Mathematical Models and Applications

Multivariate Analysis

Pedro Cristiano Silva

Instituto Superior de Agronomia

June 4, 2024

Outline

- **LINEAR ALGEBRA**
- **PRINCIPAL COMPONENT ANALYSIS**
- LINEAR DISCRIMINANT ANALYSIS
- **CLUSTER ANALYSIS**

Some bibliography (to be completed)

C J. Cadima, *Introduction to Multivariate Statistics*, slides (2021/22)

LL P. Legendre, L. Legendre, *Numerical Ecology* (2003)

E B. S. Everitt, *Cluster Analysis* (1993)

G B. Grün, *Model-based Clustering*, arXiv: 1807.01987 (2018)

HA L. Hubert and P. Arabie, *Comparing Partitions*, Journal of Classification 2 (1985)

MRS C. Manning, P. Raghavan, H. Shutze, *An Introduction to Information Retrieval* (2009)

R A. C Rencher, *Methods of Multivariate Analysis* (2002)

TK S. Theodoridis and K. Koutroumbas, Pattern Recognition (2009)

LMP L. Lebart, A. Morineau and M. Piron, *Statistique exploratoire multidimensionnelle*, Dunod (1995)

HLP F. Husson F., S. Lê S. and J. Pagès *Exploratory Multivariate Analysis by Example Using R*, Chapman & Hall (2010).

Notations

- Non bold letters (upper or lower case) represent scalar quantities: x, y, A,...
- Lower case bold letters represent vectors $\mathbf{x}$, $\mathbf{y}$, $\vec{\mathbf{x}}$, $\vec{\mathbf{y}}$,...
- Upper case bold letters represent matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{X}$, $\mathbf{Y}$,...

LINEAR ALGEBRA

Eigenvalues and eigenvectors

Definition

$\mathbf{A}_{p\times p} = [a_{ij}]$ a square matrix of order p. A vector $\mathbf{v} \in \mathbb{R}^p$, $\mathbf{v} \neq \vec{0}$, is called an **eigenvector** *of* **A** *if there is $\lambda \in \mathbb{R}$ such that $\mathbf{Av} = \lambda\mathbf{v}$. λ is called the corresponding* **eigenvalue**.

Example

$$\mathbf{A} = \begin{bmatrix} 3 & 0 & 2 \\ 0 & -1 & 1 \\ 2 & 0 & 0 \end{bmatrix}, \qquad \mathbf{v} = \begin{bmatrix} 10 \\ 1 \\ 5 \end{bmatrix}$$

We have

$$\mathbf{Av} = \begin{bmatrix} 3 & 0 & 2 \\ 0 & -1 & 1 \\ 2 & 0 & 0 \end{bmatrix} \begin{bmatrix} 10 \\ 1 \\ 5 \end{bmatrix} = \begin{bmatrix} 40 \\ 4 \\ 20 \end{bmatrix} = 4 \begin{bmatrix} 10 \\ 1 \\ 5 \end{bmatrix} = 4\mathbf{v}$$

Hence **v** *is an eigenvector of* **A** *associated to the eigenvalue $\lambda = 4$.*

Eigenvalues and eigenvectors (cont.)

- The spectrum of $\mathbf{A}$, denoted $\sigma(\mathbf{A})$, is the collection of the p eigenvalues of $\mathbf{A}$ (including repetitions), i.e., the collection of p roots (real and complex) of its **characteristic polynomial**, $p_{\mathbf{A}}(x) = \det(\mathbf{A} - x\mathbf{I}_p)$ (which has degree p)
- The eigenspace associated with an eigenvalue λ, denoted $E(\lambda)$, is the linear space spanned by the eigenvectors associated with λ
- The trace of $\mathbf{A}$, denoted $\text{tr}(\mathbf{A})$, is the sum of all diagonal elements of $\mathbf{A}$ and equals the sum of all eigenvalues of $\mathbf{A}$ (including repetitions):

$$\text{tr}(\mathbf{A}) = a_{11} + a_{22} + \cdots + a_{pp} = \sum_{\lambda \in \sigma(\mathbf{A})} \lambda$$

- The determinant of $\mathbf{A}$ (not defined here) equals the product of all eigenvalues of $\mathbf{A}$ (including repetitions):

$$\det \mathbf{A} = \prod_{\lambda \in \sigma(\mathbf{A})} \lambda$$

$\mathbf{A}$ is invertible $\Leftrightarrow \det(\mathbf{A}) \neq 0 \Leftrightarrow 0$ is not an eigenvalue of $\mathbf{A}$

Example revisited

Returning to the example of slide 6 we have the the following:

- $\sigma(\mathbf{A}) : -1, \ -1, \ 4$
- $\text{tr}(\mathbf{A}) = 3 + (-1) + 0 = 2$ corresponds to the sum of its diagonal elements which is also equal to sum of its eigenvalues (counting with repetitions): $(-1) + (-1) + 4 = 2$
- $\det(\mathbf{A}) = (-1) \times (-1) \times 4 = 4 \neq 0$ which is equal to the product of its eigenvalues (counting with repetitions)
- $E(-1) = \langle (1, 1, 0) \rangle$ has dim=1
- $E(4) = \langle (0, 1, 5) \rangle$ has dim=1

Since $\dim E(-1) + \dim E(4) = 2 < 3 = p$, $\mathbf{A}$ is not **diagonalizable**, i.e., we cannot find an invertible matrix $\mathbf{P}$ and a diagonal matrix $\mathbf{\Lambda}$ such that $\mathbf{A} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^{-1}$

Exercise

Verify that $(1, 1, 0)$ is an eigenvector of $\mathbf{A}$ associated to the eigenvalue $\lambda = -1$

Eigenvalues and eigenvectors - R

R

```
A=matrix(c(3,0,2,0,-1,1,2,0,0),ncol=3,byrow=TRUE)

A

EV<-eigen(A) # eigenvalues and eigenvectors of A

det(A) # determinant of A

tr<-sum(diag(A)) # trace of A

tr
```

Orthonormal sets

Definition

Given $\mathbf{v}_1, \ldots, \mathbf{v}_q \in \mathbb{R}^p$ *with* $q \leq p$ *we say that* $\{\mathbf{v}_1, \ldots, \mathbf{v}_q\}$ *is an* **orthonormal set** *if*

$$\|\mathbf{v}_i\| = 1, \forall i \quad \textit{and} \quad \mathbf{v}_i \perp \mathbf{v}_j \quad (i \neq j)$$

If $q = p$, $\{\mathbf{v}_1, \ldots, \mathbf{v}_q\}$ *is called an* **orthonormal basis** *of* $\mathbb{R}^p$

Denoting by $\mathbf{V}_{p\times q} = [\, \mathbf{V}_1 \quad \cdots \quad \mathbf{V}_q \,]$ the matrix whose columns are the q vectors, $\mathbf{v}_1, \ldots, \mathbf{v}_q$, we have the following:

- $\{\mathbf{v}_1, \ldots, \mathbf{v}_q\}$ is an orthonormal set iff $\mathbf{V}^T\mathbf{V} = \mathbf{I}_q$
- $\{\mathbf{v}_1, \ldots, \mathbf{v}_p\}$ is an orthonormal basis iff $\mathbf{V}^T\mathbf{V} = \mathbf{V}\mathbf{V}^T = \mathbf{I}_p$ iff $\mathbf{V}^{-1} = \mathbf{V}^T$

In the later case we can write for all $\mathbf{u} \in \mathbb{R}^p$,

$$\mathbf{u} = \underbrace{(\mathbf{u}^T\mathbf{v}_1)\mathbf{v}_1}_{\text{proj}_{\mathbf{v}_1}(\mathbf{u})} + \cdots + \underbrace{(\mathbf{u}^T\mathbf{v}_p)\mathbf{v}_p}_{\text{proj}_{\mathbf{v}_p}(\mathbf{u})} \tag{1}$$

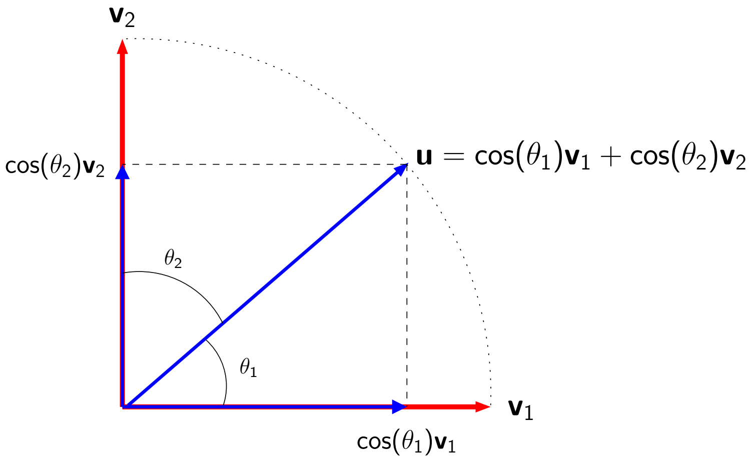
Orthonormal basis

If $\|\mathbf{u}\| = 1$ and $\{\mathbf{v}_1, \dots, \mathbf{v}_p\}$ is an orthonormal basis of $\mathbb{R}^p$ we have, applying (1) of slide 10,

$$\boxed{\mathbf{u} = \cos(\theta_1)\mathbf{v}_1 + \cdots + \cos(\theta_p)\mathbf{v}_p}$$

with $u^T u = \cos^2(\theta_1) + \cdots + \cos^2(\theta_p) = 1$, where θ_i, $i = 1, \dots, p$, denotes the angle between $\mathbf{u}$ and $\mathbf{v}_i$

The case $p = 2$:


$\mathbf{v}_2$
$\cos(\theta_2)\mathbf{v}_2$
$\mathbf{u} = \cos(\theta_1)\mathbf{v}_1 + \cos(\theta_2)\mathbf{v}_2$
θ_2
θ_1
$\mathbf{v}_1$
$\cos(\theta_1)\mathbf{v}_1$

Interlude: matrix multiplications

If $\mathbf{A}_{m\times n} = \begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_n \\ | & | & & | \end{bmatrix}$ and $\mathbf{B}_{n\times p} = \begin{bmatrix} -\mathbf{b}_1^T- \\ -\mathbf{b}_2^T- \\ \vdots \\ -\mathbf{b}_n^T- \end{bmatrix}$, with

$\mathbf{a}_j \in \mathbb{R}^m$ and $\mathbf{b}_j \in \mathbb{R}^p$, $j = 1, \dots, n$, then

$$AB = \sum_{j=1}^{n} \mathbf{a}_j \mathbf{b}_j^T$$

Example

$$\left[\begin{array}{c|c} 1 & 3 \\ 2 & 4 \end{array}\right] \left[\begin{array}{ccc} 1 & 2 & 0 \\ \hline 3 & -1 & 1 \end{array}\right] = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \begin{bmatrix} 1 & 2 & 0 \end{bmatrix} + \begin{bmatrix} 3 \\ 4 \end{bmatrix} \begin{bmatrix} 3 & -1 & 1 \end{bmatrix}$$
$$= \begin{bmatrix} 10 & -1 & 3 \\ 14 & 0 & 4 \end{bmatrix}$$

Note that if $\mathbf{b} = (\beta_1, \dots, \beta_n)$ one gets, $\mathbf{Ab} = \mathbf{A}\begin{bmatrix} \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} = \sum_{j=1}^n \beta_j \mathbf{a}_j$

Eigenvalue decomposition of a symmetric matrix

Theorem

Let $\mathbf{A}$ *be a symmetric matrix* ($\mathbf{A}^T = \mathbf{A}$) *of order* p. *Then we can find matrices* $\mathbf{V}_{p\times p}$ *and* $\mathbf{\Lambda}_{p\times p}$, *such that*

$$\boxed{\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T} \tag{2}$$

where:

