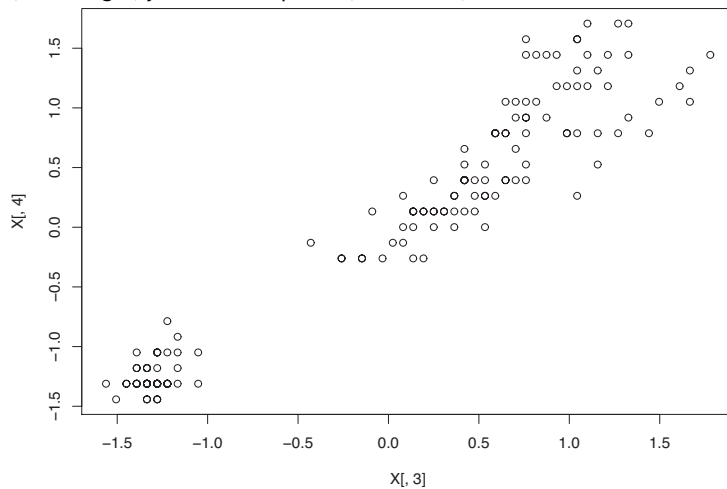
De forma similar, el código R:

```
> plot(X[,1], X[,2])
```

es el encargado de generar el diagrama de dispersión de las variables escaladas longitud del sépalo y anchura del sépalo que recoge la figura 3. Claramente,

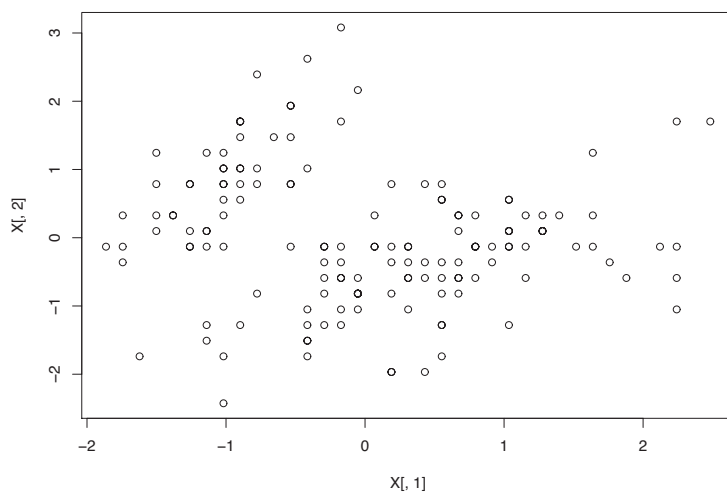
se puede observar que ahora no se establece una tendencia y que la nube de puntos se distribuye por todo el diagrama.

Figura 2. Diagrama de dispersión de las variables escaladas longitud del pétalo (Petal.Length) y anchura del pétalo (Petal.Width).



Fuente: elaboración propia

Figura 3. Diagrama de dispersión de las variables escaladas longitud del sépalo (Sepal.Length) y anchura del sépalo (Sepal.Width).



Fuente: elaboración propia

## Diagonalización de la matriz de covarianzas

Diagonalizar la matriz de covarianzas significa calcular los valores propios y los vectores propios. Con los vectores propios construiremos la matriz llamada  $P$ , que es, en esencia, la matriz asociada a la transformación que nos proyecta los datos originales en el espacio generado por las componentes principales. Las componentes principales son los vectores propios de la matriz de covarianzas. Con R esto se puede hacer de forma muy sencilla:

```
> eigCX<-eigen(CX)
```



La variable `eigCX` es una estructura que contiene los valores propios y los vectores propios asociados:

```
> eigCX

eigen() decomposition
$values
[1] 2.91849782 0.91403047 0.14675688 0.02071484

$vectors
      [,1]      [,2]      [,3]      [,4]
[1,] 0.5210659 -0.37741762 0.7195664 0.2612863
[2,] -0.2693474 -0.92329566 -0.2443818 -0.1235096
[3,] 0.5804131 -0.02449161 -0.1421264 -0.8014492
[4,] 0.5648565 -0.06694199 -0.6342727 0.5235971
```

La variable `eigCX$values` es un vector de cuatro componentes que contiene los valores propios en orden decreciente. Así, `eigCX$vectors` es la matriz **P** en la que cada columna contiene los vectores propios. De hecho, la primera columna es la primera componente principal; la segunda columna, la segunda componente principal, y así sucesivamente. Cabe destacar que **R** calcula los vectores propios unitarios, es decir, con norma 1.

