Descomposició en valors singulars: introducció i aplicacions

Estudi de cas i guia de resolució en R

PID_00262390

Francesc Pozo Montero
Jordi Ripoll Missé

Universitat Oberta de Catalunya

Francesc Pozo Montero

Llicenciat en Matemàtiques per la Universitat de Barcelona (2000) i doctor en Matemàtica Aplicada per la Universitat Politècnica de Catalunya (2005). Ha estat professor associat a la Universitat Autònoma de Barcelona i professor associat, col·laborador i actualment professor agregat a la Universitat Politècnica de Catalunya. A més, és cofundador del Grup d'Innovació Matemàtica E-learning (GIMEL), responsable de diversos projectes d'innovació docent i autor de diverses publicacions. Com a membre del grup de recerca consolidat CoDAlab, centra la recerca en la teoria de control i les aplicacions en enginyeria mecànica i civil, com també en l'ús de la ciència de dades per al monitoratge de la integritat estructural i per al monitoratge de la condició, sobretot en turbines eòliques.

Jordi Ripoll Missé

Llicenciat en Matemàtiques i doctor en Ciències Matemàtiques per la Universitat de Barcelona (2005). Professor col·laborador de la Universitat Oberta de Catalunya des del 2011 i professor del Departament d'Informàtica, Matemàtica Aplicada i Estadística de la Universitat de Girona (UdG) des del 1996, on actualment és professor agregat i desenvolupa tasques de recerca en l'àmbit de la biologia matemàtica (models amb equacions en derivades parcials i dinàmica evolutiva). També ha estat professor i tutor de la UNED en dues etapes, primer al centre associat de Terrassa i actualment al de Girona. Ha participat en nombrosos projectes d'innovació docent, especialment pel que fa a l'aprenentatge de les matemàtiques en línia.

L'encàrrec i la creació d'aquest recurs d'aprenentatge UOC han estat coordinats per la professora: Cristina Cano Bastidas (2019)

Primera edició: febrer 2019

Av. Tibidabo, 39-43, 08035 Barcelona
Disseny: Manel Andreu
Realització editorial: Oberta UOC Publishing, SL

Índex

1. Estudi de cas

Iris

Les dades van ser recollides per Edgar Anderson l'any 1935 i publicades en l'article "The irises of the Gaspe Peninsula", *Bulletin of the American Iris Society*, 59, p. 2-5.

Per a aquest estudi de cas tindrem en compte el conjunt de dades Iris que ja hem treballat en el mòdul "Descomposició en valors singulars: introducció i aplicacions". Aquestes dades proporcionen les mesures en centímetres de les variables de longitud i amplada del sèpal i de longitud i amplada del pètal, respectivament, per a cinquanta flors de cadascuna de les espècies *Iris setosa*, *Iris versicolor* i *Iris virginica*. En el mòdul "Descomposició en valors singulars: introducció i aplicacions" hem considerat una mostra de quinze flors, cinc de cada tipus. En aquest estudi de cas considerem la totalitat de la mostra. Els valors d'aquestes variables per a les cinc primeres flors de cada tipus són a la taula 1.

Taula 1. Longitud i amplada del sèpal i longitud i amplada del pètal (en centímetres)

flor	long. sèpal	ampl. sèpal	long. pètal	ampl. pètal	tipus
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*
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*
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*

Font: Edgar Anderson (1935). "The irises of the Gaspe Peninsula"

En aquest estudi de cas emprarem el programari lliure `R` per fer l'anàlisi de les components principals. Ho farem, a més, de dues maneres: *a*) seguint els passos de la tècnica i *b*) aplicant directament la instrucció `prcomp` i analitzant i interpretant els resultats.

En primer lloc, carreguem el conjunt de dades `iris` i en mostrem el resultat:

```
> 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
```

Per a les 150 flors que hem considerat, podem veure el valor de les variables:

1) longitud del sèpal
2) amplada del sèpal
3) longitud del pètal
4) amplada del pètal

També podem veure una cinquena columna, corresponent al tipus de flor o espècie (*setosa*, *versicolor* o *virginica*).