- $\mathbf{V} = [\, \mathbf{v}_1 \;\; \mathbf{v}_2 \;\; \cdots \;\; \mathbf{v}_p \,]$ *verify* $\mathbf{V}^T\mathbf{V} = \mathbf{V}\mathbf{V}^T = \mathbf{I}_p$: *matrix of (unit and pairwise orthogonal) eigenvectors of* $\mathbf{A}$
- $\mathbf{\Lambda} = diag(\lambda_1, \lambda_2, \ldots, \lambda_p)$ *with* $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$: *diagonal matrix containing the corresponding (real) eigenvalues of* $\mathbf{A}$ ($\mathbf{A}\mathbf{v}_i = \lambda_i \mathbf{v}_i$)

Using the decomposition of a matrix product in terms of sums of columns and rows products described in slide 12, we can rewrite (2) as,

$$\boxed{\mathbf{A} = \lambda_1 \mathbf{v}_1 \mathbf{v}_1^T + \lambda_2 \mathbf{v}_2 \mathbf{v}_2^T + \cdots + \lambda_p \mathbf{v}_p \mathbf{v}_p^T\,,}$$

which is called the **spectral decomposition** of $\mathbf{A}$

Singular value decomposition of an arbitrary matrix

Theorem (Compact SVD)

Let $\mathbf{A}$ *be matrix of type* $N \times p$ *and rank* r. *Then we can find matrices* $\mathbf{U}_{N\times r}$, $\mathbf{\Delta}_{r\times r}$ *and* $\mathbf{V}_{p\times r}$, *such that*

$$\boxed{\mathbf{A} = \mathbf{U}\mathbf{\Delta}\mathbf{V}^T} \tag{3}$$

where:

- $\mathbf{U} = [\, \mathbf{u}_1 \;\; \cdots \;\; \mathbf{u}_r \,]$ *verify* $\mathbf{U}^T\mathbf{U} = \mathbf{I}_r$: *matrix of (unit and pairwise orthogonal) left singular vectors of* $\mathbf{A}$
- $\mathbf{V} = [\, \mathbf{v}_1 \;\; \cdots \;\; \mathbf{v}_r \,]$ *verify* $\mathbf{V}^T\mathbf{V} = \mathbf{I}_r$: *matrix of (unit and pairwise orthogonal) right singular vectors of* $\mathbf{A}$
- $\mathbf{\Delta} = diag(\delta_1, \ldots, \delta_r)$ *with* $\delta_1 \geq \cdots \geq \delta_r > 0$: *diagonal matrix of the nonzero singular values of* $\mathbf{A}$ ($\mathbf{A}\mathbf{v}_i = \delta_i \mathbf{u}_i$ *and* $\mathbf{A}^T\mathbf{u}_i = \delta_i \mathbf{v}_i$)

Using the results of slide 12 we can rewrite (3) as,

$$\boxed{\mathbf{A} = \delta_1 \mathbf{u}_1 \mathbf{v}_1^T + \delta_2 \mathbf{u}_2 \mathbf{v}_2^T + \cdots + \delta_r \mathbf{u}_r \mathbf{v}_r^T\,,}$$

which is called the **singular value decomposition** of $\mathbf{A}$

PRINCIPAL COMPONENT ANALYSIS

Some statistics - univariate case

Given the vectors $\mathbf{x}_i = (x_1, \ldots, x_N)$ and $\mathbf{y} = (y_1, \ldots, y_N) \in \mathbb{R}^N$, of observations of two variables across N individuals, we define:

- (sample) mean of $\mathbf{x}$:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- (sample) variance of $\mathbf{x}$:

$$s_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

- (sample) covariance between $\mathbf{x}$ and $\mathbf{y}$:

$$s_{xy}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{N-1}(\mathbf{x}^*)^T\mathbf{y}^*,$$

  where $\mathbf{x}^* = (x_1 - \bar{x}, \ldots, x_N - \bar{x})$ and $\mathbf{y}^* = (y_1 - \bar{y}, \ldots, y_N - \bar{y})$ are the corresponding centered vectors

- (sample) linear correlation coefficient between $\mathbf{x}$ and $\mathbf{y}$:

$$r_{xy}^2 = \frac{s_{xy}^2}{s_x s_y}$$

Variable's cloud and individual's cloud

Let $\mathbf{X}_{N\times p} = [\, x_{ij} \,]$ be a data matrix where:

- each column of $\mathbf{X}$ represents the observations of some variable across N individuals. We write,

  $$\mathbf{X}_{N\times p} = [\mathbf{x}_1 \;\cdots\; \mathbf{x}_p] \quad \text{with} \quad \mathbf{x}_j = (x_{1j}, \ldots, x_{Nj}) \in \mathbb{R}^N, \quad j = 1, \ldots, p$$

  The columns $\mathbf{x}_1, \ldots, \mathbf{x}_p$ define a cloud of p points in $\mathbb{R}^N$ - variable's cloud

- each row of $\mathbf{X}$ represents the observations of a single individual w.r.t. p variables:

  $$\mathbf{X}^T_{p\times N} = [\mathbf{x}^1 \;\cdots\; \mathbf{x}^N] \quad \text{with} \quad \mathbf{x}^i = (x_{i1}, \ldots, x_{ip}) \in \mathbb{R}^p, \quad i = 1, \ldots, N$$

  The rows $\mathbf{x}^1, \ldots, \mathbf{x}^N$ define a cloud of N points in $\mathbb{R}^p$ - individuals's cloud

Row and column space

Summary statistics - multivariate case

- The sample mean vector of X, $\mathbf{x}^G = (\bar{x}_1, \ldots, \bar{x}_p)$ with $\bar{x}_j = \frac{1}{N}\sum_{i=1}^{N} x_{ij}$, is the individuals's cloud center of gravity, that is,

$$\mathbf{x}^G = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}^i \in \mathbb{R}^p$$

- The (sample) covariance matrix of $\mathbf{X}$ is

$$\mathbf{S} = [s_{jk}^2] = \frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{x}^i - \mathbf{x}^G)(\mathbf{x}^i - \mathbf{x}^G)^T,$$

with the covariance between variables j and k equal to

$$s_{jk}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$$

- The inertia (total variability) of $\mathbf{X}$ is given by $\mathrm{tr}(\mathbf{S}) = s_{11}^2 + \cdots + s_{kk}^2$, that is,

$$\frac{1}{N-1}\sum_{i=1}^{N}\|\mathbf{x}^i - \mathbf{x}^G\|^2 = \frac{1}{2N(N-1)}\sum_{i,j=1}^{N}\|\mathbf{x}^i - \mathbf{x}^j\|^2$$

Centered data matrix and covariance

For each $j = 1, \ldots, p$, the centered vector of the N observations of variable j is

$$\mathbf{x}_j^* = (x_{1j} - \bar{x}_j, \ldots, x_{Nj} - \bar{x}_j) \in \mathbb{R}^N$$

The (sample) covariance s_{jk}^2 can then be written, using the centered variables $\mathbf{x}_j^*$ and $\mathbf{x}_k^*$, as a simple inner product (in $\mathbb{R}^N$) divided by $N-1$,

$$s_{jk}^2 = cov(\mathbf{x}_j, \mathbf{x}_k) = \frac{1}{N-1}(\mathbf{x}_j^*)^T\mathbf{x}_k^* \tag{4}$$

Likewise, if we define the centered data matrix as

$$\mathbf{X}^* = [\,\mathbf{x}_1^* \;\cdots\; \mathbf{x}_p^*\,],$$

i.e.,

$$(\mathbf{X}^*)^T = [\,(\mathbf{x}^1 - \mathbf{x}^G) \;\;\cdots\;\; (\mathbf{x}^N - \mathbf{x}^G)\,],$$

the covariance matrix $\mathbf{S} = [s_{jk}^2]$ of $\mathbf{X}$ can be written as

$$\boxed{\mathbf{S} = \tfrac{1}{N-1}(\mathbf{X}^*)^T\mathbf{X}^*}$$

Standardized data matrix and correlation

For each $j = 1, \ldots, p$, the vector of the N observations of standardized variable j is

$$\mathbf{z}_j = \left(\frac{x_{1j} - \bar{x}_j}{s_j}, \ldots, \frac{x_{Nj} - \bar{x}_j}{s_j} \right) = \left(\frac{x^*_{1j}}{s_j}, \ldots, \frac{x^*_{Nj}}{s_j} \right) \in \mathbb{R}^N$$

and we obtain the corresponding standardized data matrix,

$$\mathbf{Z} = [\, \mathbf{z}_1 \ \cdots \ \mathbf{z}_p \,]$$

- The (sample) linear correlation coefficient between variables j and k is

$$r_{jk} = \frac{s^2_{jk}}{s_j s_k} = \frac{1}{N-1} \sum_{i=1}^{N} \left(\frac{x_{ij} - \bar{x}_j}{s_j} \right) \left(\frac{x_{ik} - \bar{x}_k}{s_k} \right) = \frac{1}{N-1} \mathbf{z}_j^T \mathbf{z}_k$$

- Hence the (sample) correlation matrix $\mathbf{R} = [r_{ij}]$ of $\mathbf{X}$ equals the covariance matrix of the standardized data matrix, i.e.,

$$\boxed{\mathbf{R} = \tfrac{1}{N-1} \mathbf{Z}^T \mathbf{Z}}$$

- The total variability (inertia) of $\mathbf{Z}$ is

$$\text{tr}(\mathbf{R}) = r_{11} + \cdots + r_{pp} = p$$

Principal component analysis - motivation

- Principal component analysis (PCA) is a statistical multvariate method that aims to reduce the dimensionality of a dataset $\mathbf{X}$ while preserving its information, i.e., the data set total variability, as much as possible
- This goal is achieved by defining a set of uncorrelated variables, called **principal components**, that are linear combinations of the original (or standardized) variables, in such a way that the first few principal components explain the maximum proportion of the data set total variability
- The dimension reduction is (particularly) effective when the original variables are (highly) correlated
- PCA is probably the most widely used multivariate statistical method

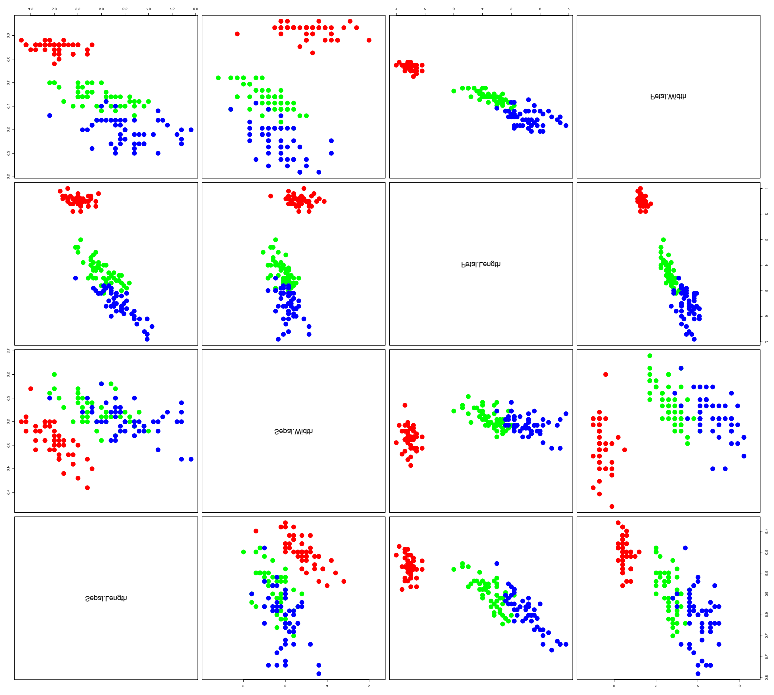
Example: iris flower data set

- The well known iris flower data set consists of 4 measurements, sepal and petal lengths and widths, *SL*,*SW*,*PL*,*PW* (in cm), containing 50 iris flowers of each one of the following three species, setosa, versicolor and virginica

