Principal Component Analysis

Jan Graffelman[1]

[1]Department of Statistics and Operations Research
Universitat Politècnica de Catalunya
Barcelona, Spain

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

jan.graffelman@upc.edu

February 22, 2020

Contents

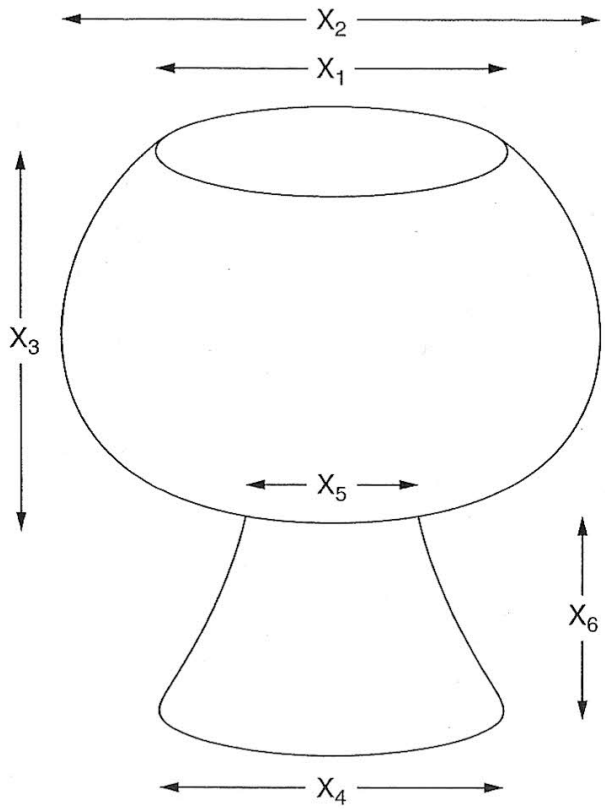
A bit of history

- Pearson, K. (1901) On lines and planes of closest fit to systems of points in space *Philosophical Magazine* **6**(2): 559-572.
- Hotelling, H. (1933) Analysis of a complex of statistical variables into principal components, *Journal of Educational Psychology*, **24**: 417-441,498-520.
- Gabriel, K. R. (1971) The biplot graphic display of matrices with application to principal component analysis, *Biometrika*, **58**(3): 453-467.

Archeological goblets from Thailand (Manly, 1989)

	X1	X2	X3	X4	X5	X6
1	13	21	23	14	7	8
2	14	14	24	19	5	9
3	19	23	24	20	6	12
4	17	18	16	16	11	8
5	19	20	16	16	10	7
6	12	20	24	17	6	9
7	12	19	22	16	6	10
8	12	22	25	15	7	7
9	11	15	17	11	6	5
10	11	13	14	11	7	4
11	12	20	25	18	5	12
12	13	21	23	15	9	8
13	12	15	19	12	5	6
14	13	22	26	17	7	10
15	14	22	26	15	7	9
16	14	19	20	17	5	10
17	15	16	15	15	9	7
18	19	21	20	16	9	10
19	12	20	26	16	7	10
20	17	20	27	18	6	14
21	13	20	27	17	6	9
22	9	9	10	7	4	3
23	8	8	7	5	2	2
24	9	9	8	4	2	2
25	12	19	27	18	5	12

Measurements on archeaological goblets (cm)



Download Goblets.dat

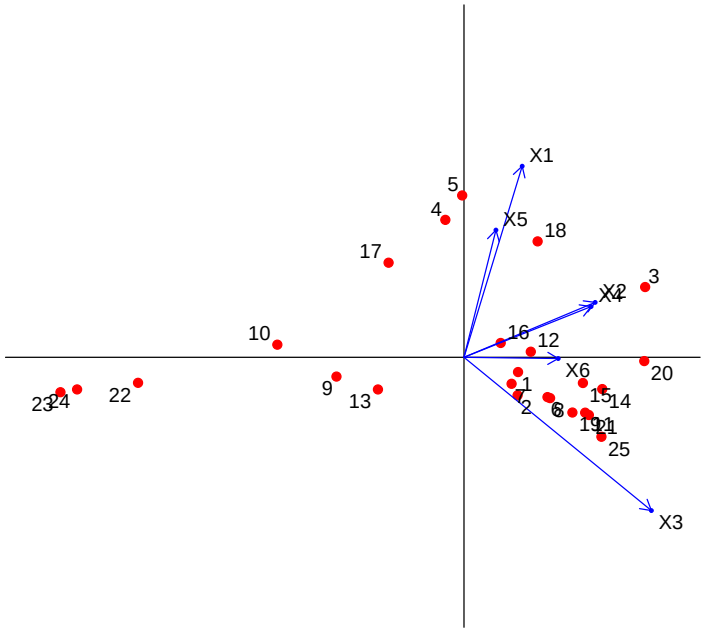
PCA objectives

Main goals:

- Reduce the number of variables
- A picture of the data matrix (biplot)

PCA objectives

	X1	X2	X3	X4	X5	X6
1	13	21	23	14	7	8
2	14	14	24	19	5	9
3	19	23	24	20	6	12
4	17	18	16	16	11	8
5	19	20	16	16	10	7
6	12	20	24	17	6	9
7	12	19	22	16	6	10
8	12	22	25	15	7	7
9	11	15	17	11	6	5
10	11	13	14	11	7	4
11	12	20	25	18	5	12
12	13	21	23	15	9	8
13	12	15	19	12	5	6
14	13	22	26	17	7	10
15	14	22	26	15	7	9
16	14	19	20	17	5	10
17	15	16	15	15	9	7
18	19	21	20	16	9	10
19	12	20	26	16	7	10
20	17	20	27	18	6	14
21	13	20	27	17	6	9
22	9	9	10	7	4	3
23	8	8	7	5	2	2
24	9	9	8	4	2	2
25	12	19	27	18	5	12



Theory PCA (1)

We search for linear combinations of the original variables

$$\begin{aligned} F_1 &= a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \\ F_2 &= a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p \\ &\vdots \\ F_p &= a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p \end{aligned}$$

Subject to:

- $F_1, F_2, \ldots F_p$ uncorrelated
- $\text{Var}(F_1)$ maximal
- $\text{Var}(F_1) \geq \text{Var}(F_2) \geq \cdots \geq \text{Var}(F_p)$
- $a_{i1}^2 + a_{i2}^2 + \cdots + a_{ip}^2 = 1 \qquad (-1 \leq a_{ij} \leq 1)$