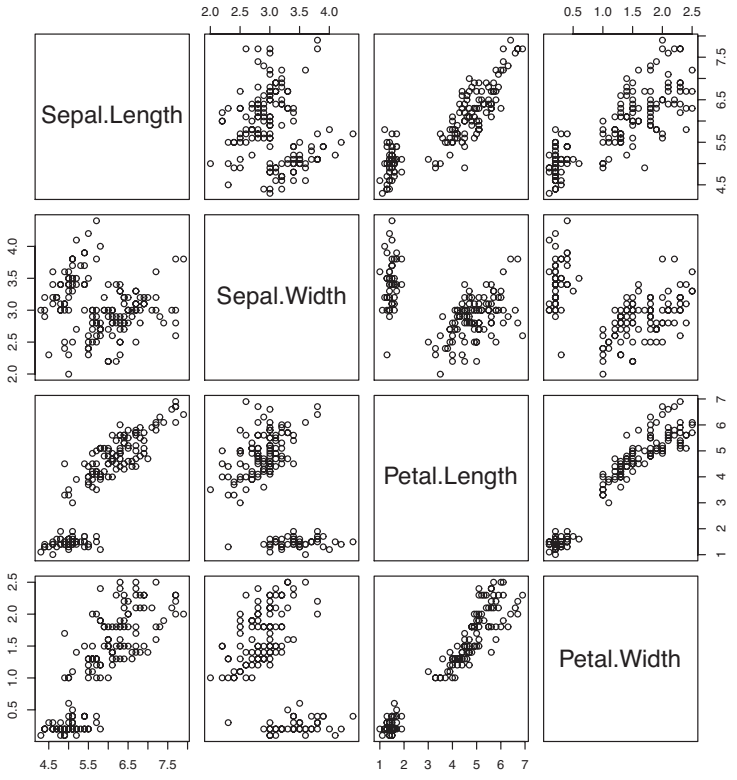
Una primera aproximació a l'estudi d'aquest conjunt de dades podria ser fer els diagrames de dispersió per a tots els parells de variables numèriques, és a dir, eliminant la variable que ens indica el tipus de flor:

Diagrama de dispersió

Un diagrama de dispersió (en anglès, *scatter plot*) mostra gràficament la relació entre dues variables quantitatives.

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

Figura 1. Diagrama de dispersió de les variables longitud del sèpal (Sepal. Length), amplada del sèpal (Sepal. Width), longitud del pètal (Petal. Length) i amplada del pètal (Petal. Width).



Font: elaboració pròpia

Fixeu-vos que en la figura 1 hem generat una matriu de dimensió 4×4 amb la instrucció `plot(iris[,1:4])`. A la diagonal principal surt el nom de les quatre variables. A la resta dels elements de la matriu de diagrames de dispersió hi ha els diagrames de dispersió de cada parell de variables. Per exemple, a la fila 1, columna 2 trobem el diagrama de dispersió de la variable de longitud del sèpal *versus* la variable d'amplada del sèpal. A la fila 2, columna 1 trobem el mateix diagrama de dispersió però canviant els eixos, és a dir, com a variable horitzontal hi ha l'amplada del sèpal i com a variable vertical hi ha la longitud del sèpal.

Una segona observació de la figura 1 ens permet veure que:

1) Hi ha una forta correlació entre les variables de longitud del pètal (Petal. Length) i d'amplada del pètal (Petal. Width). Aquesta correlació és observable perquè el núvol de punts marca una certa línia recta de pendent positiu. Això es tradueix en el fet que, quan la longitud del pètal creix, l'amplada del pètal també creix, i viceversa.

2) En contraposició a l'observació anterior, el núvol de punts corresponent a les variables de longitud del sèpal (Sepal. Length) i d'amplada del sèpal (Sepal. Width) ja no marca una tendència tan clara. De fet, sembla que els punts estiguin distribuïts de manera aleatòria.

Llevat d'aquestes dues observacions, no es poden afirmar amb certesa gaires coses més. Per tant, podrem extreure'n poques conclusions. Amb tot, els càlculs que farem després ens ajudaran a confirmar o a desestimar aquestes impressions.

Recordeu que en el procediment d'anàlisi de les components principals intervenen aquestes etapes:

1) escalat de dades,

2) càlcul de la matriu de covariàncies,

3) diagonalització de la matriu de covariàncies, que inclou el càlcul dels valors i vectors propis.