¿Qué pasa si sumamos los valores propios?

```
> sum(eigCX$values)
```

```
[1] 4
```

Obtendremos como resultado 4, un valor que coincide con la traza de la matriz de covarianzas y, de hecho, es el número de variables originales.

Si el total de la varianza, es decir, la suma de los valores propios, es  $\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 4$ , la aportación de la primera componente es:

$$\frac{\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4} \times 100\%.$$

Para nuestro ejemplo, la primera componente principal es capaz de retener un porcentaje de variabilidad igual a:

$$\frac{\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4} \times 100\% = \frac{2.91849782}{4} \times 100\% = 72.96244550\%.$$

De la misma manera, el resto de componentes principales son capaces de retener el siguiente porcentaje de variabilidad:

$$\frac{\lambda_2}{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4} \times 100\% = \frac{0.91403047}{4} \times 100\% = 22.85076175\%$$

$$\frac{\lambda_3}{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4} \times 100\% = \frac{0.14675688}{4} \times 100\% = 3.668922000\%$$

$$\frac{\lambda_4}{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4} \times 100\% = \frac{0.02071484}{4} \times 100\% = 0.5178710000\%$$

Las cuatro componentes principales retienen, respectivamente, un 72.96 %, un 22.85 %, un 3.67 % y un 0.52 %. Es decir, se necesitan tres variables originales (75 % de la información) para obtener, aproximadamente, la misma cantidad de información que con una variable nueva, la primera componente principal. La suma de la variabilidad de las dos primeras componentes principales es 95.81 %. Esto implica que las dos primeras componentes principales son capaces de retener casi toda la información que tenían las cuatro variables originales, ya que solo se pierde un 4.19 % de la información. En otras palabras, la información original necesitaba seis diagramas de dispersión dos a dos para visualizarse. Ahora, basta con un único diagrama de dispersión de las dos primeras componentes principales, que conseguimos con el código siguiente:

```
> P<-eigCX$vector
> T<-X%*%P
> plot(T[1:50,1],T[1:50,2],xlim=c(-3,3),ylim=c(-3,3),
+ col="red", xlab=NULL, ylab=NULL, ann=FALSE)
> par(new=TRUE)
> plot(T[51:100,1],T[51:100,2],xlim=c(-3,3),ylim=c(-3,3),
+ col="blue", xlab=NULL, ylab=NULL, ann=FALSE)
> par(new=TRUE)
> plot(T[101:150,1],T[101:150,2],xlim=c(-3,3),ylim=c(-3,3),
+ col="green", xlab=NULL, ylab=NULL, ann=FALSE)
```

En relación con el código anterior, la instrucción

```
> P<-eigCX$vector
```

sirve para asociar a la variable `P` la matriz que contiene los vectores propios, es decir, la matriz de la transformación lineal asociada a la proyección de los datos originales sobre el espacio vectorial generado por las componentes principales. Con la instrucción

```
> T<-X%*%P
```

calculamos las coordenadas de los datos originales en el espacio vectorial generado por las componentes principales. En otras palabras: son las coordenadas

de los datos proyectados. El número de datos es, por supuesto, igual que el de los datos originales. En este sentido,  $T$  es una matriz que contiene 150 filas, una por flor, y cuatro columnas, una por variable original. Estos 150 puntos del espacio proyectado son:

	[, 1]	[, 2]	[, 3]	[, 4]
[1, ]	-2.25714118	-0.478423832	0.127279624	0.024087508
[2, ]	-2.07401302	0.671882687	0.233825517	0.102662845
[3, ]	-2.35633511	0.340766425	-0.044053900	0.028282305
[4, ]	-2.29170679	0.595399863	-0.090985297	-0.065735340
[5, ]	-2.38186270	-0.644675659	-0.015685647	-0.035802870
[6, ]	-2.06870061	-1.484205297	-0.026878250	0.006586116
[7, ]	-2.43586845	-0.047485118	-0.334350297	-0.036652767
[8, ]	-2.22539189	-0.222403002	0.088399352	-0.024529919
[9, ]	-2.32684533	1.111603700	-0.144592465	-0.026769540
[10, ]	-2.17703491	0.467447569	0.252918268	-0.039766068
[11, ]	-2.15907699	-1.040205867	0.267784001	0.016675503
[12, ]	-2.31836413	-0.132633999	-0.093446191	-0.133037725
[13, ]	-2.21104370	0.726243183	0.230140246	0.002416941
[14, ]	-2.62430902	0.958296347	-0.180192423	-0.019151375
[15, ]	-2.19139921	-1.853846555	0.471322025	0.194081578
[16, ]	-2.25466121	-2.677315230	-0.030424684	0.050365010
[17, ]	-2.20021676	-1.478655729	0.005326251	0.188186988
[18, ]	-2.18303613	-0.487206131	0.044067686	0.092779618
[19, ]	-1.89223284	-1.400327567	0.373093377	0.060891973
[20, ]	-2.33554476	-1.124083597	-0.132187626	-0.037630354
[21, ]	-1.90793125	-0.407490576	0.419885937	0.010884821
[22, ]	-2.19964383	-0.921035871	-0.159331502	0.059398340
[23, ]	-2.76508142	-0.456813301	-0.331069982	0.019582826
[24, ]	-1.81259716	-0.085272854	-0.034373442	0.150636353
[25, ]	-2.21972701	-0.136796175	-0.117599566	-0.269238379
[26, ]	-1.94532930	0.623529705	0.304620475	0.043416203
[27, ]	-2.04430277	-0.241354991	-0.086075649	0.067454082
[28, ]	-2.16133650	-0.525389422	0.206125707	0.010241084
[29, ]	-2.13241965	-0.312172005	0.270244895	0.083977887
[30, ]	-2.25769799	0.336604248	-0.068207276	-0.107918349
[31, ]	-2.13297647	0.502856075	0.074757996	-0.048027970
[32, ]	-1.82547925	-0.422280389	0.269564311	0.239069476
[33, ]	-2.60621687	-1.787587272	-0.047070727	-0.228470534
[34, ]	-2.43800983	-2.143546796	0.082392024	-0.048053409
[35, ]	-2.10292986	0.458665270	0.169706329	0.028926042
[36, ]	-2.20043723	0.205419224	0.224688852	0.168343905
[37, ]	-2.03831765	-0.659349230	0.482919584	0.195702902
[38, ]	-2.51889339	-0.590315163	-0.019370918	-0.136048774
[39, ]	-2.42152026	0.901161067	-0.192609402	-0.009705907
[40, ]	-2.16246625	-0.267981199	0.175296561	0.007023875
[41, ]	-2.27884081	-0.440240541	-0.034778398	0.106626042

[42,]	-1.85191836	2.329610745	0.203552303	0.288896090
[43,]	-2.54511203	0.477501017	-0.304745527	-0.066379077
[44,]	-1.95788857	-0.470749613	-0.308567588	0.176501717
[45,]	-2.12992356	-1.138415464	-0.247604064	-0.150539117
[46,]	-2.06283361	0.708678586	0.063716370	0.139801160
[47,]	-2.37677076	-1.116688691	-0.057026813	-0.151722682
[48,]	-2.38638171	0.384957230	-0.139002234	-0.048671707
[49,]	-2.22200263	-0.994627669	0.180886792	-0.014878291
[50,]	-2.19647504	-0.009185585	0.152518539	0.049206884
[51,]	1.09810244	-0.860091033	0.682300393	0.034717469
[52,]	0.72889556	-0.592629362	0.093807452	0.004887251
[53,]	1.23683580	-0.614239894	0.552157058	0.009391933
[54,]	0.40612251	1.748546197	0.023024633	0.065549239
[55,]	1.07188379	0.207725147	0.396925784	0.104387166
[56,]	0.38738955	0.591302717	-0.123776885	-0.240027187
[57,]	0.74403715	-0.770438272	-0.148472007	-0.077111455
[58,]	-0.48569562	1.846243998	-0.248432992	-0.040384912
[59,]	0.92480346	-0.032118478	0.594178807	-0.029779844
[60,]	0.01138804	1.030565784	-0.537100055	-0.028366154
[61,]	-0.10982834	2.645211115	0.046634215	0.013714785
[62,]	0.43922201	0.063083852	-0.204389093	0.039992104
[63,]	0.56023148	1.758832129	0.763214554	0.045578465
[64,]	0.71715934	0.185602819	0.068429700	-0.164256922
[65,]	-0.03324333	0.437537419	-0.194282030	0.108684396
[66,]	0.87248429	-0.507364239	0.501830204	0.104593326
[67,]	0.34908221	0.195656268	-0.489234095	-0.190869932
[68,]	0.15827980	0.789451008	0.301028700	-0.204612265
[69,]	1.22100316	1.616827281	0.480693656	0.225145511
[70,]	0.16436725	1.298259939	0.172260719	-0.051554138
[71,]	0.73521959	-0.395247446	-0.614467782	-0.083006045
[72,]	0.47469691	0.415926887	0.264067576	0.113189079
[73,]	1.23005729	0.930209441	0.367182178	-0.009911322
[74,]	0.63074514	0.414997441	0.290921638	-0.273304557
[75,]	0.70031506	0.063200094	0.444537765	0.043313222
[76,]	0.87135454	-0.249956017	0.471001057	0.101376117
[77,]	1.25231375	0.076998069	0.724727099	0.039556002
[78,]	1.35386953	-0.330205463	0.259955701	0.066604931
[79,]	0.66258066	0.225173502	-0.085577197	-0.036318171
[80,]	-0.04012419	1.055183583	0.318506304	0.064571834
[81,]	0.13035846	1.557055553	0.149482697	-0.009371129
[82,]	0.02337438	1.567225244	0.240745761	-0.032663020
[83,]	0.24073180	0.774661195	0.150707074	0.023572390
[84,]	1.05755171	0.631726901	-0.104959762	-0.183354200
[85,]	0.22323093	0.286812663	-0.663028512	-0.253977520
[86,]	0.42770626	-0.842758920	-0.449129446	-0.109308985
[87,]	1.04522645	-0.520308714	0.394464890	0.037084781
[88,]	1.04104379	1.378371048	0.685997804	0.136378719