Theory PCA

- All coefficients and eigenvalues can be obtained by the spectral decomposition of the covariance matrix:

$$\mathbf{S} = \mathbf{A}\mathbf{D}_\lambda\mathbf{A}'.$$

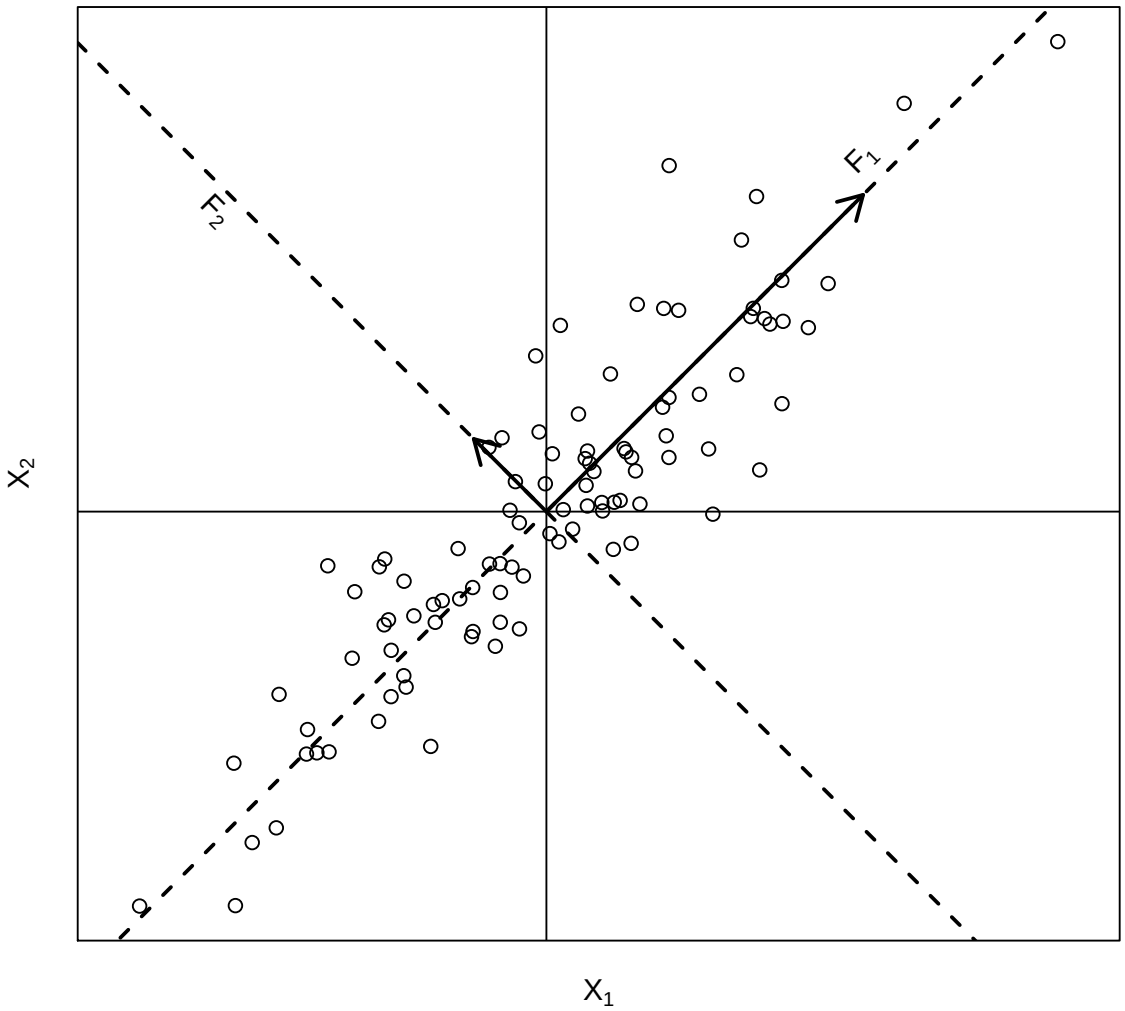
- The principal components are obtained as:

$$\underset{(n\times p)}{\mathbf{F}_p} = \underset{(n\times p)}{\mathbf{X}_c}\ \underset{(p\times p)}{\mathbf{A}}$$

- Eigenvalues correspond to variances of the principal components because:

$$\frac{1}{n-1}\mathbf{F}_p'\mathbf{F}_p = \frac{1}{n-1}(\mathbf{X}_c\mathbf{A})'\mathbf{X}_c\mathbf{A} = \frac{1}{n-1}\mathbf{A}'\mathbf{X}_c{}'\mathbf{X}_c\mathbf{A} = \mathbf{A}'\mathbf{S}\mathbf{A} = \mathbf{A}'\mathbf{A}\mathbf{D}_\lambda\mathbf{A}\mathbf{A}' = \mathbf{D}_\lambda$$

Geometric Interpretation ($p = 2$)



Theory PCA

- An alternative way to perform PCA is by the singular value decomposition (SVD) of the centred data matrix.

$$\mathbf{X}_c = \mathbf{UDA}'.$$

- The principal components are obtained as:

$$\mathbf{F}_p = \mathbf{X}_c\mathbf{A} = \mathbf{UD}$$

- Squared singular values relate to the variance of the principal components because:

$$\frac{1}{n-1}\mathbf{F}_p{}'\mathbf{F}_p = \frac{1}{n-1}(\mathbf{UD})'\mathbf{UD} = \frac{1}{n-1}\mathbf{D}^2 = \mathbf{D}_\lambda$$

- The SVD approach is very convenient for biplot construction.

What is a biplot?

A biplot is a powerful tool for the graphical exploration of multivariate data (e.g. pattern and outlier detection).

A biplot is a multivariate generalization of the scatter plot.

A biplot differs from a scatterplot in some ways:

- It has typically more than 2 axes
- The axes are not perpendicular, but tend to be oblique.
- The data matrix is not represented exactly, but approximately.

A biplot is a joint display of the rows and the columns a matrix that is optimal in the least squares sense.

Making a biplot

In order to make a biplot, the matrix to be represented needs to be factored

$$\mathbf{X}_{n\times p} = \mathbf{F}_{n\times r}\mathbf{G}'_{r\times p} \tag{1}$$

into a matrix of row markers ($\mathbf{F}$) and a matrix of column markers ($\mathbf{G}$). Note that this factorization also exists in an ordinary scatter plot:

$$\mathbf{X}_{n\times 2} = \mathbf{X}_{n\times 2}\mathbf{I}_{2\times 2}$$

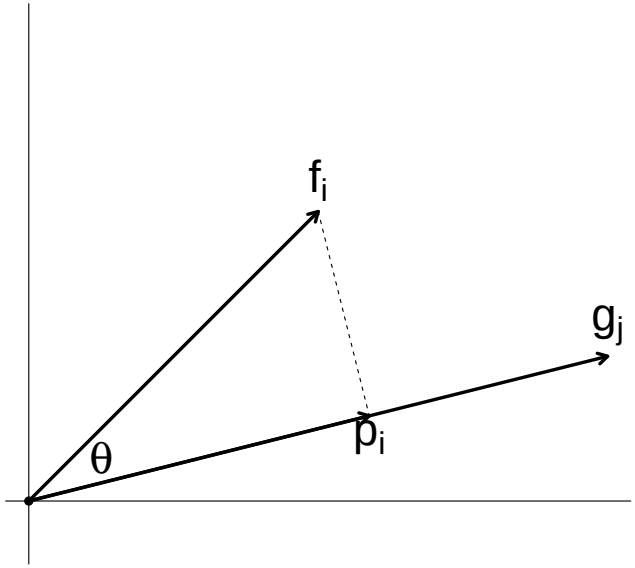
The factorization (1) is not unique:

$$\mathbf{X}_{n\times p} = \mathbf{F}_{n\times r}\mathbf{T}\mathbf{T}^{-1}\mathbf{G}'_{r\times p} = \tilde{\mathbf{F}}_{n\times r}\tilde{\mathbf{G}}'_{r\times p} \tag{2}$$

where $\mathbf{T}$ is any non-singular linear transformation.