Escalat de dades

Per a l'escalat de dades farem servir la instrucció `scale` de la manera següent:

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

Observeu que guardem en la variable `X` les variables numèriques del conjunt de dades. L'opció `center = TRUE` significa que a tots els elements d'una mateixa columna els restarem la seva mitjana aritmètica. L'opció `scale = TRUE` significa que tots els elements d'una mateixa columna seran dividits per la seva desviació tipus. El resultat és una matriu de 150 files i 4 columnes:

```
      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
```

En mostrar per pantalla el contingut de la matriu esglaonada `X`, es poden veure els valors de les mitjanes aritmètiques per columnes, i les desviacions típiques, també per columnes.

Càlcul de la matriu de covariàncies

La matriu de covariàncies també es pot calcular de dues maneres, és a dir, per mitjà de la definició o per mitjà de la instrucció `cov`:

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

o bé

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

El resultat és el mateix en tots dos 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
```

A més, podem observar que:

1) Els elements de la diagonal principal són tots igual a 1. Això és així perquè les nostres dades han estat escalades o normalitzades i, per tant, la variància de totes és 1.

2) La matriu de covariàncies és una matriu simètrica. Això és així perquè la covariància és simètrica, és a dir:

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

3) Les variables 3 i 4, corresponents a la longitud i a l'amplada del pètal, estan molt relacionades, ja que la seva covariància és 0.9628654 (molt propera a 1). Les variables 1 i 3 —longitud del sèpal i del pètal, respectivament— també estan significativament relacionades, tot i que amb menor proporció, ja que la seva covariància és 0.8717538.

4) Contràriament, les variables 1 i 2 —longitud i amplada del sèpal, respectivament— no estan gaire relacionades. En efecte, la seva covariància és –0.1175698.

Aquesta alta correlació es pot observar a les figures 2 i 3. En particular, amb relació a la figura 2, el codi de R següent

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

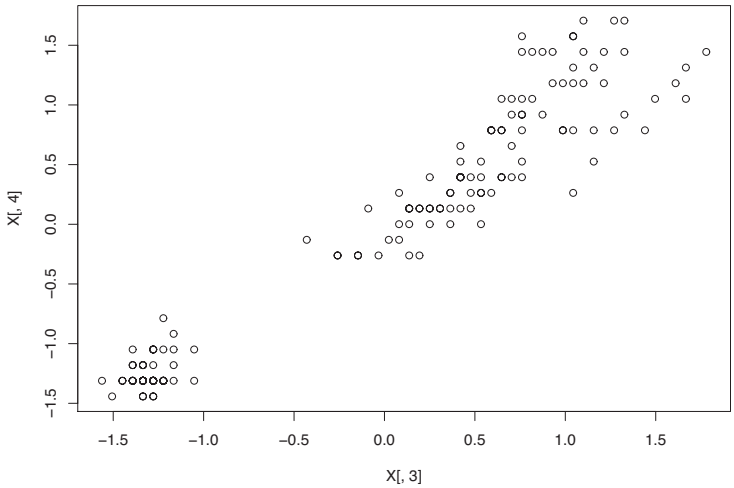
és l'encarregat de generar el diagrama de dispersió de les variables escalades sobre la longitud del pètal i l'amplada del pètal. El núvol de punts corresponent segueix una línea recta amb pendent positiu. Dit d'una altra manera: el núvol de punts es distribueix de forma lineal. Com hem afirmat abans, això implica que un augment en la variable que hem representat horitzontalment (longitud del pètal) comporta un augment en la mateixa proporció de la variable que hem representat verticalment (amplada del pètal). Ara bé, aquest diagrama de dispersió no estableix una relació causal. Simplement, totes dues variables tenen una alta correlació.

De forma similar, el codi R:

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

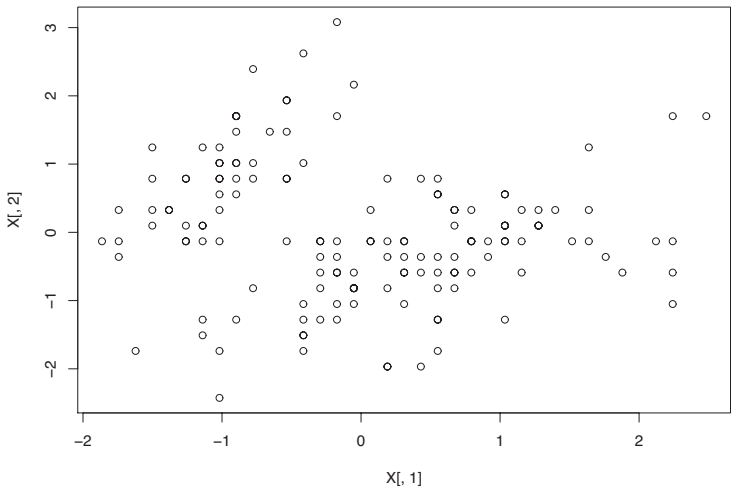
és l'encarregat de generar el diagrama de dispersió de les variables escalades sobre la longitud del sèpal i l'amplada del sèpal que recull la figura 3. Clarament, es pot observar que ara no s'estableix una tendència i que el núvol de punts es distribueix per tot el diagrama.

Figura 2. Diagrama de dispersió de les variables escalades sobre la longitud del pètal (Petal. Length) i l'amplada del pètal (Petal. Width).



Font: elaboració pròpia

Figura 3. Diagrama de dispersió de les variables escalades sobre la longitud del sèpal (Sepal. Length) i l'amplada del sèpal (Sepal. Width).



Font: elaboració pròpia

Diagonalització de la matriu de covariàncies

Diagonalitzar la matriu de covariàncies significa calcular-ne els valors propis i els vectors propis. Amb els vectors propis construirem la matriu anomenada **P**, que és, en essència, la matriu associada a la transformació que ens projecta les dades originals a l'espai generat per les components principals. Les components principals són els vectors propis de la matriu de covariàncies. Amb R això es pot fer d'una manera molt senzilla:

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

La variable `eigCX` és una estructura que conté els valors propis i els vectors propis associats:

```
> 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$values` és un vector de quatre components que conté els valors propis en ordre decreixent. De la mateixa manera, `eigCX$vectors` és la matriu $\mathbf{P}$ en què cada columna conté els vectors propis. De fet, la primera columna és la primera component principal; la segona columna, la segona component principal, i així successivament. Cal destacar que `R` calcula els vectors propis unitaris, és a dir, amb norma 1.

Què passa si sumem els valors propis?

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

```
[1] 4
```

Obtindrem com a resultat 4, un valor que coincideix amb la traça de la matriu de covariàncies i, de fet, és el nombre de variables originals.

Si el total de la variància, és a dir, la suma dels valors propis, és $\lambda_1+\lambda_2+\lambda_3+\lambda_4 = 4$, l'aportació de la primera component és:

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

Per al nostre exemple, la primera component principal és capaç de retenir un percentatge de variabilitat 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 mateixa manera, la resta de les components principals són capaces de retenir el percentatge de variabilitat següent:

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

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

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

Les quatre components principals retenen un 72.96%, un 22.85%, un 3.67% i un 0.52%, respectivament. És a dir, calen tres variables originals (75% de la informació) per obtenir, aproximadament, la mateixa quantitat d'informació que amb una sola variable nova, la primera component principal. La suma de la variabilitat de les dues primeres components principals és 95.81%. Això implica que les dues primeres components principals són capaces de retenir gairebé tota la informació que tenien les quatre variables originals, ja que només es perd un 4.19% de la informació. Dit d'una altra manera, la informació original necessitava sis diagrames de dispersió dos a dos per visualitzar-se. Ara, n'hi ha prou amb un únic diagrama de dispersió de les dues primeres components principals, que aconseguim amb el codi següent:

```
> P<-eigCX$vectors
> 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 relació amb el codi anterior, la instrucció

```
> P<-eigCX$vectors
```

serveix per associar a la variable `P` la matriu que conté els vectors propis, és a dir, la matriu de la transformació lineal associada a la projecció de les dades originals sobre l'espai vectorial generat per les components principals. En la instrucció

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

calculem les coordenades de les dades originals en l'espai vectorial generat per les components principals. Dit d'una altra manera: són les coordenades de les

dades projectades. El nombre de dades és, per descomptat, igual que el de les dades originals. En aquest sentit, `T` és una matriu que conté 150 files, una per flor, i quatre columnes, una per variable original. Aquests 150 punts de l'espai projectat són:

```
             [,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
```

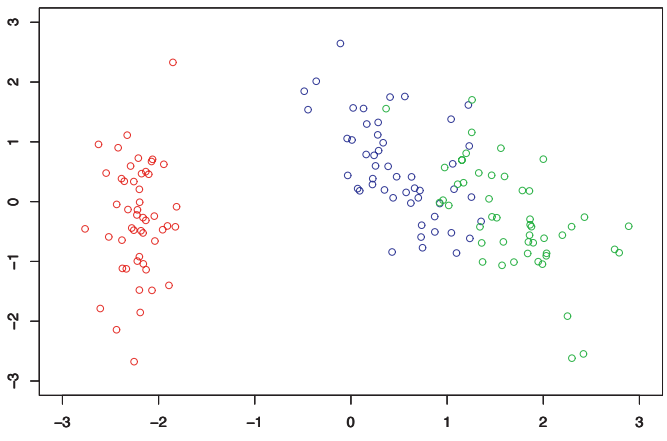
```
[135,]  1.19845835  0.808606364  0.164173760 -0.487849130
[136,]  2.78942561 -0.853942542  0.541093785  0.294893130
[137,]  1.57099294 -1.065013214 -0.942695700  0.035486875
[138,]  1.34179696 -0.421020154 -0.180271551 -0.214702016
[139,]  0.92173701 -0.017165594 -0.415434449  0.005220919
[140,]  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 continuació, amb el codi

```
> 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)
```

representem gràficament els primers cinquanta punts en el pla generat per les dues primeres components principals. Amb les opcions `xlim` i `ylim` definim els rangs horitzontal i vertical, respectivament. L'opció `col="red"` serveix per representar aquests punts de color vermell. Amb les opcions `xlab = NULL` i `ylab = NULL` no declarem cap etiqueta pels eixos i amb `ann = FALSE` també evitem que `R` posi cap anotació per defecte. Finalment, el codi `par (new=TRUE)` serveix per "congelar" el gràfic generat, a l'espera de representar gràficament la resta dels punts. El resultat es pot veure a la figura 4.

Figura 4. Diagrama de dispersió de les dues primeres components principals. Els colors representen el tipus o espècie de flor: *setosa* (vermell), *versicolor* (blau) i *virginica* (verd).



Font: elaboració pròpia

Aquesta figura, on hem representat gràficament les dues primeres components principals de les 150 flors, mostra clarament tres núvols de punts:

- El núvol de punts de color vermell, que correspon a l'espècie *setosa*.
- El núvol de punts de color blau, que correspon a l'espècie *versicolor*.
- El núvol de punts de color verd, que correspon a l'espècie *virginica*.

Gràcies a l'anàlisi de components principals, les 150 flors han quedat representades en un únic diagrama de dispersió: l'eix horitzontal inclou la primera component principal, i l'eix vertical, la segona. El conjunt de punts corresponent a *setosa* (vermell) ha quedat perfectament clusteritzat. En el cas dels conjunts de punts corresponents a les espècies *versicolor* (blau) i *virginica* (verd), tots dos clústers tenen alguna intersecció; amb tot, la diferència és clarament visible.

En molts casos, l'anàlisi de components principals permet descobrir patrons, clústers o classificacions visuals que, altrament, serien molt difícils de detectar. Per tant, l'anàlisi de components principals és una tècnica de la ciència de dades molt estesa per detectar defectes o per classificar conjunts de dades multidimensionals.

Trobareu dos exemples molt interessants sobre l'aplicació de l'anàlisi de components principals en la detecció de defectes en estructures en aquests articles:

- **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, núm. 5, pàg. 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, núm. 2).

En el primer article, l'anàlisi de components principals és suficient per classificar els diferents tipus de danys que hi ha en l'estructura que es considera. En el segon cas, a part de l'anàlisi de components principals, cal aplicar el contrast d'hipòtesi per poder classificar el tipus de dany.