- Hence the iris flower dataset defines a cloud of 150 points in $\mathbb{R}^4$. We can try to have a geometrical grasp of the shape of this 4-dimensional cloud by projecting it on a two dimensional space (plane), using all possible combinations of two variables

Example: iris flower data set

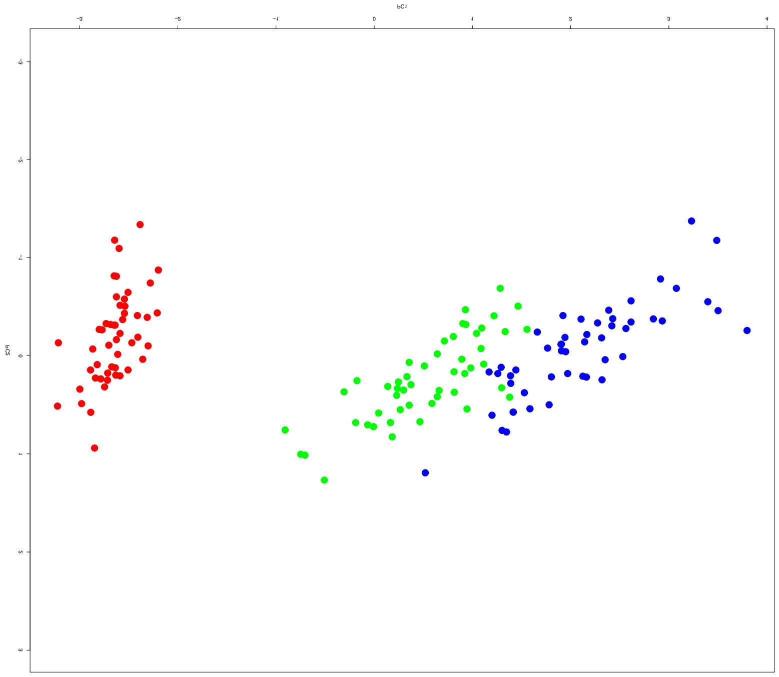
```
pairs(iris[-5],asp=TRUE,pch=16,col=c(rep("red",50),
      rep("green",50),rep("blue",50)))
```



Best 2-dimensional representation using PCA

- Another approach is to define new synthetic uncorrelated variables that are linear combinations of the original iris flowers measurements, the so-called principal components (PC), in such a way each PC explains, as much as possible, of the total dataset variability
- The projection of the cloud of iris flowers on the plane associated with the first two PCs, called principal factorial plane (PFP), explains 98.1% of the iris dataset variability and thus provides an excellent 2-dimensional portray of the original cloud of iris flowers
- This is actually the best representation among all 2-dimensional representations of the iris flower dataset, in the sense that it is the 2-dimensional representation that retains the largest amount of the dataset variability

Best two-dimensional representation of the iris flowers



Eigenvalues of the covariance matrix

Let $\mathbf{X}_{N\times p}$ be a data matrix and $\mathbf{S} = \frac{1}{N-1}(\mathbf{X}^*)^T\mathbf{X}^*$ the corresponding covariance matrix. Then:

- $\mathbf{S}$ is symmetric ($\mathbf{S}^T = \mathbf{S}$)
- $\mathbf{x}^T\mathbf{S}\mathbf{x}$ is a **semi-definite positive** quadratic form, that is,

$$\mathbf{x}^T\mathbf{S}\mathbf{x} \geq 0, \quad \forall \mathbf{x} \in \mathbb{R}^p$$

- the eigenvalues $\lambda_1, \ldots, \lambda_p$ of $\mathbf{S}$ are nonnegative real numbers and we may assume that

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$$

- If moreover all variables are globally non-correlated then all eigenvalues of $\mathbf{S}$ are strictly positive real numbers. In this case $\mathbf{S}$ is invertible, $\mathbf{x}^T\mathbf{S}\mathbf{x}$ is a definite positive quadratic form, which amounts to say that

$$\mathbf{x}^T\mathbf{S}\mathbf{x} > 0, \quad \forall \mathbf{x} \in \mathbb{R}^p,\ \mathbf{x} \neq \vec{\mathbf{0}}$$

Linear combinations

Consider again the data set $\mathbf{X}_{N\times p} = [\,\mathbf{x}_1 \ \cdots \ \mathbf{x}_p\,]$ containing the observations of p variables across N individuals.

A **linear combination** of the p columns $\mathbf{x}_1, \ldots, \mathbf{x}_p$ of X as the form

$$\mathbf{y} = \alpha_1\mathbf{x}_1 + \cdots + \alpha_p\mathbf{x}_p = [\,\mathbf{x}_1 \ \cdots \ \mathbf{x}_p\,]\begin{bmatrix}\alpha_1\\ \vdots\\ \alpha_p\end{bmatrix} = \mathbf{X}\mathbf{a},$$

where

$$\mathbf{a} = (\alpha_1, \ldots, \alpha_p) = \begin{bmatrix}\alpha_1\\ \vdots\\ \alpha_p\end{bmatrix} \in \mathbb{R}^p,$$

is the vector of coefficients (**loadings**) (see slide 12)

Covariance between linear combinations

Given $\mathbf{a}, \mathbf{b} \in \mathbb{R}^p$, the covariance between the linear combinations $\mathbf{Xa}$ and $\mathbf{Xb}$ is

$$cov(\mathbf{Xa}, \mathbf{Xb}) = \mathbf{a}^T \mathbf{S} \mathbf{b} \tag{5}$$

Actually, using (4) of slide 20 we have,

$$\begin{aligned} cov(\mathbf{Xa}, \mathbf{Xb}) &= \frac{1}{N-1}[(\mathbf{Xa})^*]^T (\mathbf{Xb})^* \overset{exercise}{=} \frac{1}{N-1}(\mathbf{X}^*\mathbf{a})^T \mathbf{X}^*\mathbf{b} \\ &= \frac{1}{N-1}\mathbf{a}^T (\mathbf{X}^*)^T \mathbf{X}^*\mathbf{b} = \mathbf{a}^T \frac{1}{N-1}(\mathbf{X}^*)^T \mathbf{X}^*\mathbf{b} \\ &= \mathbf{a}^T \mathbf{S} \mathbf{b} \end{aligned}$$

In particular, $var(\mathbf{Xa}) = \mathbf{a}^T \mathbf{S} \mathbf{a}$

Exercise

Prove that centering a linear combination of variables $\mathbf{x}_1, \dots, \mathbf{x}_p$ is equivalent to the linear combination of the centered variables $\mathbf{x}_1^*, \dots, \mathbf{x}_p^*$ with the same coefficients, i.e.,

$$(\mathbf{Xa})^* = (\alpha_1 \mathbf{x}_1 + \cdots + \alpha_p \mathbf{x}_p)^* = \alpha_1\, \mathbf{x}_1^* + \cdots + \alpha_p\, \mathbf{x}_p^* = \mathbf{X}^*\mathbf{a},$$

where $\mathbf{X} = [\, \mathbf{x}_1 \ \cdots \ \mathbf{x}_p \,]$, $\mathbf{X}^* = [\, \mathbf{x}_1^* \ \cdots \ \mathbf{x}_p^* \,]$ *and* $\mathbf{a} = (\alpha_1, \dots, \alpha_p) \in \mathbb{R}^p$

First principal component - formulation

To define the first principal component we seek a linear combination of the p observed variables $\mathbf{x}_1, \dots, \mathbf{x}_p$ that maximizes the variance, that is, we want to solve the following problem:

$$\text{determine } \mathbf{a} \in \mathbb{R}^p \text{ such that } \ var(\mathbf{Xa}) = \mathbf{a}^T \mathbf{S} \mathbf{a} \ \text{ is maximum}$$

Without further restrictions on vector $\mathbf{a}$ the problem is ill-posed since if we multiply the vector of coefficients $\mathbf{a}$ by a scalar λ we obtain

$$var(\mathbf{X}(\lambda\mathbf{a})) = \lambda\mathbf{a}^T \mathbf{S} \lambda\mathbf{a} = \lambda^2 \mathbf{a}^T \mathbf{S} \mathbf{a} = \lambda^2 var(\mathbf{Xa}),$$

which shows that the variance of a linear combination can be arbitrarily large. To overcome this issue we reformulate the problem as follows:

$$\boxed{\text{determine } \mathbf{a} \in \mathbb{R}^p \text{ with } \|\mathbf{a}\| = 1 \ : \ var(\mathbf{Xa}) = \mathbf{a}^T \mathbf{S} \mathbf{a} \text{ is maximum}} \tag{6}$$

The previous problem can be equivalently formulated as the problem of maximizing the so-called **Rayleigh-Ritz ratio** *(cf. slides Prof. Cadima)*

$$\boxed{\textit{determine } \mathbf{a} \in \mathbb{R}^p \setminus \{\vec{0}\} \ : \ \frac{\mathbf{a}^T \mathbf{S} \mathbf{a}}{\mathbf{a}^T \mathbf{a}} \ \textit{is maximum}} \tag{7}$$

First principal component (cont.)

The covariance matrix $\mathbf{S}$ admits a spectral decomposition (see slide 13) of the form,

$$\mathbf{S} = \lambda_1 \mathbf{v}_1 \mathbf{v}_1^T + \lambda_2 \mathbf{v}_2 \mathbf{v}_2^T + \cdots + \lambda_p \mathbf{v}_p \mathbf{v}_p^T \tag{8}$$

where $\mathbf{v}_1, \ldots, \mathbf{v}_p \in \mathbb{R}^p$ are unit and pairwise orthogonal eigenvectors of $\mathbf{S}$ associated to (sorted) real eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$

By the results of slide 11, we have for all $\mathbf{a} \in \mathbb{R}^p$, $\|\mathbf{a}\| = 1$,

$$\mathbf{a} = \cos(\theta_1)\mathbf{v}_1 + \cdots + \cos(\theta_p)\mathbf{v}_p, \tag{9}$$

with

$$\cos^2\theta_1 + \cdots + \cos^2\theta_p = 1, \tag{10}$$

where θ_i denotes the angle between the vectors $\mathbf{a}$ and $\mathbf{v}_i$, , $i = 1, \ldots, p$

First principal component (cont.)

Applying (8), (9) and (10) from the previous slide, along with relations $\lambda_1 \geq \cdots \geq \lambda_p > 0$, $\|\mathbf{v}_i\|^2 = \mathbf{v}_i^T \mathbf{v}_i = 1$ for all i and $\mathbf{v}_i^T \mathbf{v}_j = 0$, $i \neq j$, we obtain by straightforward computations (since all inner products envolving $\mathbf{v}_i$ and $\mathbf{v}_j$, $j \neq i$, vanish),

$$\begin{aligned} \mathbf{a}^T\mathbf{S}\mathbf{a} &= \lambda_1 \cos^2\theta_1 + \cdots + \lambda_p \cos^2\theta_p \\ &\leq \lambda_1 \cos^2\theta_1 + \cdots + \lambda_1 \cos^2\theta_p \\ &= \lambda_1(\cos^2\theta_1 + \cdots + \cos^2\theta_p) = \lambda_1 \end{aligned}$$

Thus $var(\mathbf{X}\mathbf{a}) = \mathbf{a}^T\mathbf{S}\mathbf{a} \leq \lambda_1$ (the largest eigenvalue of $\mathbf{S}$). Taking $\mathbf{a} = \mathbf{v}_1$, we get

$$\mathbf{a}^T\mathbf{S}\mathbf{a} = \mathbf{v}_1^T\mathbf{S}\mathbf{v}_1 = \lambda_1 \underbrace{\cos^2\theta_1}_{1} + \lambda_2 \underbrace{\cos^2\theta_2}_{0} + \cdots + \lambda_1 \underbrace{\cos^2\theta_p}_{0} = \lambda_1$$

The maximum variance of a linear combination $\mathbf{X}\mathbf{a}$, with unit vector of coefficients $\mathbf{a}$, of $\mathbf{x}_1, \ldots, \mathbf{x}_p$, is attained along the direction of a unit eigenvector $\mathbf{v}_1$ of $\mathbf{S}$ associated with the largest eigenvalue λ_1. Hence the first principal component is

PC_1 : $\mathbf{y}_1 = \mathbf{X}\mathbf{v}_1$ *with maximum variance* λ_1

The larger the value of λ_1, the more the cloud of points is elongated along the PC_1
O critério da ACP (maximizar variância) corresponde a procurar combinações lineares dos vectores de comprimento máximo (com soma 1 de quadrados dos coeficientes).