Biplots and scalar product

In a biplot data values are approximated by the scalar product between two vectors.



$$\cos\theta = \frac{\| \mathbf{p} \|}{\| \mathbf{f}_i \|} = \frac{\mathbf{f}_i' \mathbf{g}_j}{\| \mathbf{f}_i \| \| \mathbf{g}_j \|} \qquad \| \mathbf{p} \| = \frac{\mathbf{f}_i' \mathbf{g}_j}{\| \mathbf{g}_j \|} \qquad x_{ij} \approx \mathbf{f}_i' \mathbf{g}_j = \| \mathbf{p}_i \| \cdot \| \mathbf{g}_j \|$$

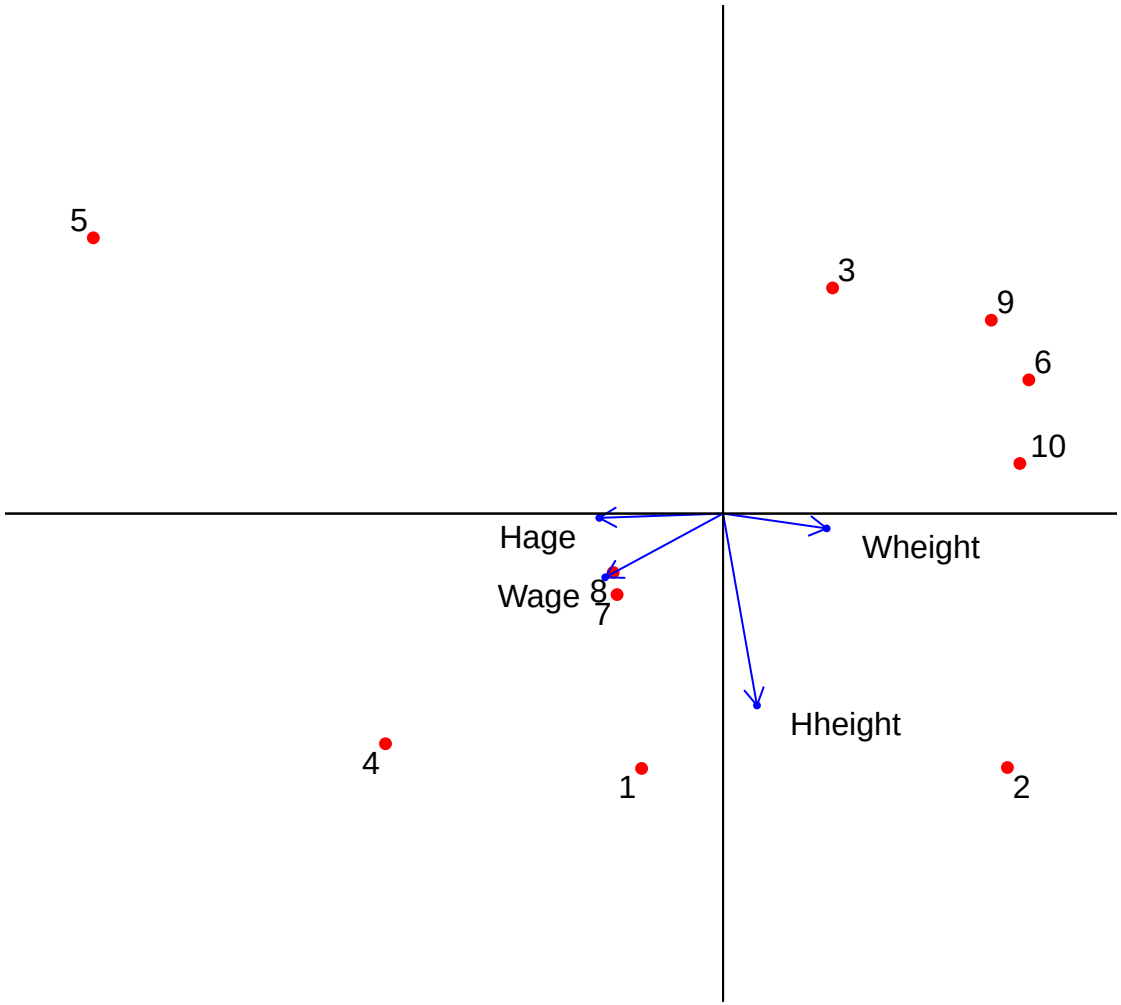
Example of a biplot

	X			
	Age H	Height H	Age W	Height W
1	49	1809	43	1590
2	25	1841	28	1560
3	40	1659	30	1620
4	52	1779	57	1540
5	58	1616	52	1420
6	32	1695	27	1660
7	43	1730	52	1610
8	47	1740	43	1580
9	31	1685	23	1610
10	26	1735	25	1590