[89,]	0.06935597	0.218770433	-0.290605718	-0.146653279
[90,]	0.28253073	1.324886147	-0.089111491	0.008876070
[91,]	0.27814596	1.116288852	-0.094172116	-0.269753497
[92,]	0.62248441	-0.024839814	0.020412763	-0.147193289
[93,]	0.33540673	0.985103828	0.198724011	0.006508757
[94,]	-0.36097409	2.012495825	-0.105467721	0.019505467
[95,]	0.28762268	0.852873116	-0.130452657	-0.107043742
[96,]	0.09105561	0.180587142	-0.128547696	-0.229191812
[97,]	0.22695654	0.383634868	-0.155691572	-0.132163118
[98,]	0.57446378	0.154356489	0.270743347	-0.019794366
[99,]	-0.44617230	1.538637456	-0.189765199	0.199278855
[100,]	0.25587339	0.596852285	-0.091572385	-0.058426315
[101,]	1.83841002	-0.867515056	-1.002044077	-0.049085303
[102,]	1.15401555	0.696536401	-0.528389994	-0.040385459
[103,]	2.19790361	-0.560133976	0.202236658	0.058986583
[104,]	1.43534213	0.046830701	-0.163083761	-0.234982858
[105,]	1.86157577	-0.294059697	-0.394307408	-0.016243853
[106,]	2.74268509	-0.797736709	0.580364827	-0.101045973
[107,]	0.36579225	1.556289178	-0.983598122	-0.132679346
[108,]	2.29475181	-0.418663020	0.649530452	-0.237246445
[109,]	1.99998633	0.709063226	0.392675073	-0.086221779
[110,]	2.25223216	-1.914596301	-0.396224508	0.104488870
[111,]	1.35962064	-0.690443405	-0.283661780	0.107500284
[112,]	1.59732747	0.420292431	-0.023108991	0.058136869
[113,]	1.87761053	-0.417849815	-0.026250468	0.145926073
[114,]	1.25590769	1.158379741	-0.578311891	0.098826244
[115,]	1.46274487	0.440794883	-1.000517746	0.274738504
[116,]	1.58476820	-0.673986887	-0.636297054	0.191222383
[117,]	1.46651849	-0.254768327	-0.037306280	-0.154811637
[118,]	2.41822770	-2.548124795	0.127454475	-0.272892966
[119,]	3.29964148	-0.017721580	0.700957033	0.045037725
[120,]	1.25954707	1.701046715	0.266643612	-0.064963167
[121,]	2.03091256	-0.907427443	-0.234015510	0.167390481
[122,]	0.97471535	0.569855257	-0.825362161	0.027662914
[123,]	2.88797650	-0.412259950	0.854558973	-0.126911337
[124,]	1.32878064	0.480202496	0.005410239	0.139491837
[125,]	1.69505530	-1.010536476	-0.297454114	-0.061437911
[126,]	1.94780139	-1.004412720	0.418582432	-0.217609339
[127,]	1.17118007	0.315338060	-0.129503907	0.125001677
[128,]	1.01754169	-0.064131184	-0.336588365	-0.008625505
[129,]	1.78237879	0.186735633	-0.269754304	0.030983849
[130,]	1.85742501	-0.560413289	0.713244682	-0.207519953
[131,]	2.42782030	-0.258418706	0.725386035	-0.017863520
[132,]	2.29723178	-2.617554417	0.491826144	-0.210968943
[133,]	1.85648383	0.177953334	-0.352966242	0.099675959
[134,]	1.11042770	0.291944582	0.182875741	-0.185721512
[135,]	1.19845835	0.808606364	0.164173760	-0.487849130