Second principal component

For the second principal component PC$_2$, we seek a linear combination of $\mathbf{x}_1, \dots, \mathbf{x}_p$, $\mathbf{Xa}$, with $\|\mathbf{a}\| = 1$, that maximizes $var(\mathbf{Xa}) = \mathbf{a}^T\mathbf{Sa}$ and is uncorrelated with PC$_1$, i.e.,

$$cov(\mathbf{Xa}, \mathbf{Xv}_1) = \mathbf{a}^T\mathbf{Sv}_1 = \lambda_1\mathbf{a}^T\mathbf{v}_1 = 0.$$

. Hence we want to solve the following problem (assuming $\lambda_1 > 0$):

$$\boxed{\text{determine } \mathbf{a} \in \mathbb{R}^p \text{ with } \begin{cases} \|\mathbf{a}\| = 1 \\ \mathbf{a} \perp \mathbf{v_1} \end{cases} : \; var(\mathbf{Xa}) = \mathbf{a}^T\mathbf{Sa} \text{ is maximum}}$$

Since $\mathbf{a} \perp \mathbf{v}_1 \Leftrightarrow \cos\theta_1 = 0$, we seek $\mathbf{a} = \cos(\theta_2)\mathbf{v}_2 + \cdots + \cos(\theta_p)\mathbf{v}_p$, with $\cos^2(\theta_2) + \cdots + \cos^2(\theta_p) = 1$ and we obtain similarly,

$$\begin{array}{rcl} \mathbf{a}^T\mathbf{Sa} & = & \lambda_2\cos^2\theta_2 + \cdots + \lambda_p\cos^2\theta_p \\ & \leq & \lambda_2(\cos^2\theta_2 + \cdots + \cos^2\theta_p) = \lambda_2 \end{array}$$

Taking $\mathbf{a} = \mathbf{v}_2$ (a unit eigenvector of $\mathbf{S}$ associated with the second largest eigenvalue λ_2 and orthogonal to $\mathbf{v}_1$), one gets

$$\mathbf{a}^T\mathbf{Sa} = \lambda_2$$

The second PC is thus defined by a unit eigenvector $\mathbf{v}_2$ of $\mathbf{S}$, associated with the second largest eigenvalue λ_2 and orthogonal to the vector $\mathbf{v}_1$

$$PC_2 : \quad \mathbf{y}_2 = \mathbf{Xv}_2 \quad \textit{with maximum variance equal to } \lambda_2$$

Principal components

In general, to define the j-th principal component PC$_j$, $j = 2, \dots, p$, we seek a linear combination of the p original observed variables, that maximizes the variance and is uncorrelated with PC$_1, \dots,$ PC$_{j-1}$ (assuming $\lambda_{j-1} > 0$:

$$\boxed{\text{determine } \mathbf{a} \in \mathbb{R}^p \text{ with } \left\{ \begin{array}{ccc} \|\mathbf{a}\| & = & 1 \\ \mathbf{a} & \perp & \mathbf{v_1} \\ & \vdots & \\ \mathbf{a} & \perp & \mathbf{v_{j-1}} \end{array} \right| \; var(\mathbf{Xa}) = \mathbf{a}^T\mathbf{Sa} \text{ is maximum}} \qquad (11)$$

We construct in this way a collection of p principal components

$$\mathbf{y}_1 = \mathbf{Xv}_1, \quad \mathbf{y}_2 = \mathbf{Xv}_2, \quad \dots, \quad \mathbf{y}_p = \mathbf{Xv}_p$$

with maximum variances,

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p > 0,$$

where $\mathbf{v}_1, \dots, \mathbf{v}_p$ are unit and pairwise orthogonal eigenvectors of $\mathbf{S}$, respectively associated to $\lambda_1, \dots, \lambda_p$, i.e., for all $j, k = 1, \dots, p$, $k \neq j$ we have

$$\|\mathbf{v}_j\| = 1, \qquad \mathbf{v}_j \perp \mathbf{v}_k, \qquad \mathbf{Sv}_j = \lambda_j\mathbf{v}_j$$

Vector of loadings

The vector $\mathbf{v}_j$ defining the j-th principal component $\mathbf{y}_j = \mathbf{X}\mathbf{v}_j$, contains the coefficients, also called loadings, of the j-th principal component w.r.t. the original observed variables $\mathbf{x}_1, \dots, \mathbf{x}_p$. In other words, writing the vector of loadings as $\mathbf{v}_j = (\alpha_1, \dots, \alpha_p)$ we obtain,

$$\mathbf{y}_j = \alpha_1 \mathbf{x}_1 + \cdots + \alpha_p \mathbf{x}_p$$

- *If the p eigenvalues of the covariance matrix $\mathbf{S}$ are pairwise distinct, i.e., $\lambda_1 > \cdots > \lambda_p > 0$, the vector of loadings defining each PC is unique up to sign: if $\mathbf{y}_j = \mathbf{X}\mathbf{v}_j$ is a solution of (11) of slide 34, then $\mathbf{y}'_j = \mathbf{X}(-\mathbf{v}_j)$ is also a solution of (11) - this is the most common situation*
- *If there are repeated eigenvalues of $\mathbf{S}$ the PCs associated with repeated eigenvalues are not uniquely determined. Actually, the vectors of loadings defining these PCs can arise from any orthonormal base of the eigenspace associated with the repeated eigenvalue and therefore can be defined in infinitely many distinct ways*

Scores matrix

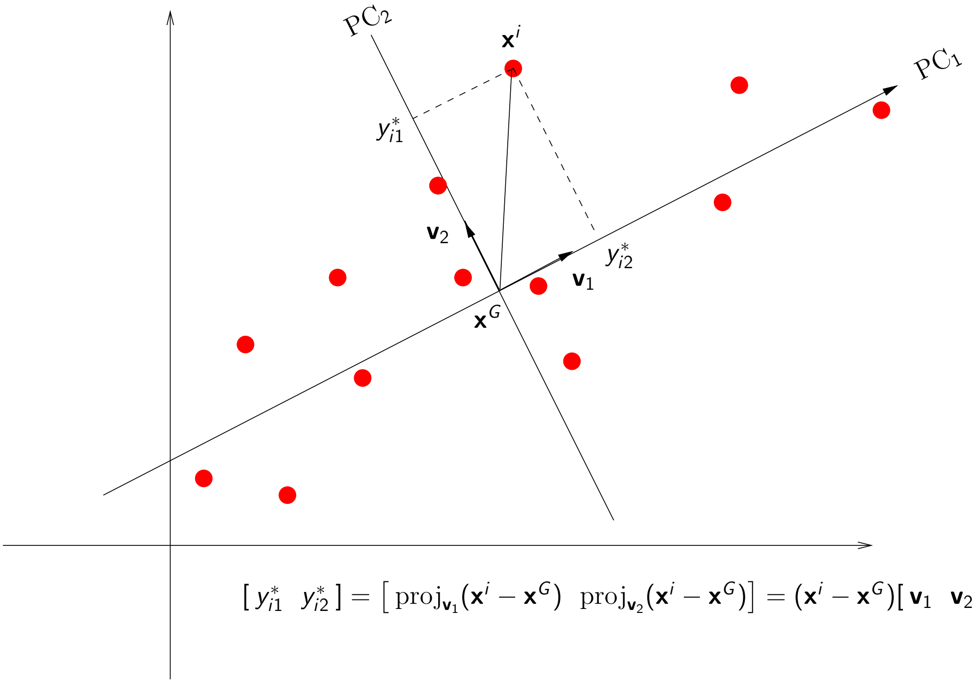
Recall that,

- $\mathbf{X}_{N\times p} = [x_{ij}]$ is the original data matrix of the p observed variables across N individuals
- $\mathbf{X}^T = [\,\mathbf{x}^1 \ \cdots \ \mathbf{x}^N\,]$, with $\mathbf{x}^i = (x_{i1}, \dots, x_{ip})$ the i-th row of $\mathbf{X}$, i.e., the coordinates of i-individual in the cloud of N points of $\mathbf{R}^p$
- $\mathbf{x}^G = (\bar{x}_1, \dots, \bar{x}_p)$ is the center of gravity (also called barycenter) of the cloud of individuals
- $\mathbf{X}^* = [x^*_{ij}]$ is the centered data matrix, where $x^*_{ij} = x_{ij} - \bar{x}_j$
- $\mathbf{x}^i - \mathbf{x}^G = (x^*_{i1}, \dots, x^*_{ip})$ the i-th row of $\mathbf{X}^*$, i.e., the vector of the coordinates of individual i in the centered cloud of N points
- $\mathbf{V} = [\,\mathbf{v}_1 \ \cdots \mathbf{v}_p\,]$ is the matrix of loadings

The matrix $\mathbf{Y}^* = [y^*_{ij}] = \mathbf{X}^*\mathbf{V}$ is called scores matrix: the rows of $\mathbf{Y}^*$ correspond to the vectors of coordinates, also called (factor) scores, of the N individuals w.r.t the new coordinate axes defined by the vectors of loadings of the PCs

The column j of $\mathbf{Y}^*$, $\mathbf{y}^*_j$, contains the values of the (centered) cloud of N individuals w.r.t the new sinthetic variable $\mathbf{y}_j$ that defined the PC_j

Scores of individual i when $p = 2$



$$[\, y_{i1}^* \;\; y_{i2}^* \,] = \left[\, \mathrm{proj}_{\mathbf{v}_1}(\mathbf{x}^i - \mathbf{x}^G) \;\; \mathrm{proj}_{\mathbf{v}_2}(\mathbf{x}^i - \mathbf{x}^G) \right] = (\mathbf{x}^i - \mathbf{x}^G)[\, \mathbf{v}_1 \;\; \mathbf{v}_2 \,]$$

Covariance of the scores matrix

- The covariance matrix of the (centred) scores matrix $\mathbf{Y}^*$, is

$$\begin{aligned} cov(\mathbf{Y}^*) &= cov(\mathbf{X}^*\mathbf{V}) = \frac{1}{N-1}(\mathbf{X}^*\mathbf{V})^T(\mathbf{X}^*\mathbf{V}) \\ &= \mathbf{V}^T\frac{1}{N-1}(\mathbf{X}^*)^T\mathbf{X}^*\mathbf{V} = \mathbf{V}^T\mathbf{S}\mathbf{V} = \mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_p) \end{aligned}$$

- The total variation of $\mathbf{Y}^*$, i.e., the dataset total variability is

$$\sum_{j=1}^{p} \mathrm{var}(\mathbf{y}_j^*) = \sum_{j=1}^{p} \lambda_j = \mathrm{tr}(\mathbf{\Lambda}) = \mathrm{tr}(\mathbf{S}) = \sum_{j=1}^{p} \mathrm{var}(\mathbf{x}_j)$$