Age and height of husband and wife for 10 couples

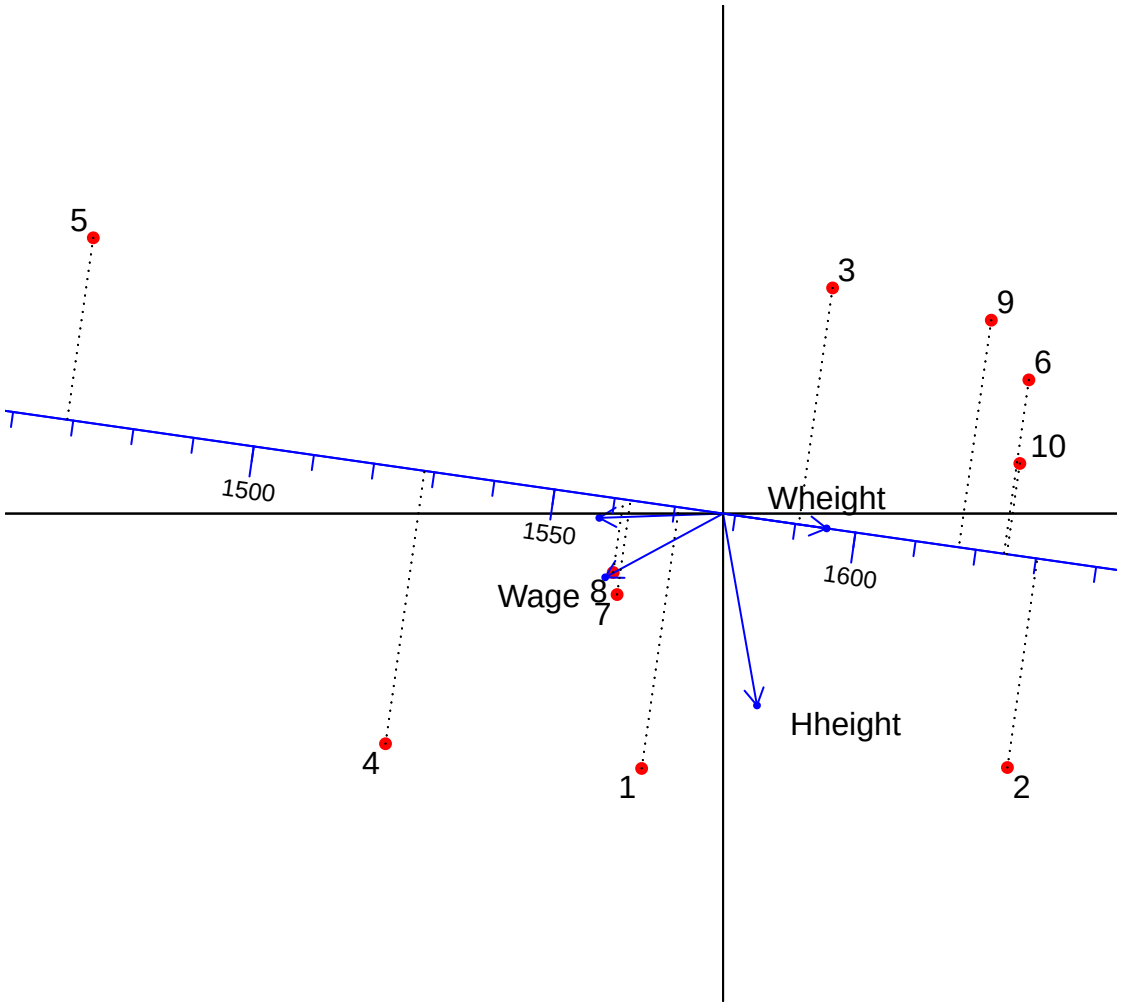
Example of a biplot

Biplot of the standardized Husband and Wife data.



Biplot interpretation

Biplot of the standardized Husband and Wife data.



Biplots in PCA

$$\mathbf{F}_p = \mathbf{X}_c \mathbf{A}$$

$$\mathbf{X}_c = \mathbf{F}_p \mathbf{A}' = \mathbf{F}_p \mathbf{G}_s{}' \qquad \mathbf{G}_s = \mathbf{A}$$

- PCA gives a biplot of the centred data matrix.
- The biplot is obtained by jointly plotting the first two principal components (first two columns of $\mathbf{F}_p$) and the first two eigenvectors (first two colums of $\mathbf{G}_s$)
- The rows of $\mathbf{F}_p$ are usually represented by dots, and the rows of $\mathbf{G}_s$ by arrows.
- The coordinates $\mathbf{F}_p$ are called principal coordinates, and the coordinates $\mathbf{G}_s$ are called standard coordinates.
- Standard coordinates satisfy $\mathbf{G}_s{}'\mathbf{G}_s = \mathbf{I}$

Biplots in PCA: alternative scaling

$$\mathbf{X}_c = \mathbf{F}_p \mathbf{D}_s{}^{-1} \mathbf{D}_s \mathbf{G}_s{}' = \mathbf{F}_s \mathbf{D}_s \mathbf{G}_s{}' = \mathbf{F}_s \mathbf{G}_p{}' \qquad \mathbf{G}_p = \mathbf{G}_s \mathbf{D}_s$$

- This biplot plots the standardized principal components $\mathbf{F}_s = \mathbf{F}_p \mathbf{D}_s{}^{-1}$
- $\mathbf{D}_s$ contains the standard deviations of the principal components

We thus have two biplots:

- $\mathbf{X}_c = \mathbf{F}_p \mathbf{G}_s{}'$ (the form biplot)
- $\mathbf{X}_c = \mathbf{F}_s \mathbf{G}_p{}'$ (the covariance biplot)

In general, form biplots focus on the representation of distances, whereas covariance biplot focus on representing correlation structure.

Some biplot properties

- In the form biplot, Euclidean distances between points approximate Euclidean distances between rows of the data matrix.
- In the covariance biplot, Euclidean distances between points approximate Mahalanobis distances between rows of the data matrix.
- In the covariance biplot, the length of an arrow approximates the standard deviation of the corresponding variable.
- In the covariance biplot, the angle between two arrows approximates the correlation between the two corresponding variables.

How many components ?

Criteria:

- Percentage of explained variance ($> 80\%$).
- Size of the eigenvalue ($> \overline{\lambda}$).
- The scree plot.
- Significance tests with the eigenvalues.

How many components ?

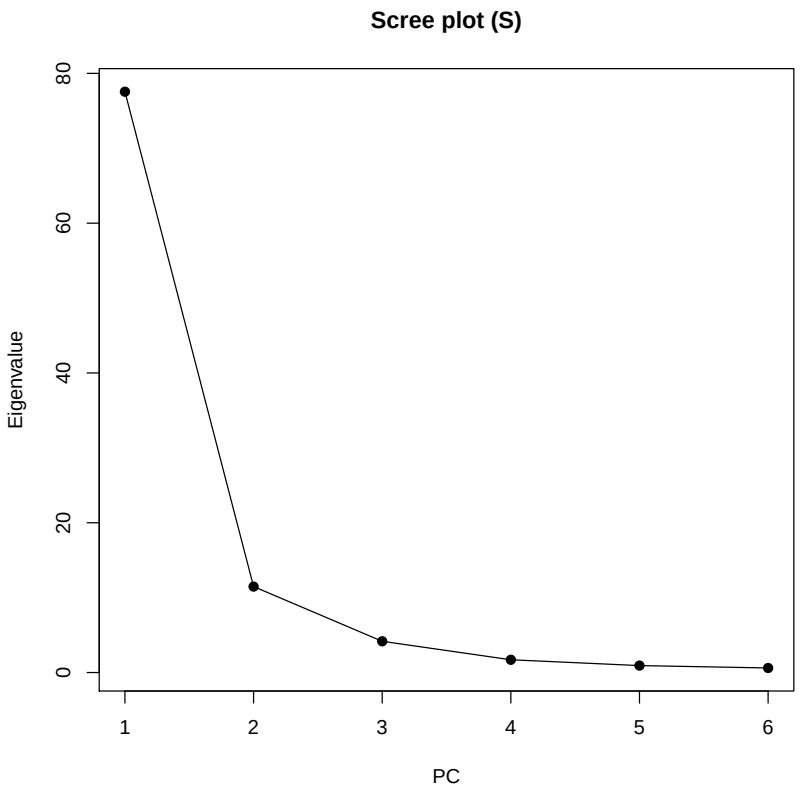
We have:

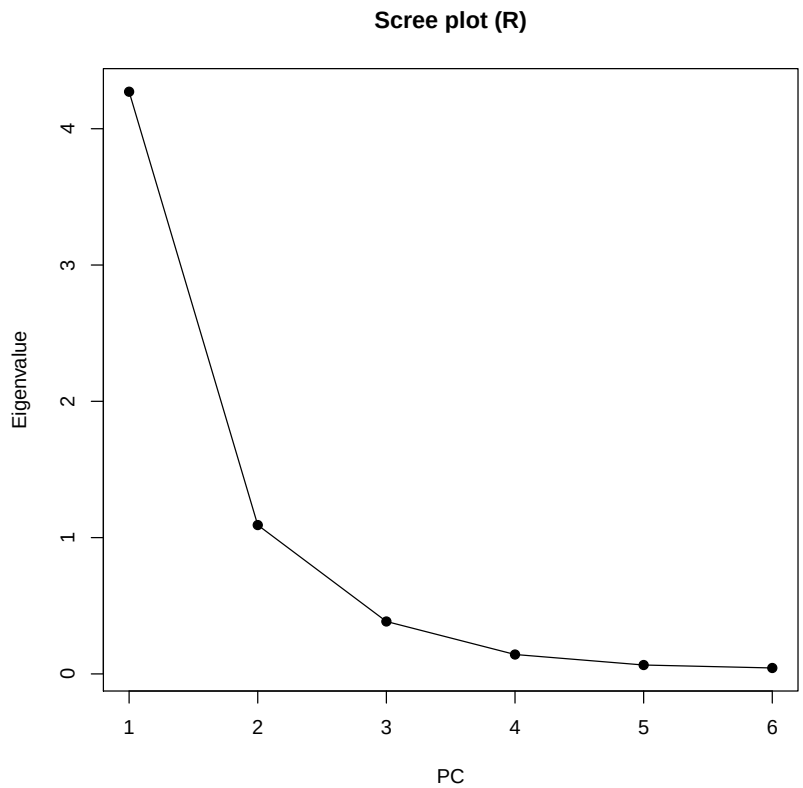
$$\text{tr}(\mathbf{S}) = \text{tr}(\mathbf{A}\mathbf{D}_\lambda\mathbf{A}') = \text{tr}(\mathbf{D}_\lambda).$$

$$\sum_{i=1}^{p} V(X_i) = \sum_{i=1}^{p} V(F_i) = \sum_{i=1}^{p} \lambda_i.$$

Component	F_1	F_2	$\cdots$	F_p
Variance	λ_1	λ_2	$\cdots$	λ_p
Fraction	$\lambda_1/\sum\lambda_i$	$\lambda_2/\sum\lambda_i$	$\cdots$	$\lambda_p/\sum\lambda_i$
Cum. Fraction	$\lambda_1/\sum\lambda_i$	$(\lambda_1+\lambda_2)/\sum\lambda_i$	$\cdots$	$\sum\lambda_i/\sum\lambda_i$

The scree plot





Types of PCA

There are two types of PCA. Computations can be based on

- the covariance matrix (**S**)
  - Not invariant w.r.t. the scale of measurement
  - The variable with the largest variance dominates
  - Some authors focus on components with $\lambda_i > \overline{\lambda}$
- the correlation matrix (**R**)
  - Invariant w.r.t. the scale of measurement
  - All variables have equal weight
  - Some authors focus on components with $\lambda_i > 1$

Interpretation

Components can be interpreted with the aid of

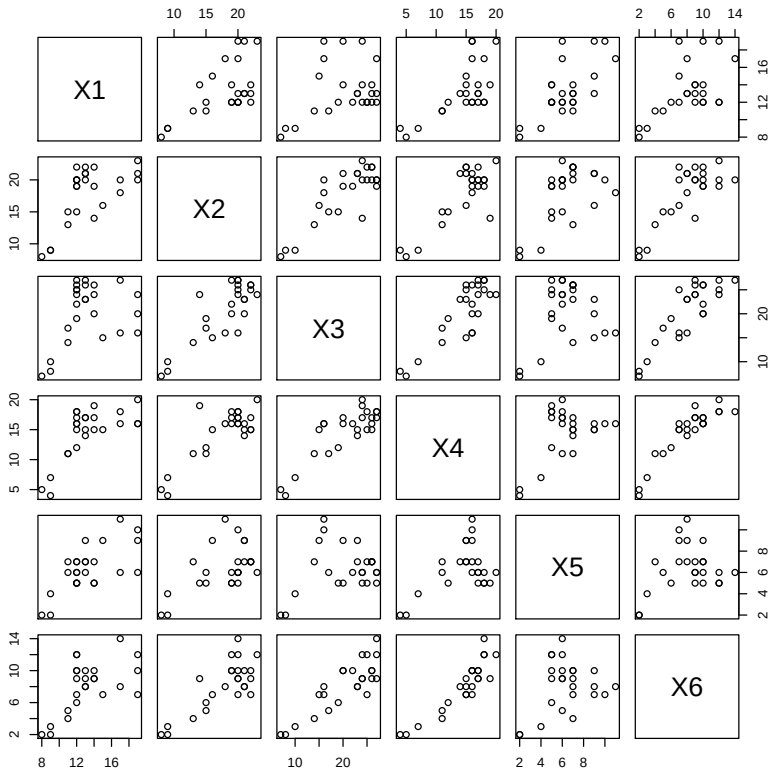
- the coefficients
- the correlations between variables and components
- the biplot

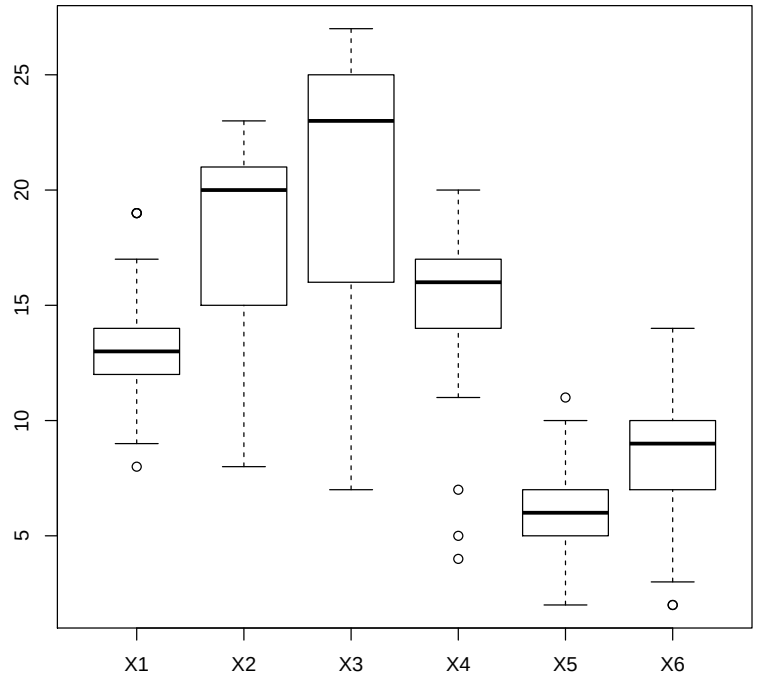
If the aim is to get a picture of the data matrix, then interpretation of the components may not be needed.

Some notes on goodness-of-fit

- From the eigenvalues of the analysis the overall goodness-of-fit of a k-dimensional solution can be calculated.
- One can also compute the goodness-of-fit of individual rows and columns of the data matrix.
- Goodness-of-fit of the variables can also be computed as R^2 in a regression onto the principal components.