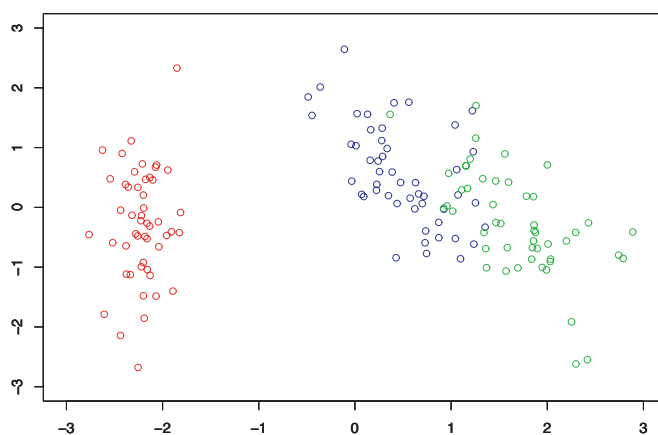
```
[136,] 2.78942561 -0.853942542 0.541093785 0.294893130
[137,] 1.57099294 -1.065013214 -0.942695700 0.035486875
[138,] 1.34179696 -0.421020154 -0.180271551 -0.214702016
[139,] 0.92173701 -0.017165594 -0.415434449 0.005220919
[140,] 1.84586124 -0.673870645 0.012629804 0.194543500
[141,] 2.00808316 -0.611835930 -0.426902678 0.246711805
[142,] 1.89543421 -0.687273065 -0.129640697 0.468128374
[143,] 1.15401555 0.696536401 -0.528389994 -0.040385459
[144,] 2.03374499 -0.864624030 -0.337014969 0.045036251
[145,] 1.99147547 -1.045665670 -0.630301866 0.213330527
[146,] 1.86425786 -0.385674038 -0.255418178 0.387957152
[147,] 1.55935649 0.893692855 0.026283300 0.219456899
[148,] 1.51609145 -0.268170747 -0.179576781 0.118773236
[149,] 1.36820418 -1.007877934 -0.930278721 0.026041407
[150,] 0.95744849 0.024250427 -0.526485033 -0.162533529
```

A continuación, con el código

```
> plot(T[1:50,1],T[1:50,2],xlim=c(-3,3),ylim=c(-3,3),
+ col="red", xlab = NULL, ylab=NULL, ann = FALSE)
> par(new=TRUE)
```

representamos gráficamente los primeros cincuenta puntos en el plano generado por las dos primeras componentes principales. Con las opciones `xlim` e `ylim` definimos los rangos horizontal y vertical, respectivamente. La opción `col="red"` sirve para representar estos puntos de color rojo. Con las opciones `xlab = NULL` e `ylab = NULL` no declaramos ninguna etiqueta para los ejes y con `ann = FALSE` también evitamos que R ponga alguna anotación por defecto. Finalmente, el código `par (new=TRUE)` sirve para “congelar” el gráfico generado, a la espera de representar gráficamente el resto de puntos. El resultado se puede ver en la figura 4.

Figura 4. Diagrama de dispersión de las dos primeras componentes principales. Los colores representan el tipo o especie de flor: *setosa* (rojo), *versicolor* (azul) y *virginica* (verde).



Fuente: elaboración propia

Esta figura, donde hemos representado gráficamente las dos primeras componentes principales de las 150 flores, muestra claramente tres nubes de puntos:

- La nube de puntos de color rojo, que corresponde a la especie *setosa*.
- La nube de puntos de color azul, que corresponde a la especie *versicolor*.
- La nube de puntos de color verde, que corresponde a la especie *virginica*.

Gracias al análisis de componentes principales, las 150 flores han quedado representadas en un único diagrama de dispersión: el eje horizontal incluye la primera componente principal, y el eje vertical, la segunda. El conjunto de puntos correspondiente a *setosa* (rojo) ha quedado perfectamente clusterizado. En el caso de los conjuntos de puntos correspondientes a las especies *versicolor* (azul) y *virginica* (verde), ambos clústeres tienen alguna intersección; sin embargo, la diferencia es claramente visible.