- The quality of the reduction obtained by keeping the first k PCs $(1 \le k \le p)$ is assessed by the proportion of variability explained by the first k PCs:

$$\frac{\lambda_1 + \cdots + \lambda_k}{\lambda_1 + \cdots + \lambda_p}$$

Covariance and correlation

- The covariance between the variable $\mathbf{x}_j$ and the PC $\mathbf{y}_k$ is

$$\begin{aligned} cov(\mathbf{x}_j, \mathbf{y}_k) &= \frac{1}{N-1}[(\mathbf{X}\mathbf{e}_j)^*]^T(\mathbf{X}\mathbf{v}_k)^* = \frac{1}{N-1}(\mathbf{X}^*\mathbf{e}_j)^T(\mathbf{X}^*\mathbf{v}_k) \\ &= \mathbf{e}_j^T\frac{1}{N-1}(\mathbf{X}^*)^T\mathbf{X}^*\mathbf{v}_k = \mathbf{e}_j^T\mathbf{S}\mathbf{v}_k = \mathbf{e}_j^T\lambda_k\mathbf{v}_k \\ &= \lambda_k\mathbf{e}_j^T\mathbf{v}_k = \lambda_k\mathbf{v}_{jk} \end{aligned}$$

  where $\mathbf{v}_{jk} = \mathbf{e}_j^T\mathbf{v}_k$ is j-th component of $\mathbf{v}_k$, i.e., (j,k)-entry of the loadings matrix $\mathbf{V}$

- The correlation between $\mathbf{x}_j$ and $\mathbf{y}_k$ is

$$cor(\mathbf{x}_j, \mathbf{y}_k) = \frac{cov(\mathbf{x}_j, \mathbf{y}_k)}{\sqrt{\mathbf{x}_j}\sqrt{\mathbf{y}_k}} = \frac{\lambda_k\mathbf{v}_{jk}}{s_j\sqrt{\lambda_k}} = \frac{\sqrt{\lambda_k}\mathbf{v}_{jk}}{s_j}$$

- The contribution of individual i to the construction of PC$_k$ is the proportion of the variance of PC$_k$ that is due to individual i (in %):

$$ctr_{i,k} = \frac{\frac{1}{N-1}(y_{i,k}^*)^2}{\lambda_k} \times 100 = \frac{(y_{i,k}^*)^2}{\sum_{j=1}^N (y_{j,k}^*)^2} \times 100$$

Example: iris flower dataset revisited

R

```
X=iris[-5] # non standardized
head(X)
iris.acp<-prcomp(X) # performs PCA on the covariance
matrix
summary(iris.acp)
iris.acp$sdev # std accounted by the PCs
sum(iris.acp$sdev[1]^2) # total variance
iris.acp$rot # matrix of loadings
iris.acp$x # matrix of scores
plot(iris.acp$x[,1:2],asp=TRUE,pch=16,col=c(rep("red",50),
rep("green",50),rep("blue",50))) #
```

Importance of the PC components

The R command `summary(iris.acp)` gives, for each j, the standard deviation $\sqrt{\lambda_j}$ associated with PC$_j$, the proportion of the total variance explained by PC$_j$, $\frac{\lambda_j}{\sum_k \lambda_k}$, and the cumulative variance explained by the first j PCs:

	PC1	PC2	PC3	PC4
Standard deviation	2.0563	0.49262	0.2797	0.15439
Proportion of Variance	0.9246	0.05307	0.0171	0.00521
Cumulative Proportion	0.9246	0.97769	0.9948	1.00000

Thus we have that:

- The cloud of points projected on the line associated with the first PC explains about 92% of the dataset's total variability
- The cloud of points projected on the plane associated with the first two PCs (principal factorial plane - PFP) explains about 98% of the total variability of the dataset,
- and so on. . .

More on the R command `prcomp`

`iris.acp$sdev` give, for each j, the standard deviation $\sqrt{\lambda_j}$ associated with PC$_j$:
2.0562689 0.4926162 0.2796596 0.1543862

`sum(iris.acp$sdev[1]^2)` gives the dataset total variance: 4.572957

`iris.acp$rotation` returns the matrix of loadings, where column j contains the coefficients of the PC$_j$, $\mathbf{y}_j$, written as linear combination of the original observed variables $\mathbf{x}_1, \dots, \mathbf{x}_4$:

	PC1	PC2	PC3	PC4
Sepal.Length	0.3614	-0.6566	0.5820	0.3155
Sepal.Width	-0.0845	-0.7302	-0.5979	-0.3197
Petal.Length	0.8567	0.1734	-0.0762	-0.4798
Petal.Width	0.3583	0.0755	-0.5458	0.7537

The first PC (for instance), is a linear combination of the observed measurements as:

$$\begin{aligned}\mathbf{y}_1 &= 0.3614\,\text{Sepal.Length} - 0.0845\,\text{Sepal.Width} + 0.8567\,\text{Petal.Length} + 0.3583\,\text{Petal.Widt}\\ &\approx 0.3614\,\text{Sepal.Length} + 0.8567\,\text{Petal.Length} + 0.3583\,\text{Petal.Width}\end{aligned}$$

which represents a kind of overall measurement of the iris flowers that explains a large amount ($\geq 90\%$) of the total variability of the iris dataset

More on the R command prcomp - loadings

The columns of the loading matrix are unit eigenvectors of **S** and pairwise orthogonal

R

```
V<-iris.acp$rotation
S <- cov(S)
round(t(V)%*% V,10) # gives the identity matrix
v1 <- V[,1]
lambda1 <- iris.acp$sdev[1]^2
S %*% v1
lambda1%*% v1
```

More on the R command prcomp - scores

`iris.acp$x` returns the matrix of factor scores, where each row i contains coordinates of the individual i w.r.t. the PCs, i.e., w.r.t. the new synthetic variables $\mathbf{y}_1, \ldots, \mathbf{y}_4$:

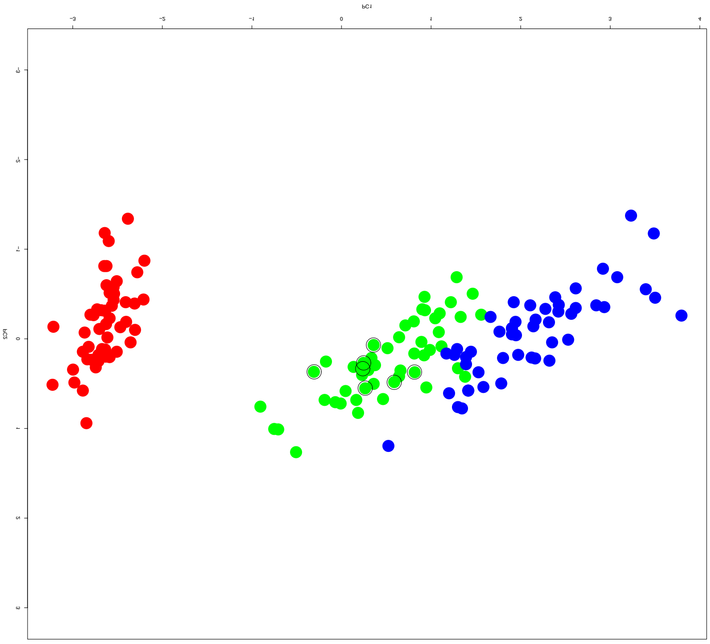
	PC1	PC2	PC3	PC4
	-2.68413	-0.31940	0.02791	0.00226
	-2.71414	0.17700	0.21046	0.09903
	-2.88899	0.14495	-0.01790	0.01997
	-2.74534	0.31830	-0.03156	-0.07558
	⋮	⋮	⋮	⋮

More on the R command `prcomp` - scores

R

```
N <- 150 X.G <- colMeans(X) # iris's cloud center of gravity
Xc <- scale(X,scale=FALSE) # centred iris data matrix
Yc <- Xc %*% V # scores matrix
head(Yc) ; head(iris.acp$x) # should be equal!
sum(iris.acp$x[,1]^2)/(N-1) ; iris.acp$sdev[1]^2
# contributions of each individual to the 1st PC
Yc[,1]*Yc[,1]/sum(Yc[,1]*Yc[,1])
vspace.25ex
# individuals with contribution above the average
Yc[,1]*Yc[,1]/sum(Yc[,1]*Yc[,1]>1/150
# quality of the representation of individual i in each PC
Yc[1,]*Yc[1,]/sum(Yc[1,]*Yc[1,])
# quality of the representation of individual i in the PFP
(Yc[1,1]*Yc[1,1]+Yc[1,2]*Yc[1,2])/sum(Yc[1,]*Yc[1,])
cos2<-matrix(0,ncol=4,nrow=150)
for (i in 1:150) { cos2[i,]<-Yc[i,]*Yc[i,]/sum(Yc[i,]*Yc[i,]) }
sort(rowSums(cos2[,1:2]))
order(rowSums(cos2[,1:2]))
plot(iris.acp$x[,1:2],pch=16,col=c(rep("red",50),
rep("green",50),rep("blue",50)),asp=TRUE)
points(iris.acp$x[rowSums(cos2[,1:2])<.7,1:2],pch=1)
```

Representation of the iris flower dataset in the PFP



Drawbacks of the PCA on the covariance matrix

- Variable(s) with much larger variance(s) tend to dominate the first principal component. Actually, by (6) in slide 30 the first PC is $\mathbf{Xa}$ such that $\mathbf{a} = (\alpha_1, \dots, \alpha_p)$ maximizes

$$var(\mathbf{Xa}) = \sum_{i=1}^{p} \alpha_i^2 var(\mathbf{x}_i) + 2\sum_{i<j} \alpha_i \alpha_j cov(\mathbf{x}_i, \mathbf{x}_j), \quad \text{with} \quad \|\mathbf{a}\| = \sum_{i=1}^{p} \alpha_i^2 = 1$$

- The PCs are invariant under orthogonal transformations of the variables (e.g. rotations), but not under differentiated change of scales in each variable. As a consequence the PCA is highly dependent on the units of measurements - this is a major drawback
- Another important drawback when there are distinct units of measurements is how to a interpret a PC if the PC is a linear combination of variables expressed in totally different units of measurements, say, for instance temperature and weight?