Goblets: some exploratory analysis





Descriptive statistics

	N	N*	Mean	Stdev	Med	Min	Max
X1	25	0	13.28	3.01	13	8	19
X2	25	0	17.84	4.34	20	8	23
X3	25	0	20.44	6.08	23	7	27
X4	25	0	14.60	4.14	16	4	20
X5	25	0	6.36	2.16	6	2	11
X6	25	0	8.12	3.14	9	2	14

	X1	X2	X3	X4	X5	X6
X1	9.04	8.13	6.33	8.41	4.48	5.55
X2	8.13	18.81	22.11	14.89	5.43	10.85
X3	6.33	22.11	36.92	21.23	3.29	16.36
X4	8.41	14.89	21.23	17.17	4.36	11.84
X5	4.48	5.43	3.29	4.36	4.66	1.96
X6	5.55	10.85	16.36	11.84	1.96	9.86

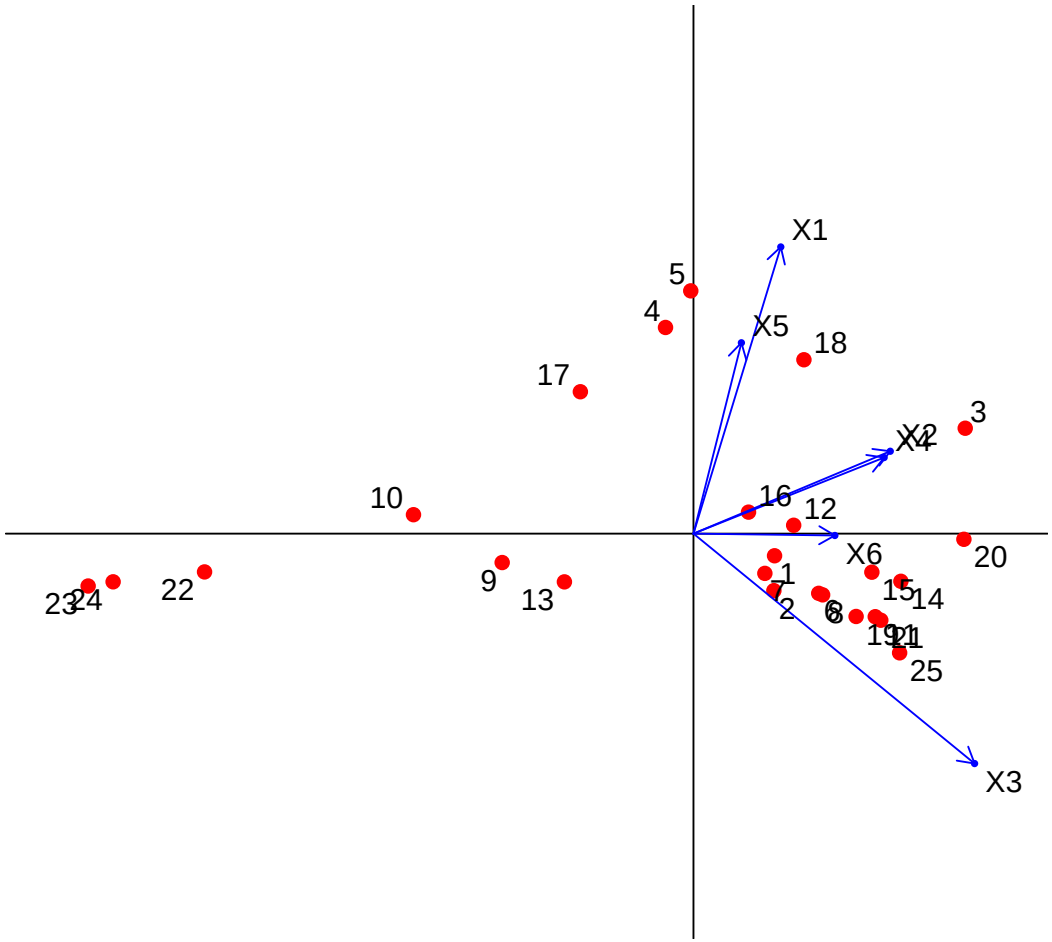
	X1	X2	X3	X4	X5	X6
X1	1.00	0.62	0.35	0.67	0.69	0.59
X2	0.62	1.00	0.84	0.83	0.58	0.80
X3	0.35	0.84	1.00	0.84	0.25	0.86
X4	0.67	0.83	0.84	1.00	0.49	0.91
X5	0.69	0.58	0.25	0.49	1.00	0.29
X6	0.59	0.80	0.86	0.91	0.29	1.00

PCA of the covariance matrix

	PC1	PC2	PC3	PC4	PC5	PC6
X1	0.20	0.67	-0.23	-0.27	0.61	-0.11
X2	0.46	0.19	0.62	-0.44	-0.37	-0.22
X3	0.66	-0.54	0.11	0.16	0.49	0.06
X4	0.44	0.18	-0.45	0.46	-0.39	-0.45
X5	0.11	0.44	0.39	0.60	-0.01	0.52
X6	0.33	0.00	-0.45	-0.36	-0.31	0.68

	PC1	PC2	PC3	PC4	PC5	PC6
eigenvalue	77.56	11.48	4.17	1.71	0.93	0.61
fraction	0.80	0.12	0.04	0.02	0.01	0.01
cumulative	0.80	0.92	0.97	0.98	0.99	1.00

Biplot (principal components)