En muchos casos, el análisis de componentes principales permite descubrir patrones, clústeres o clasificaciones visuales que, de otro modo, serían muy difíciles de detectar. Por lo tanto, el análisis de componentes principales es una técnica de la ciencia de datos muy extendida para detectar defectos o para clasificar conjuntos de datos multidimensionales.

Encontraréis dos ejemplos muy interesantes sobre la aplicación del análisis de componentes principales en la detección de defectos en estructuras en estos artículos:

- **L. E. Mujica; J. Rodellar; A. Fernández; A. Güemes** (2011). “Q-statistic and T2-statistic PCA-based measures for damage assessment in structures”. *Structural Health Monitoring* (vol. 10, n.º 5, págs. 539-553).
- **L. E. Mujica; M. Ruiz; F. Pozo; J. Rodellar; A. Güemes** (2013). “A structural damage detection indicator based on principal component analysis and statistical hypothesis testing”. *Smart Materials and Structures* (vol. 23, n.º 2).

En el primer artículo, el análisis de componentes principales basta para clasificar los diferentes tipos de daños que existen en la estructura que se considera. En el segundo caso, aparte del análisis de componentes principales, hay que aplicar el contraste de hipótesis para poder clasificar el tipo de daño.

El capítulo “Damage and fault detection of structures using principal component analysis and hypothesis testing”, del libro *Advances in Principal Component Analysis* (págs. 137-191), presenta el estado del arte de la aplicación del análisis de componentes principales en el campo de la monitorización de la integridad estructural, así como en la monitorización de la condición, es decir, detección, identificación y clasificación de daños o errores en estructuras diversas.

## Análisis de componentes principales y R

El código que hemos usado antes para calcular las componentes principales y la proyección de los datos originales sobre el espacio vectorial generado por estas componentes se puede simplificar si utilizamos directamente la instrucción que ofrece R. En efecto, con el código:

```
> data(`iris`)
> iris.pca<-prcomp(iris[,1:4],center = TRUE, scale = TRUE)
```

cargamos de nuevo los datos denominados *iris* y, mediante la instrucción `prcomp`, calculamos el análisis de componentes principales. Notad que es fundamental añadir las opciones `center = TRUE` y `scale = TRUE` para escalar los datos. ¿Qué contiene la estructura `iris.pca`? Lo veremos a continuación.

```
> iris.pca

Standard deviations (1, ..., p=4):
[1] 1.7083611 0.9560494 0.3830886 0.1439265

Rotation (n x k) = (4 x 4):
               PC1          PC2          PC3
Sepal.Length  0.5210659 -0.37741762  0.7195664
Sepal.Width   -0.2693474 -0.92329566 -0.2443818
Petal.Length  0.5804131 -0.02449161 -0.1421264
Petal.Width   0.5648565 -0.06694199 -0.6342727
               PC4
Sepal.Length  0.2612863
Sepal.Width   -0.1235096
Petal.Length -0.8014492
Petal.Width   0.5235971
```

La estructura `iris.pca` tiene cinco campos:

- 1) `sdev`, un vector que contiene las raíces cuadradas de los valores propios de la matriz de covarianzas;
- 2) `rotation`, una matriz que contiene las componentes principales y, por lo tanto, una matriz que en el módulo “Descomposición en valores singulares: introducción y aplicaciones” hemos denominado **P**;
- 3) `center`, un vector que contiene la media aritmética de cada variable en el conjunto de datos original;
- 4) `scale`, un vector que contiene la desviación típica de cada variable en el conjunto de datos original;
- 5) `x`, una matriz que contiene las proyecciones de los datos originales sobre el espacio vectorial generado por las componentes principales y, por lo tanto, una matriz que en el módulo “Descomposición en valores singulares: introducción y aplicaciones” hemos denominado **T**.