When the variables have different units of measurements or very different variances it is advisable or even mandatory to standardize (i.e., to center and reduce the variables to unit variance) prior to perform the PCA. This amounts to compute the eigenvectors of the correlation matrix of $\mathbf{X}$

PCA on the correlation matrix

Let $\mathbf{X}_{N\times p} = [\mathbf{x}_{ij}]$ be the usual data matrix and $\mathbf{Z}_{N\times p} = [\mathbf{z}_{ij}]$, be the corresponding data matrix of the standardized variables $z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$

- The covariance matrix of the standardized data $\mathbf{Z}$ is

$$\mathbf{R} = cov(\mathbf{Z}) = \frac{1}{N-1}\mathbf{Z}^T\mathbf{Z},$$

which corresponds to the correlation matrix of $\mathbf{X}$

- The PCs are now given by $\mathbf{Y}_j = \mathbf{Z}\mathbf{v}_j$ where $\mathbf{v_1}, \dots, \mathbf{v_p}$ are unit and pairwise orthogonal eigenvectors of $\mathbf{R}$ associated with eigenvalues $\lambda_1 \geq \dots \geq \lambda_p > 0$
- The total variance is now the number of variables:

$$p = \sum_{i=1}^{p} var(\mathbf{z}_j) = \lambda_1 + \dots + \lambda_p$$

- The correlation coefficient between $\mathbf{z}_j$ and $\mathbf{y}_k$ reduces to

$$cor(\mathbf{z}_j, \mathbf{y}_k) = \sqrt{\lambda_k}\mathbf{v}_{kj}$$

Interpretation of the results in space of variables

Each standardized variable $\mathbf{z}_j$ and each PC $\mathbf{y}_k$, can be represented as vectors in $\mathbb{R}^N$. This allows to reinterpret geometrically some of the previous statistics:

- The variables $\mathbf{z}_j$, $j = 1, \ldots, p$, lie in a hypersphere of radius $\sqrt{N-1}$:

  $$\|\mathbf{z}_j\|^2 = \mathbf{z}_j^T \mathbf{z}_j = (N-1)var(\mathbf{z}_j) = N-1$$

  More generally, the length of centered variable is also proportional to its standard deviation (exercise)
- The length of each PC is proportional to its variance:

  $$\begin{aligned} \|\mathbf{y}_k\|^2 &= \mathbf{y}_k^T \mathbf{y}_k = (\mathbf{Z}\mathbf{v}_k)^T(\mathbf{Z}\mathbf{v}_k) \\ &= \mathbf{v}_k^T \mathbf{Z}^T \mathbf{Z} \mathbf{v}_k = (N-1)\mathbf{v}_k^T \mathbf{R} \mathbf{v}_k \\ &= (N-1)\lambda_k = (N-1)var(\mathbf{y}_k) \end{aligned}$$

- The correlation coefficient between $\mathbf{z}_j$ and $\mathbf{y}_k$ is the cosine of the angle θ_{jk} between the variables $\mathbf{z}_j$ and $\mathbf{y}_k$:

  $$\begin{aligned} cor(\mathbf{z}_j, \mathbf{y}_k) &= \frac{cov(\mathbf{z}_j, \mathbf{y}_k)}{\sqrt{var(\mathbf{z}_j)}\sqrt{var(\mathbf{y}_k)}} = \frac{\frac{\mathbf{z}_j^T \mathbf{y}_k}{N-1}}{\frac{\|\mathbf{z}_k\|}{\sqrt{N-1}}\sqrt{\lambda_k}} \\ &= \frac{\mathbf{z}_j^T \mathbf{y}_k}{\|\mathbf{z}_k\|(\sqrt{N-1}\sqrt{\lambda_k})} = \frac{\mathbf{z}_j^T \mathbf{y}_k}{\|\mathbf{z}_j\|\,\|\mathbf{y}_k\|} = \cos(\theta_{jk}) \end{aligned}$$

- The correlation coefficient between $\mathbf{z}_j$ and $\mathbf{z}_k$ is the cosine of the angle between the vectors representing these variables (exercise)

How many PCs ?

No exact answer can be given. Some empirical rules are listed below:

- To define a cutoff %: to consider a given cumulative percentage of the total variation (usually between 70% and 90%) and to choose the smallest number m of PC such that the % of explained variance by the first m PCs exceeds the chosen %.
- Scree plot: to look for a elbow point in the scree plot of the variance
- Kaiser's rule (for PCA on correlation matrix): to retain the PCs with variance greater than the average value 1: the PCs with variance inferior to 1 contain less information than the original variables and are not worthing to retain. (for the PCA on the covariance matrix, the cutoff value 1 should be replaced by the average of the PCs variances)
- Jolliffe's variant of Kaiser's rule (for PCA on correlation matrix): is a more conservative rule that proposes a cutoff value of 0.7
- Broken-stick model: a unit stick is randomly broken into p segments. The expected length of the k-th largest segment is $\ell_k^* = \frac{1}{p}\sum_{j=k}^{p}\frac{1}{j}$. This rule retains the PCs while the variance of each PC_k keeps above the length ℓ_k

PCA on the correlation matrix - summary

- All variables have the same variance and therefore their importance is equalized
- The cloud of individuals tend to have a more rounded shape
- The PCA tends to reflect existing correlation patterns among variables
- The first PC tends to be dominated by groups of variables that are highly correlated. Actually, by (6) in slide 30 applied to the correlation matrix $\mathbf{Z}$ we deduce that the first PC is $\mathbf{Za}$ such that $\mathbf{a} = (\alpha_1, \ldots, \alpha_p)$ maximizes

$$var(\mathbf{Za}) = \sum_{i=1}^{p} \alpha_i^2 + 2\sum_{i<j} \alpha_i \alpha_j cor(\mathbf{z}_i, \mathbf{z}_j), \quad \text{with} \quad \|\mathbf{a}\| = \sum_{i=1}^{p} \alpha_i^2 = 1$$

- The PCs can be interpreted since they are linear combinations of dimensionless variables
- The number of PCs that are necessary to explain a given proportion of the dataset total variability is usually higher compared to the PCA on the covariance matrix

A more geometrical approach to PCA using SVD

Applying the SVD to the centered data matrix $\mathbf{X}^*$ we obtain

$$\boxed{\mathbf{X}^* = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}^T = \textstyle\sum_{j=1}^{r} \delta_j \mathbf{u}_j \mathbf{v}_j^T}$$

where

- $\boldsymbol{\Delta}_{r\times r} = \text{diag}(\delta_1, \ldots, \delta_r)$ is the diagonal matrix containing the (positive) singular values of $\mathbf{Z}$ with $\delta_1 \geq \delta_2 \geq \cdots \geq \delta_r > 0$
- $\mathbf{U}_{N\times r} = [\, \mathbf{u}_1 \ \cdots \ \mathbf{u}_r]$, with $\mathbf{u}_1, \ldots, \mathbf{u}_r \in \mathbb{R}^N$, is the matrix of left singular vectors of $\mathbf{Z}$
- $\mathbf{V}_{p\times r} = [\, \mathbf{v}_1 \ \cdots \ \mathbf{v}_r]$, with $\mathbf{v}_1, \ldots, \mathbf{v}_r \in \mathbb{R}^p$, is the matrix of right singular vectors of $\mathbf{Z}$
- $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}_r$, that is, the left and right singular vectors, are unit and pairwise orthogonal vectors vectors

For each $k = 1, \ldots, r$, we have a rank k linear approximation of $\mathbf{X}^*$,

$$\mathbf{X}(k) = \sum_{j=1}^{k} \delta_j \mathbf{u}_j \mathbf{v}_j^T = \mathbf{U}(k)\boldsymbol{\Delta}(k)\mathbf{V}(k)^T$$

Here $\mathbf{U}(k)$ and $\mathbf{V}(k)$ are the submatrices containing the first k columns of $\mathbf{U}$ and $\mathbf{V}$, respectively, and $\boldsymbol{\Delta}(k)$ is the diagonal matrix containing the first k singular values

Best rank k-linear approximation

For instance, we have the following rank one and rank two linear approximations,

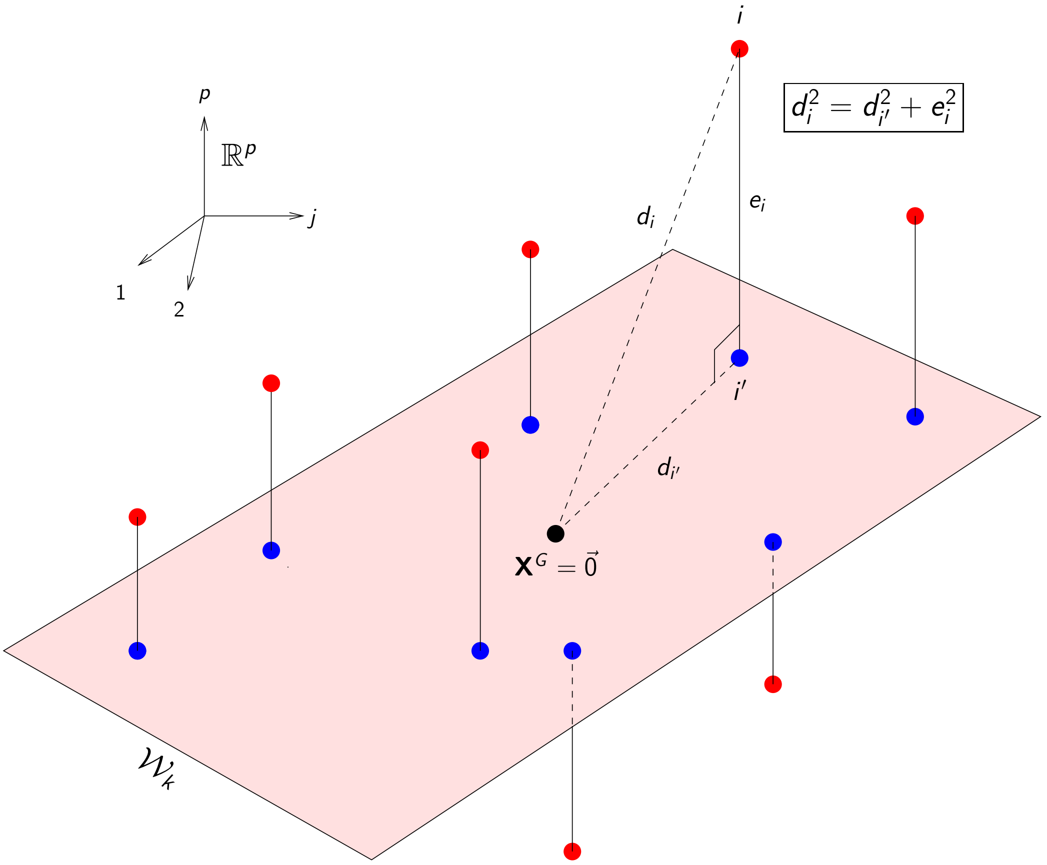
$$\begin{aligned}\mathbf{X}(1) &= \delta_1 \mathbf{u}_1 \mathbf{v}_1^T = \mathbf{U}(1)\mathbf{\Delta}(1)\mathbf{V}(1)^T \\ \mathbf{X}(2) &= \delta_1 \mathbf{u}_1 \mathbf{v}_1^T + \delta_2 \mathbf{u}_1 \mathbf{v}_2^T = \mathbf{U}(2)\mathbf{\Delta}(2)\mathbf{V}(2)^T\end{aligned}$$

All rows of $\mathbf{X}(k)$ are linear combinations of $\mathbf{v}_1^T, \ldots, \mathbf{v}_k^T$. Moreover:

- For each k, the cloud of N points defined by the rows of $\mathbf{X}(k)$ lie in a k-dimension linear subspace $\mathcal{W}(k)$ of $\mathbb{R}^p$ (generated by the vectors $\mathbf{v}_1 \ldots, \mathbf{v}_k$), that is close to the cloud of centered points defined by the rows of $\mathbf{X}^*$
- Denoting by i the point defined by row i of $\mathbf{X}^*$ (a red dot in next slide) and by i' the corresponding k-approximated point in $\mathcal{W}(k)$ (corresponding projected blue dot), which is defined by the row i of $\mathbf{X}(k)$, we have that $i - i'$ is a linear combination of $\mathbf{v}_j$, $j > k$, and thus orthogonal to the linear space $\mathcal{W}(k)$
- Denoting by d_i the distance between i and the origin (center of gravity), by $d_{i'}$ the distance between i' and the origin and setting $e_i = d(i, i')$, we have a decomposition

$$d_i^2 = d_{i'}^2 + e_i^2 \qquad (12)$$

Best fitting k-dimensional linear space



Best k-dimensional fitting

- The cloud of (blue) points $\mathbf{X}(k)$ gives the best rank k approximation of $\mathbf{X}^*$, corresponding to the best fitting k-dimensional linear space in terms of least square distances, between the centered cloud of points defined by $\mathbf{X}^*$ and the cloud of the projected points in the k-dimensional space, $\mathbf{X}(k)$. In other words it minimizes the sum of square distances $\sum_i e_i^2$ (Eckart-Young's Theorem)
- Using the decomposition (12) of the slide 52 we obtain,

$$\begin{aligned}\underbrace{var(\mathbf{X}^*)}_{\text{total var.}} &= \frac{1}{N-1}\sum_i d_i^2 = \frac{1}{N-1}\sum_{i'} d_{i'}^2 + \frac{1}{N-1}\sum_i e_i^2\\ &= \underbrace{var(\mathbf{X}(k))}_{\text{explain. var.}} + \underbrace{\frac{1}{N-1}\sum_i e_i^2}_{\text{unexplain. var.}}\end{aligned}$$

Therefore the optimal solution in the sense of the least square distances, minimizes the variance that is left unexplained, i.e., maximizes the variance of the cloud of N points projected in a k-dimensional space (explained variance) - main goal of PCA!