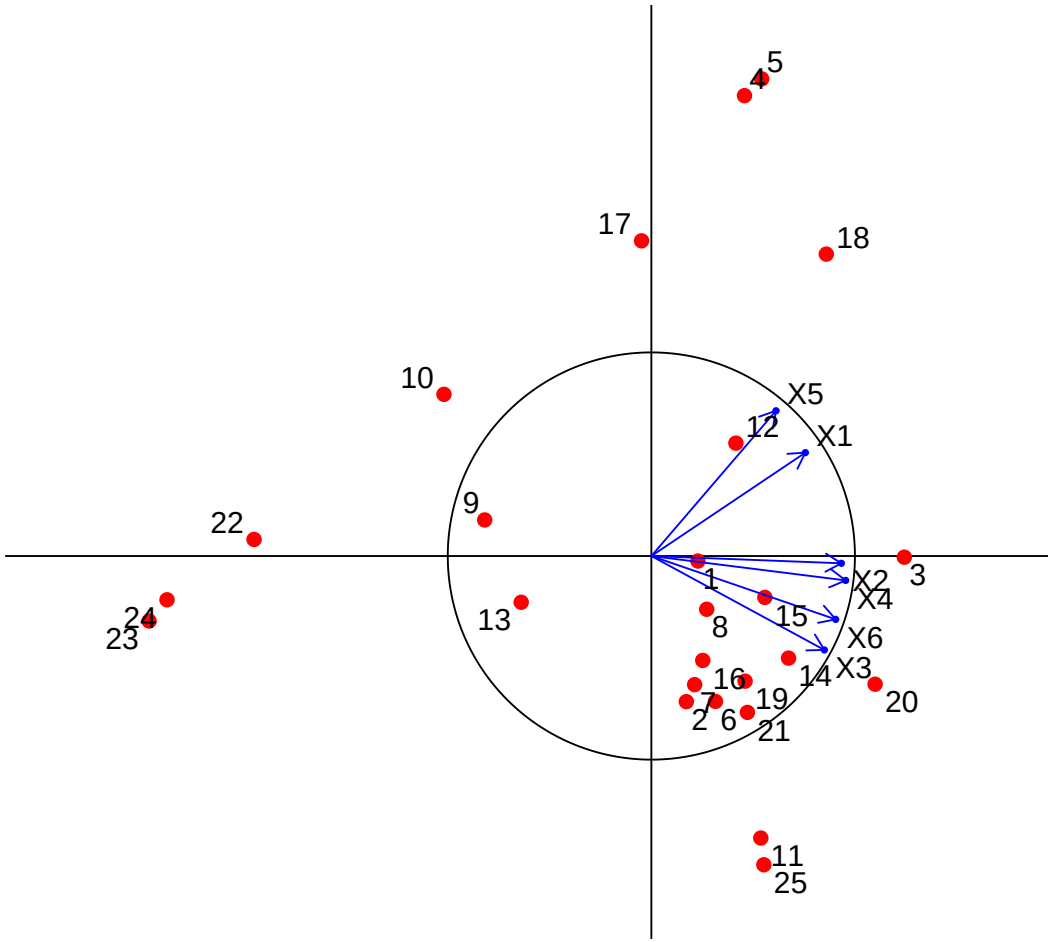


PCA of the correlation matrix

	PC1	PC2	PC3	PC4	PC5	PC6
X1	0.37	0.49	-0.62	-0.32	0.28	0.26
X2	0.45	-0.03	0.38	-0.67	-0.08	-0.44
X3	0.41	-0.44	0.32	0.02	0.38	0.62
X4	0.46	-0.11	-0.16	0.54	0.38	-0.56
X5	0.30	0.68	0.49	0.36	-0.22	0.16
X6	0.44	-0.30	-0.33	0.13	-0.76	0.13

	PC1	PC2	PC3	PC4	PC5	PC6
eigenvalue	4.27	1.09	0.38	0.14	0.07	0.04
fraction	0.71	0.18	0.06	0.02	0.01	0.01
cumulative	0.71	0.89	0.96	0.98	0.99	1.00

Biplot (standardized principal components)

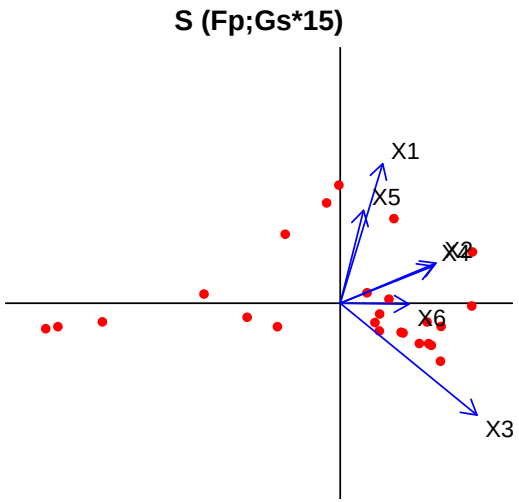


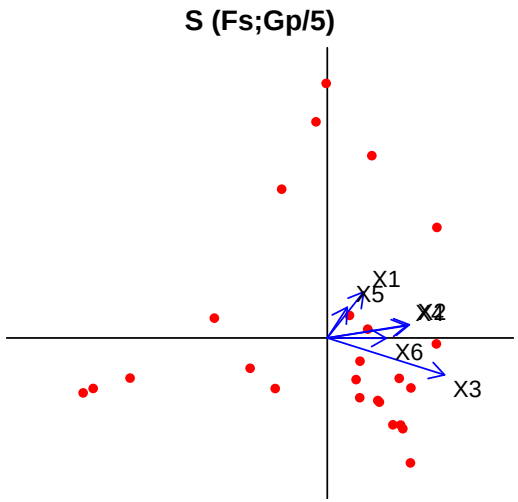
PCA in R

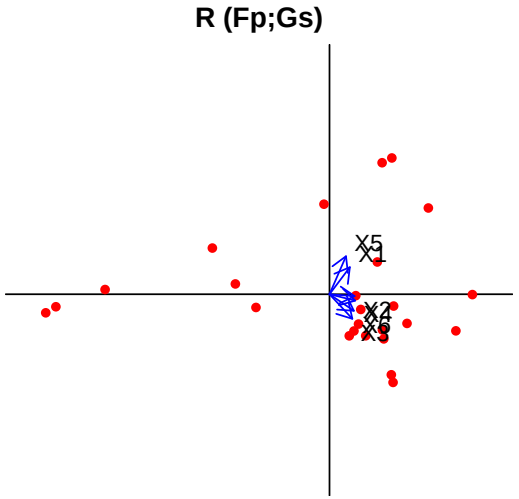
Try:

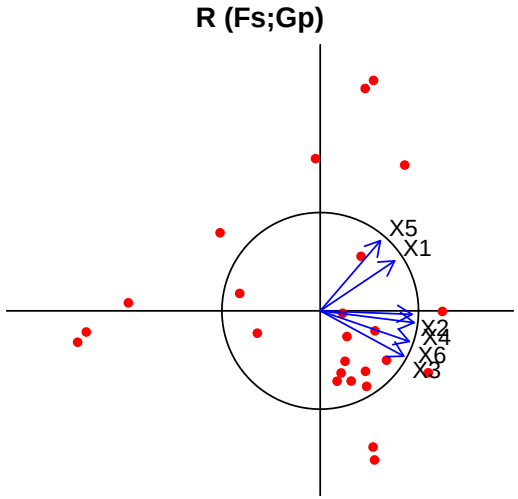
```
> X <- read.table("http://www-eio.upc.es/~jan/Data/goblets.dat",header=TRUE)
> ?princomp
> out <- princomp(X)
> summary(out)
> biplot(out)

> install.packages("FactoMineR")
> library("FactoMineR")
> res.pca = PCA(X, scale.unit=TRUE, ncp=5, graph=T)

> install.packages("ade4")
> library("ade4")
> pca1 <- dudi.pca(X,scannf=FALSE,nf=3)
> scatter(pca1)
```