Así pues, para hacer un diagrama de dispersión como el de la figura 4, se debería usar este código:



```

> data(`iris`)
> iris.pca<-prcomp(iris[,1:4],center = TRUE, scale = TRUE)
> T<-iris.pca$x # definimos la variable T
> plot(T[1:50,1],T[1:50,2],xlim=c(-3,3),ylim=c(-3,3),
+ col="red", xlab = NULL, ylab=NULL, ann = FALSE)
> par(new=TRUE)
> plot(T[51:100,1],T[51:100,2],xlim=c(-3,3),ylim=c(-3,3),
+ col="blue", xlab=NULL, ylab=NULL, ann = FALSE)
> par(new=TRUE)
> plot(T[101:150,1],T[101:150,2],xlim=c(-3,3),ylim=c(-3,3),
+ col="green", xlab=NULL, ylab=NULL, ann = FALSE)

```

También hay muchas opciones en R para representar gráficamente las proyecciones sobre el espacio vectorial de las dos primeras componentes principales. A continuación adjuntamos una que necesita cargar la librería `ggfortify`:

```

> data(`iris`)
> df<-iris[,1:4]
> df2<-iris
> library(ggfortify)
> iris.pca<-prcomp(df,center = TRUE, scale=TRUE)
> autoplot(iris.pca,data = df2,colour='Species',
+ loadings = TRUE, loadings.label = TRUE,
+ loadings.label.size = 3, scale = 0)

```

El resultado se puede observar en la figura 5. Del código hay que destacar el uso de la instrucción `autoplot`, con las opciones que comentaremos a continuación:

- `data=df2` contiene los datos originales, incluyendo la quinta columna correspondiente a las etiquetas (en este caso, la especie asociada a cada punto);
- `color='Species'` agrupa por colores los puntos de la misma especie;
- `loadings = TRUE` incorpora los vectores con la influencia de las dos componentes principales en las cuatro variables originales;
- `loadings.label = TRUE` incorpora los nombres de las variables originales en el diagrama de dispersión;
- `loading.label.size` establece la medida de los nombres de las variables originales;
- `scale = 0` mantiene la escala de los datos originales que ya han sido escalados en la creación del modelo del análisis de componentes principales.

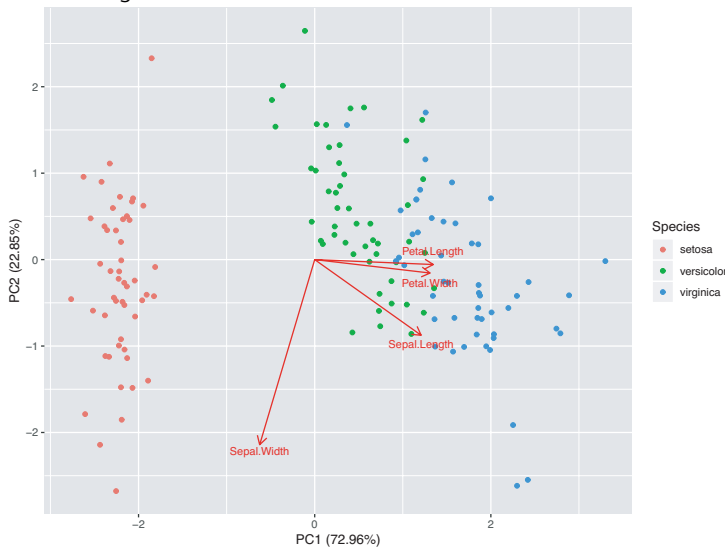
Observando la figura 5 también podemos extraer la información visual siguiente:

**1)** Se puede observar la contribución de cada una de las cuatro variables originales a las dos primeras componentes principales. Si nos fijamos, por ejemplo, en la primera componente principal (la dirección horizontal), las variables que más intervienen son la longitud del sépalo (`Sepal.Length`) y la longitud y anchura del pétalo (`Petal.Length`, `Petal.Width`). En el caso de la segunda com-

ponente principal (la dirección vertical), la variable que tiene más peso es la anchura del sépalo (Sepal.Width).

2) También vemos que las flechas que indican las direcciones de las variables sobre longitud y anchura del pétalo (Petal.Length, Petal.Width) están prácticamente superpuestas. Recordemos que, en este caso, la covarianza entre estas dos variables es 0.9628654, que representa un valor muy próximo a 1. Es decir, como habíamos dicho, estas variables están altamente relacionadas.

Figura 5. Diagrama de dispersión de las dos primeras componentes principales. Los colores representan el tipo o especie de flor: *setosa* (rojo), *versicolor* (verde) y *virginica* (azul). Además, se incluye la influencia de las dos primeras componentes principales en las cuatro variables originales.



Fuente: elaboración propia

### Clasificación a partir del análisis de componentes principales

Veamos un ejemplo sencillo de aplicación del análisis que hemos llevado a cabo hasta ahora para clasificar una nueva flor desconocida. Imaginemos, pues, que ahora tenemos una flor con las características que recoge la tabla 2.