Equivalence between the EVD and SVD approaches

We shall assume all singular values positive (otherwise we have to work with a slight different version of the SVD decomposition):

$$(\mathbf{X}^*)^T\mathbf{X}^* = (\mathbf{U\Delta V}^T)^T(\mathbf{U\Delta V}^T) = \mathbf{V\Delta}^T\mathbf{U}^T\mathbf{U\Delta V}^T = \mathbf{V\Delta}^2\mathbf{V}^T,$$

which is equivalent to say that,

$$\mathbf{S} = \mathbf{V}\left(\frac{1}{\sqrt{N-1}}\mathbf{\Delta}\right)^2\mathbf{V}^T \tag{13}$$

Hence the PC loadings, i.e., the eigenvectors of $\mathbf{S}$, are the right singular vectors of $\mathbf{X}^*$ and the corresponding PC standard deviations, the singular values of $\mathbf{X}^*$ divided by $\sqrt{N-1}$. The PC factor scores are given by

$$\mathbf{Y}^* = \mathbf{X}^*\mathbf{V} = \mathbf{U\Delta V}^T\mathbf{V} = \mathbf{U\Delta},$$

and the left singular vectors verify

$$\mathbf{U} = \mathbf{X}^*\mathbf{V\Delta}^{-1} = \mathbf{Y}^*\mathbf{\Delta}^{-1},$$

where $\mathbf{Y}^*\mathbf{\Delta}^{-1}$ is a matrix of normalized scores (more precisely, with constant standard deviations $\frac{1}{\sqrt{N-1}}$)

One can consider, alternatively, the SVD of $\frac{1}{\sqrt{N-1}}(\mathbf{X}^*)^T\mathbf{X}^*$. In this case the PCs variances λ_j are the squared singular values δ_j^2 of $\frac{1}{\sqrt{N-1}}(\mathbf{X}^*)^T\mathbf{X}^*$ (see the slides of Prof. Cadima)

comparing PCA via EVD and via SVD in R

R

```
# EVD APPROACH TO PCA
X<-iris[-5] # can be replaced by your own dataset or standardized
X.pca <- prcomp(X) # computes the PCA of X
loadings <- X.pca$rotation # eigenvectors of S=cov(X) (loadings)

sdev <- X.pca$sdev
# standard deviations of the PCs (square roots of the eigenvalues of S)

scores <- X.pca$x # scores (coordinates of the individuals w.r.t PCs)

# SVD APPROACH TO PCA
Xc <- scale(X,scale=FALSE) # Xc = centered X ou t(t(X)-colMeans(X))
X.svd<-svd(Xc) # computes the SVD of Xc
left.sing <- X.svd$u # left singular vectors of Xc
singvalues <- X.svd$d # singular values of Xc
right.sing <- X.svd$v # right singular vectors of Xc

# EQUIVALENCE BETWEEN EVD AND SVD APPROACHES

sdev; singvalues/sqrt(N-1)
# eigenvalues of S = square of sing values of Xc (divided by N-1)

head(loadings); head(right.sing) # loadings = right sing vectors

head(scores) ; head(left.sing%*%diag(singvalues))
# scores = normalized left sing vectors
```

A very useful decomposition. . .

Any matrix $\mathbf{C}_{N\times p}$ of rank r can be decomposed as a

$$\mathbf{C} = \mathbf{A}\,\mathbf{B}^T = \sum_{i=1}^{r} \mathbf{a}_i \mathbf{b}_i^T ,$$

where $\mathbf{A} = [\,\mathbf{a}_1 \;\cdots\; \mathbf{a}_r\,]$ and $\mathbf{B} = [\,\mathbf{b}_1 \;\cdots\; \mathbf{b}_r\,]$, with $\mathbf{a}_i \in \mathbf{R}^N$ and $\mathbf{b}_i \in \mathbf{R}^p$

In particular, any matrix $\mathbf{C}$ of rank one, i.e., with proportional rows and proportional columns, can be decomposed as:

$$\mathbf{C} = \mathbf{a}\,\mathbf{b}^T = \begin{bmatrix} a_1 \\ \vdots \\ a_N \end{bmatrix} \begin{bmatrix} b_1 & \cdots & b_p \end{bmatrix}, \quad \text{with}$$

$$\mathbf{a} = \begin{bmatrix} a_1 \\ \vdots \\ a_N \end{bmatrix} = (a_1, \ldots, a_N) \in \mathbb{R}^N, \quad \mathbf{b} = \begin{bmatrix} b_1 \\ \vdots \\ b_p \end{bmatrix} = (b_1, \ldots, b_N) \in \mathbb{R}^p$$

The decomposition is not unique. For instance,

$$\mathbf{C} = \begin{bmatrix} 2 & 4 & 6 \\ 4 & 8 & 12 \end{bmatrix} = \begin{bmatrix} 2 \\ 4 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \begin{bmatrix} 2 & 4 & 6 \end{bmatrix}$$

In the general case the decomposition can be obtained using the SVD. . .

Biplots

The biplots provide simultaneous representations of the individuals and variables of a data matrix in a low dimension space (usually of dimension two or three), using the SVD applied to the centered data matrix in order to obtain a decomposition of the type described in the previous slide

Let $\mathbf{X}^*$ be the matrix obtained by centering the p observed variables of a data matrix $\mathbf{X}_{N\times p}$ (i.e., column centering the matrix). We will assume $\mathbf{X}^*$ has rank p. Applying the SVD we can write,

$$\mathbf{X}^* = \mathbf{U\Delta V}^T \tag{14}$$

where,

- $\mathbf{U}_{N\times p}$ verifies $\mathbf{U}^T\mathbf{U} = \mathbf{I}_p$ is the matrix of left singular vectors of $\mathbf{X}^*$
- $\mathbf{V}_{p\times p}$ verifies $\mathbf{V}^T\mathbf{V} = \mathbf{I}_p$ is the matrix of right singular vectors of $\mathbf{X}^*$, i.e., the matrix of loadings of $\mathbf{X}$
- $\mathbf{\Delta}_{p\times p} = \text{diag}(\delta_1, \ldots, \delta_p)$ is a diagonal matrix containing the singular values of $\mathbf{X}^*$

Biplots (cont.)

Using the decomposition (14) of the previous slide we can decompose $\mathbf{X}^* = \mathbf{GH}^T$ in many different ways. We will refer here two of them:

- $\mathbf{G} = \mathbf{U\Delta}$ and $\mathbf{H} = \mathbf{V}$ - focuses on distances between individuals
- $\mathbf{G} = \mathbf{U}$ and $\mathbf{H} = \mathbf{V\Delta}$ - focuses on covariances/correlations between variables

In the first case, $\mathbf{G} = \mathbf{U\Delta}$ contains the left singular vectors scaled by the respective singular values which gives the factor scores (coordinates) of the individuals. Actually, the right singular vectors of $\mathbf{X}^*$ are eigenvectors of the covariance matrix $\mathbf{S}$, i.e, vectors of loadings of $\mathbf{X}$ and therefore the scores matrix is given

$$\mathbf{Y}^* = \mathbf{X}^*\mathbf{V} = \mathbf{U\Delta V}^T\mathbf{V} = \mathbf{U\Delta}$$

The matrix $\mathbf{H} = \mathbf{V}$,corresponds to the matrix of right singular vectors, i.e. to the matrix of the vectors of loadings

Biplots (cont.)

Consider now the second case, $\mathbf{G}_{N\times p} = \mathbf{U}$ and $\mathbf{H}_{p\times p} = \mathbf{V\Delta}$ and denote

$$\mathbf{G}^T = [\, \mathbf{g}^1 \ \cdots \ \mathbf{g}^N \,],$$

where $\mathbf{g}^j \in \mathbb{R}^p$ is the j-th row of $\mathbf{G}$. Similarly denote

$$\mathbf{H}^T = [\, \mathbf{h}^1 \ \cdots \ \mathbf{h}^p \,],$$

where $\mathbf{h}^k \in \mathbb{R}^p$ is the k-th row of $\mathbf{H}$ The rows of G and H are called, respectively, **markers of individuals and variables**. We have,

$$\begin{aligned} (N-1)\,\mathbf{S} &= (\mathbf{X}^*)^T\mathbf{X}^* = (\mathbf{G}\mathbf{H}^T)^T\mathbf{G}\mathbf{H}^T \\ &= \mathbf{H}\mathbf{G}^T\mathbf{G}\mathbf{H}^T = \mathbf{H}\mathbf{U}^T\mathbf{U}\mathbf{H}^T = \mathbf{H}\mathbf{H}^T \end{aligned}$$

Hence

$$(\mathbf{h}^j)^T\mathbf{h}^k = (N-1)s^2_{jk},$$

that is, the inner product between the markers $\mathbf{h}^j$ and $\mathbf{h}^k$ is proportional to the covariance between the observed variables $\mathbf{x}_j$ and $\mathbf{x}_k$. In particular, the length of each variable marker is proportional to the standard deviation of the corresponding variable and we get, denoting θ_{jk} the angle between the variable markers $\mathbf{h}^j$ and $\mathbf{h}^k$,

$$\cos(\theta_{jk}) = r_{jk}$$

Euclidean and Mahalanobis distances

The usual squared (euclidean) distance between the individuals $\mathbf{x}^i, \mathbf{x}^\ell \in \mathbf{R}^p$ is

$$d^2_{i\ell} = \|\mathbf{x}^i - \mathbf{x}^\ell\|^2 = (\mathbf{x}^i - \mathbf{x}^\ell)^T(\mathbf{x}^i - \mathbf{x}^\ell)$$

The (squared) Mahalanobis distance accounts for the dataset variability and generalizes the euclidean distance. Assuming the covariance matrix $\mathbf{S}$ invertible, the Mahalanobis distance between the individuals $\mathbf{x}^i, \mathbf{x}^\ell \in \mathbf{R}^p$ is defined as

$$\delta^2_{i\ell} = (\mathbf{x}^i - \mathbf{x}^\ell)^T\mathbf{S}^{-1}(\mathbf{x}^i - \mathbf{x}^\ell)$$