Four PCA biplots









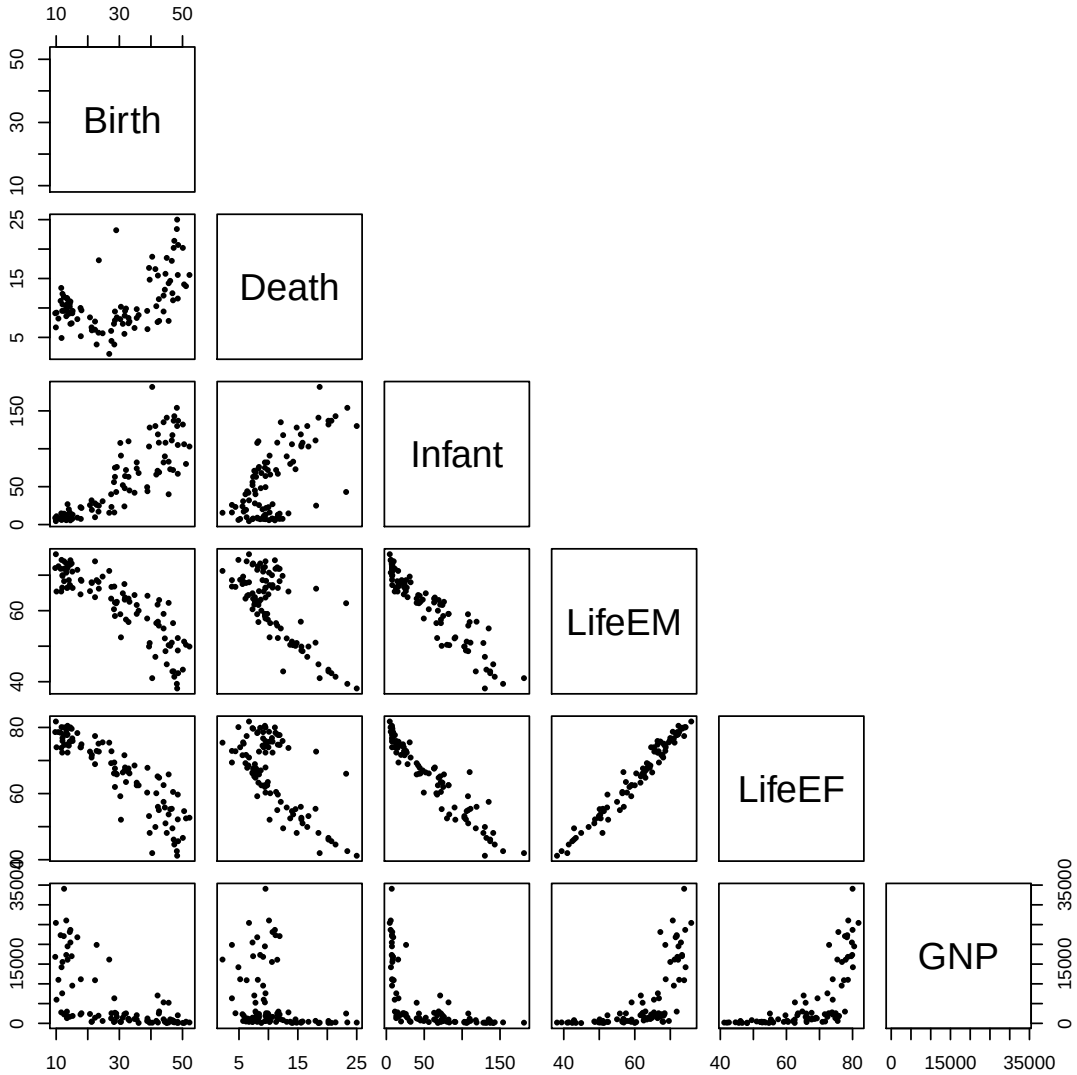
Another example: poverty data set

For 97 countries in the world the following variables are registered

- Birth: Live birth rate per 1,000 of population
- Death: Death rate per 1,000 of population
- Infant: Infant deaths per 1,000 of population under 1 year old
- LifeEM: Life expectancy at birth for males
- LifeEF: Life expectancy at birth for females
- GNP: Gross National Product per capita in U.S. dollars
- Country: Name of the country

	Birth	Death	Infant	LifeEM	LifeEF	GNP
Albania	24.70	5.70	30.80	69.60	75.50	600
Bulgaria	12.50	11.90	14.40	68.30	74.70	2250
Czechoslovakia	13.40	11.70	11.30	71.80	77.70	2980
Former_E._Germany	12.00	12.40	7.60	69.80	75.90	NA
Hungary	11.60	13.40	14.80	65.40	73.80	2780
Poland	14.30	10.20	16.00	67.20	75.70	1690
Romania	13.60	10.70	26.90	66.50	72.40	1640
Yugoslavia	14.00	9.00	20.20	68.60	74.50	NA
USSR	17.70	10.00	23.00	64.60	74.00	2242
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Scatterplot matrix

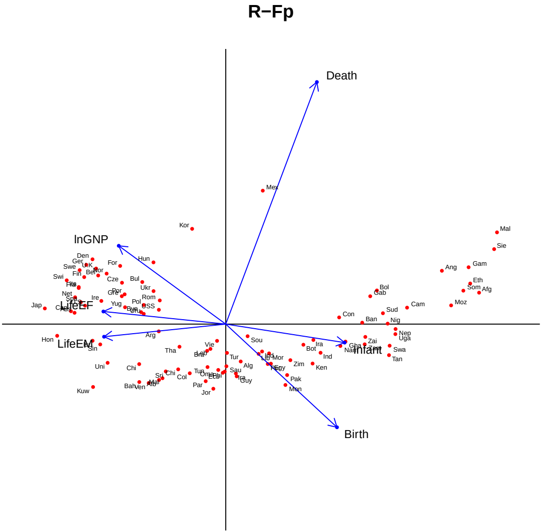


Correlation matrix and variance decomposition

	Birth	Death	Infant	LifeEM	LifeEF	lnGNP
Birth	1.00	0.49	0.86	-0.87	-0.89	-0.74
Death	0.49	1.00	0.65	-0.73	-0.69	-0.51
Infant	0.86	0.65	1.00	-0.94	-0.96	-0.79
LifeEM	-0.87	-0.73	-0.94	1.00	0.98	0.81
LifeEF	-0.89	-0.69	-0.96	0.98	1.00	0.83
lnGNP	-0.74	-0.51	-0.79	0.81	0.83	1.00

	PC1	PC2	PC3	PC4	PC5	PC6
λ	4.96	0.58	0.28	0.11	0.06	0.01
fraction	0.83	0.10	0.05	0.02	0.01	0.00
cumulative	0.83	0.92	0.97	0.99	1.00	1.00

Biplot


R−Fp
Death
lnGNP
LifeEF
LifeEM
Infant
Birth
Mex
Kor
Hun
For
Den
Swe
Swi
Fin
Bel
Nor
Cze
Bul
Ukr
Rom
Pol
USS
Yug
Ire
Net
Jap
Hon
Sin
Arg
Tha
Vie
Uni
Chi
Kuw
Bah
Ven
Col
Par
Jor
Tur
Sau
Alg
Guy
Sou
Mor
Zim
Ken
Pak
Mon
Bot
Ira
Ind
Con
Ban
Sud
Nig
Cam
Bol
Gab
Nep
Uga
Zai
Swa
Tan
Ang
Gam
Eth
Som
Afg
Moz
Mal
Sie

References

- Aluja-Banet, T. & Morineau, A. (1999) Aprender de los datos: el análisis de componentes principales. Una aproximación desde el data mining, Ediciones Universitarias de Barcelona.
- Cuadras, C. (2008) Nuevos métodos de Análisis Multivariante. Chapter 5. Download book here
- Johnson & Wichern, (2002) Applied Multivariate Statistical Analysis, 5th edition, Prentice Hall, Chapter 8.
- Jolliffe, I.T. (1986) Principal Component Analysis, Springer-Verlag, New York.
- Manly, B.F.J. (1989) Multivariate statistical methods: a primer. 3rd edition. Chapman and Hall, London. Chapter 6.
- Peña, D. (2002) Análisis de datos multivariantes. McGraw-Hill, Madrid.