Tabla 2. Longitud y anchura del sépalo y longitud y anchura del pétalo (en centímetros)

flor	long. sépalo	anchura sépalo	long. pétalo	anchura pétalo	tipo
106	6.0	2.8	4.3	1.3	desconocido

Fuente: elaboración propia

¿Cómo podemos saber a qué especie pertenece esta flor?

1) Escalamos los valores de longitud y anchura del sépalo y de longitud y anchura del pétalo, en relación con las medias aritméticas y las desviaciones típicas de los datos originales (150 flores). Denominamos  $\mathbf{u}^T$  al vector que contiene estos cuatro valores normalizados.

2) Proyectamos este vector sobre el espacio vectorial generado por las componentes principales con el producto

$$\mathbf{t}^T = \mathbf{u}^T \mathbf{P},$$

y denominamos  $\mathbf{t}^T$  al resultado.

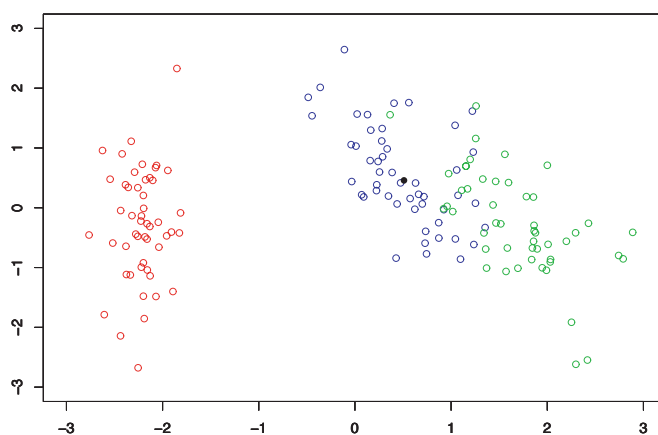
3) Añadimos las dos primeras componentes de este punto al diagrama de dispersión original y determinamos a qué clúster pertenece.

El código de R que hay que usar es:

```
> data(`iris`')
> iris.pca<-prcomp(df,center = TRUE, scale=TRUE)
> T<-iris.pca$x
> f<-c(6.0,2.8,4.3,1.3)
> f<-(f-iris.pca$center)/iris.pca$scale # escalado de los datos
> u<-f%*%iris.pca$rotation # proyección de los datos escalados
> plot(T[1:50,1],T[1:50,2],xlim=c(-3,3),ylim=c(-3,3),
+ col="red", xlab = NULL, ylab=NULL, ann = FALSE)
> par(new=TRUE)
> plot(T[51:100,1],T[51:100,2],xlim=c(-3,3),ylim=c(-3,3),
+ col="blue", xlab=NULL, ylab=NULL, ann = FALSE)
> par(new=TRUE)
> plot(T[101:150,1],T[101:150,2],xlim=c(-3,3),ylim=c(-3,3),
+ col="green", xlab=NULL, ylab=NULL, ann = FALSE)
> par(new=TRUE)
> plot(u[1],u[2],xlim=c(-3,3),ylim=c(-3,3),col="black",
+ pch=16, xlab=NULL, ylab=NULL, ann = FALSE)
```

Podemos ver el resultado en la figura 6. El punto negro corresponde al punto que tenemos que identificar. A partir de una inspección visual se podría afirmar que este punto pertenece o es más próximo al conjunto o clúster de puntos azules. Por lo tanto, clasificaríamos el punto (la flor) como *versicolor*.

Figura 6. Diagrama de dispersión de las dos primeras componentes principales. Los colores representan al tipo o especie de flor: *setosa* (rojo), *versicolor* (azul) y *virginica* (verde). El punto negro es el que debemos clasificar.



Fuente: elaboración propia

## Bibliografía

**Mujica, L. E.; Rodellar, J.; Fernández, A.; Güemes, A.** (2011). "Q-statistic and  $T^2$ -statistic PCA-based measures for damage assessment in structures". *Structural Health Monitoring* (vol. 10, n.º 5, págs. 539-553).

**Mujica, L. E.; Ruiz, M.; Pozo, F.; Rodellar, J.; Güemes, A.** (2013). "A structural damage detection indicator based on principal component analysis and statistical hypothesis testing". *Smart Materials and Structures* (vol. 23, n.º 2).

**Pozo, F.; Vidal, Y.** (2018). *Damage and fault detection of structures using principal component analysis and hypothesis testing. Advances in Principal Component Analysis* (págs. 137-191). Singapur: Springer.