The Mahalanobis distance between the individuals $\mathbf{x}^i = \mathbf{H}\mathbf{g}^i$ and $\mathbf{x}^\ell = \mathbf{H}\mathbf{g}^\ell$ is proportional to the (squared) euclidean distance between the corresponding markers $\mathbf{g}^i$ and $\mathbf{g}^\ell$. Actually, from relation (13) of slide 56, we obtain

$$(N-1)\,\mathbf{V}\mathbf{\Delta}^{-2}\mathbf{V}^T = (N-1)\,((\mathbf{X}^*)^T\mathbf{X}^*))^{-1} = \mathbf{S}^{-1}$$

and therefore

$$\begin{aligned} (N-1)(\mathbf{g}^i - \mathbf{g}^\ell)^T(\mathbf{g}^i - \mathbf{g}^\ell) &= (N-1)(\mathbf{g}^i - \mathbf{g}^\ell)^T\mathbf{\Delta}\mathbf{\Delta}^{-2}\mathbf{\Delta}(\mathbf{g}^i - \mathbf{g}^\ell) \\ &= (N-1)(\mathbf{g}^i - \mathbf{g}^\ell)^T\mathbf{\Delta}(\mathbf{V}^T\mathbf{V})\mathbf{\Delta}^{-2}(\mathbf{V}^T\mathbf{V})\mathbf{\Delta}(\mathbf{g}^i - \mathbf{g}^\ell) \\ &= (\mathbf{g}^i - \mathbf{g}^\ell)^T(\mathbf{V\Delta})^T\mathbf{S}^{-1}(\mathbf{V\Delta})(\mathbf{g}^i - \mathbf{g}^\ell) \\ &= (\mathbf{g}^i - \mathbf{g}^\ell)^T\mathbf{H}^T\mathbf{S}^{-1}\mathbf{H}(\mathbf{g}^i - \mathbf{g}^\ell) \\ &= (\mathbf{H}(\mathbf{g}^i - \mathbf{g}^\ell))^T\mathbf{S}^{-1}\mathbf{H}(\mathbf{g}^i - \mathbf{g}^\ell) \\ &= (\mathbf{x}^i - \mathbf{x}^\ell)^T\mathbf{S}^{-1}(\mathbf{x}^i - \mathbf{x}^\ell) = \delta^2_{i\ell}, \quad \text{(UFF!)} \end{aligned}$$

“Exact” interpretation of a biplot

Summarizing, we have the following “exact interpretations”:

- The cosine of the angle between two variable markers is the correlation coefficient between these variables
- The length of a variable marker is proportional to the standard deviation of the variable
- The euclidean distance between individual markers is proportional to the Mahalanobis distance between the corresponding individuals
- The coordinate of the orthogonal projection of an individual marker $\mathbf{g}^i$ onto the line defined by a variable marker $\mathbf{h}^j$ equals value of the individual on that variable divided by the standard deviation of the variable

The last property follows directly from relation $\mathbf{X}^* = \mathbf{G}\mathbf{H}^T$, which implies that $x^*_{ij} = (\mathbf{g}^i)^T\mathbf{h}^j$ and therefore,

$$\mathrm{proj}_{\mathbf{h}^j}(\mathbf{g}^i) = \frac{(\mathbf{g}^i)^T\mathbf{h}^j}{\|\mathbf{h}^j\|^2}\mathbf{h}^j = \frac{x^*_{ij}}{\|\mathbf{h}^j\|^2}\mathbf{h}^j$$

Note that $\|\mathrm{proj}_{\mathbf{h}^j}(\mathbf{g}^i)\| = \frac{|x^*_{ij}|}{\|\mathbf{h}^j\|}$

“Approximated interpretations” of a biplot

Let $\mathbf{G}^T(m) = \mathbf{U}(m)^T$ and $\mathbf{H}^T(m) = \mathbf{\Delta}(m)\mathbf{V}(m)^T$, $1 \leq m \leq p$ be the submatrices containing the first m rows of $\mathbf{G}^T$ and $\mathbf{H}^T$, resp. Denote

$$(\mathbf{G}(m))^T = [\,\mathbf{g}^1_m \ \cdots \ \mathbf{g}^N_m\,], \qquad (\mathbf{H}(m))^T = [\,\mathbf{h}^1_m \ \cdots \ \mathbf{h}^p_m\,]$$

The rows of $G(m)$ and $H(m)$ give approximations to the markers of the individuals and variables. We have:

- The cosines of the angles between variable markers are approximately equal to the correlation coefficients between these variables
- The length of a variable marker is approximately proportional to the standard deviation of the variable
- The (euclidean) distances between individual markers are approximately proportional to the Mahalanobis distance between these individuals
- The coordinate of the orthogonal projection of an individual marker $\mathbf{g}^i$ onto the line defined by a variable marker $\mathbf{h}^j$ is approximately equal to the value of the individual on that variable divided by the standard deviation of the variable

The higher the proportion of the explained variance by the first m PCs, the better the approximations in the previous points

Displaying a biplot using in R software

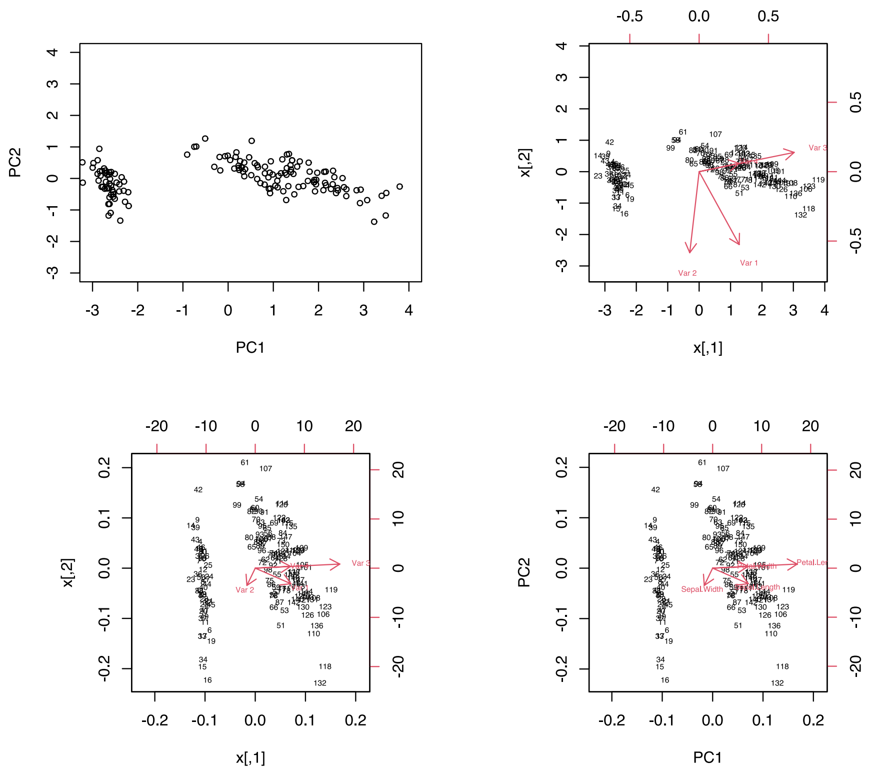
We can display the biplot of the iris flowers data set in two distinct ways, with the `biplot` function:

R

```
Xc <- scale(iris[-5],scale=FALSE) #centred iris flower dataset
iris.svd <- svd(Xc) # compute the svd UDVT̂ of the centred iris dataset
U <- iris.svd$u
V <- iris.svd$v
Delta <- diag(iris.svd$d) # creates a diagonal matrix with diagonal with
the singular values
par(mfrow=c(2,2)) # 4 simultaneous windows
plot(iris.pca$x[,1:2],asp=TRUE,pch=16) # plot
biplot(U % * % Delta, V, asp=TRUE,cex=.5) # G=U Delta; H=V
biplot(U, V% * %Delta, asp=TRUE,cex=.5) # G=U; H=V Delta
biplot(iris.acp, asp=TRUE,cex=.5) # computes the second species
```

Iris flower biplots

The output obtained by the script of the previous slide



Some notes on generalized euclidean distances

If $\mathbf{S}$ is a symmetric positive definite (hence invertible) matrix of order p, we define the (squared) **generalized euclidean distance** between the vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$ as

$$d^2_{\mathbf{S}}(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})^T \mathbf{S}^{-1} (\mathbf{x} - \mathbf{y})$$

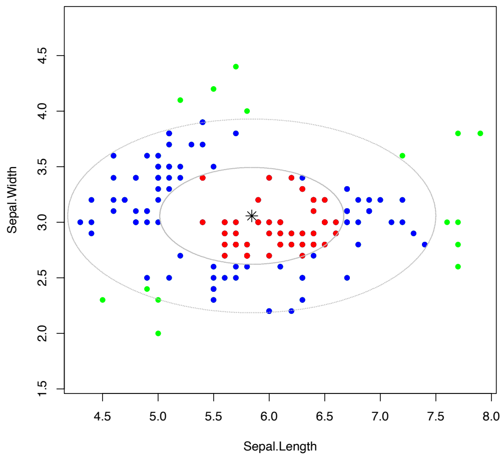
- If $\mathbf{S} = \mathbf{I}_p$, $d^2_{\mathbf{S}}(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})^T (\mathbf{x} - \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|^2$ is the usual (squared) Euclidean distance between $\mathbf{x}$ and $\mathbf{y}$
- If $\mathbf{S} = cov(\mathbf{X})$, $d^2_{\mathbf{S}}(\mathbf{x}, \mathbf{y})$ is the (squared) Mahalanobis distance between $\mathbf{x}$ and $\mathbf{y}$
- When the variables are uncorrelated, the covariance matrix $\mathbf{S}$ is a diagonal matrix containing the variances of the p variables and $d^2_{\mathbf{S}}(\mathbf{x}, \mathbf{y})$ equals the (squared) euclidean distance between the corresponding standardized variables
- The Mahalanobis distance of between an individual and the cloud's center of gravity is 'smaller' along the directions of $\mathbf{X}$ of greater variability and generalizes to the multivariate case the idea of how many standard deviations each observed vector $\mathbf{x}$ is far away from the mean. This can be useful, for instance, to detect outliers. . .

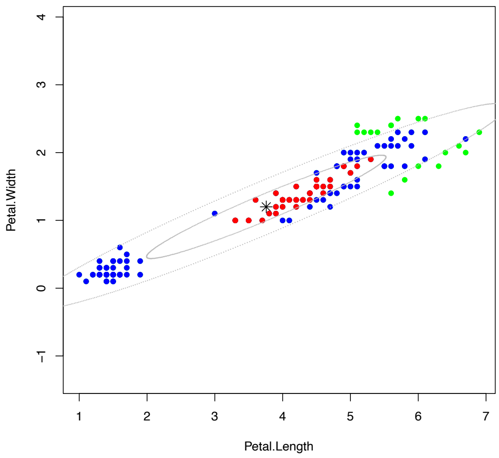
Mahalanobis distances for the iris flower data set

The variance-covariance matrices of the sepal and the petal widths are, respectively:

$$\begin{bmatrix} 0.6856935 & -0.0424340 \\ -0.0424340 & 0.1899794 \end{bmatrix}, \qquad \begin{bmatrix} 3.116278 & 1.2956094 \\ 1.295609 & 0.5810063 \end{bmatrix}$$

The iris flowers at Mahalanobis distances from the mean less than or equal to 1 are displayed in red and the iris flowers at mahalanobis distances greater than 1 and smaller than or equal to 2 displayed in blue color





Contribution and square cosine

- Recall that the contribution of individual i to a PC$_k$ is the part of the variance of PC$_k$ that is due to individual i (in %):

$$ctr_{i,k} = \frac{(y^*_{i,k})^2}{\sum_{j=1}^{N}(y^*_{j,k})^2} \times 100$$

Individuals with contributions above the average are usually more important to interpret the PC

- A related notion is the square cosine of a PC k with an individual i, which gives the contribution of the PC to the squared distance of the individual to the origin:

$$\cos^2_{i,k} = \frac{(y^*_{i,k})^2}{\sum_{j=1}^{p}(y^*_{i,j})^2}$$

Square cosines can be added together to assess the quality of representation of an individual i by its projection on the space defined by several PCs. For instance, the quality of representation of individual i in the PFP is given by,

$$\cos^2_{i,1} + \cos^2_{i,2} = \frac{(y^*_{i,1})^2 + (y^*_{i,2})^2}{\sum_{j=1}^{p}(y^*_{i,j})^2}$$

Only well represented individuals should be interpreted!

PCA interpretation - summary

- Proportion of the variance explained by a PC
- Correlation between a variable and a PC
- Contribution of an individual to a PC
- Square cosine of a PC with an individual
- Biplot