El capítol "Damage and fault detection of structures using principal component analysis and hypothesis testing", del llibre *Advances in Principal Component Analysis* (pàg. 137-191), presenta l'estat de l'art de l'aplicació de l'anàlisi de components principals en el camp del monitoratge de la integritat estructural, com també en el monitoratge de la condició, és a dir, detecció, identificació i classificació de danys o errors en estructures diverses.

Anàlisi de components principals i R

El codi que hem fet servir abans per calcular les components principals i la projecció de les dades originals sobre l'espai vectorial generat per aquestes components es pot simplificar si fem servir directament la instrucció que ofereix R. En efecte, amb el codi:

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

carreguem de nou les dades anomenades *iris* i, per mitjà de la instrucció `prcomp`, calculem l'anàlisi de components principals. Fixeu-vos que és fonamental que hi afegim les opcions `center = TRUE` i `scale = TRUE` per escalar les dades. Què conté l'estructura `iris.pca`? Vegem-ho.

```
> 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
```

L'estructura `iris.pca` té cinc camps:

1) `sdev`, un vector que conté les arrels quadrades dels valors propis de la matriu de covariàncies;

2) `rotation`, una matriu que conté les components principals i, per tant, una matriu que en el mòdul "Descomposició en valors singulars: introducció i aplicacions" hem anomenat $\mathbf{P}$;

3) `center`, un vector que conté la mitjana aritmètica de cada variable en el conjunt de dades original;

4) `scale`, un vector que conté la desviació tipus de cada variable en el conjunt de dades original;

5) `x`, una matriu que conté les projeccions de les dades originals sobre l'espai vectorial generat per les components principals i, per tant, una matriu que en el mòdul "Descomposició en valors singulars: introducció i aplicacions" hem anomenat $\mathbf{T}$.

Per tant, per fer un diagrama de dispersió com el de la figura 4, caldria utilitzar aquest codi:

```
> data(``iris'')
> iris.pca<-prcomp(iris[,1:4],center = TRUE, scale = TRUE)
> T<-iris.pca$x # definim 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)
```

També hi ha moltes opcions en `R` per representar gràficament les projeccions sobre l'espai vectorial de les dues primeres components principals. A continuació n'adjuntem una que necessita carregar la llibreria `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 resultat es pot observar a la figura 5. Del codi, cal destacar-ne l'ús de la instrucció `autoplot`, amb les opcions que comentarem tot seguit:

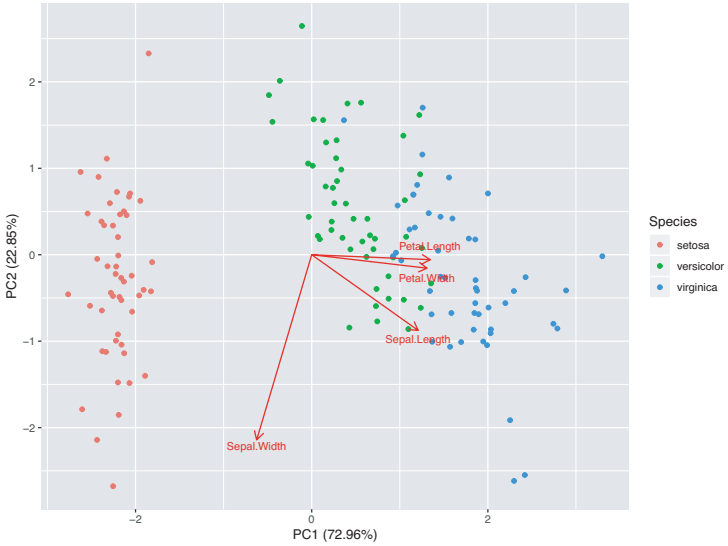
- `data=df2` conté les dades originals, incloent-hi la cinquena columna corresponent a les etiquetes (en aquest cas, l'espècie associada a cada punt);
- `color='Species'` agrupa per colors els punts de la mateixa espècie;
- `loadings = TRUE` incorpora els vectors amb la influència de les dues components principals en les quatre variables originals;
- `loadings.label = TRUE` incorpora els noms de les variables originals en el diagrama de dispersió;
- `loading.label.size` estableix la mida dels noms de les variables originals;
- `scale = 0` manté l'escala de les dades originals que ja han estat escalades en la creació del model de l'anàlisi de components principals.

Observant la figura 5 també podem extreure la informació visual següent:

1) Es pot observar la contribució de cadascuna de les quatre variables originals a les dues primeres components principals. Si ens fixem, per exemple, en la primera component principal (la direcció horitzontal), les variables que més hi intervenen són la longitud del sèpal (Sepal. Length) i la longitud i amplada del pètal (Petal. Length, Petal. Width). En el cas de la segona component principal (la direcció vertical), la variable que té més pes és l'amplada del sèpal (Sepal. Width).

2) També es pot veure que les fletxes que indiquen les direccions de les variables sobre longitud i amplada del pètal (Petal. Length, Petal. Width) estan pràcticament superposades. Recordem que, en aquest cas, la covariància entre aquestes dues variables és 0.9628654, que representa un valor molt proper a 1. És a dir, ja havíem dit que aquestes variables estan altament relacionades.

Figura 5. Diagrama de dispersió de les dues primeres components principals. Els colors representen el tipus o espècie de flor: *setosa* (vermell), *versicolor* (verd) i *virginica* (blau). A més, s'inclou la influència de les dues primeres components principals en les quatre variables originals.



Font: elaboració pròpia

Classificació a partir de l'anàlisi de components principals

Vegem un exemple senzill d'aplicació de l'anàlisi que hem dut a terme fins ara per classificar una nova flor desconeguda. Imaginem, doncs, que ens arriba una flor amb les característiques que recull la taula 2.

Taula 2. Longitud i amplada del sèpal i longitud i amplada del pètal (en centímetres)

flor	long. sèpal	ampl. sèpal	long. pètal	ampl. pètal	tipus
106	6.0	2.8	4.3	1.3	desconegut

Font: elaboració pròpia

Com podem saber a quina espècie pertany aquesta flor?

1) Escalarem els valors de la longitud i amplada del sèpal i de longitud i amplada del pètal, en relació amb les mitjanes aritmètiques i les desviacions tipus de les dades originals (150 flors). Anomenarem $\mathbf{u}^T$ el vector que conté aquests quatre valors normalitzats.

2) Projectarem aquest vector sobre l'espai vectorial generat per les components principals amb el producte

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

i anomenem $\mathbf{t}^T$ el resultat.

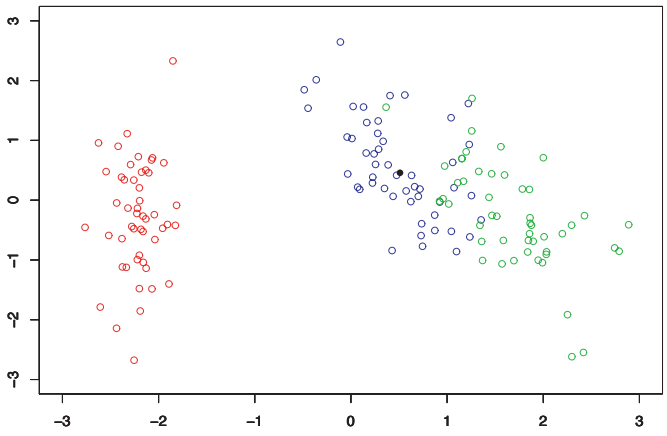
3) Afegirem les dues primeres components d'aquest punt al diagrama de dispersió original i mirarem a quin clúster pertany.

El codi de R que cal fer servir és:

```
> 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$scale # escalat de les dades
> u<-f%*%iris.pca$rotation # projecció de les dades escalades
> 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)
```

Podem veure'n el resultat a la figura 6. El punt negre correspon al punt que hem d'identificar. A partir d'una inspecció visual afirmaríem que aquest punt pertany o és més proper al conjunt o clúster de punts blaus. Per tant, classificaríem el punt (la flor) com a *versicolor*.

Figura 6. Diagrama de dispersió de les dues primeres components principals. Els colors representen el tipus o espècie de flor: *setosa* (vermell), *versicolor* (blau) i *virginica* (verd). El punt negre és el que hem de classificar.



Font: elaboració pròpia

Bibliografia

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, núm. 5, pàg. 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, núm. 2).

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