Mathematical Models and Applications

Multivariate Analysis

Pedro Cristiano Silva

Instituto Superior de Agronomia

2025/26

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 $\mathbf{A}$, $\mathbf{B}$, $\mathbf{X}$, $\mathbf{Y}$,...

Some background in LINEAR ALGEBRA

Matrix multiplications revisited

If $\mathbf{A}_{m\times n} = \begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_n \\ | & | & & | \end{bmatrix}$ with $\mathbf{a}_j \in \mathbb{R}^n$, $j = 1, \ldots, n$, and

$\mathbf{B}_{n\times p} = \begin{bmatrix} -\mathbf{b}_1^T- \\ -\mathbf{b}_2^T- \\ \vdots \\ -\mathbf{b}_n^T- \end{bmatrix}$ with $\mathbf{b}_j \in \mathbb{R}^p$, then $AB = \sum_{j=1}^n \mathbf{a}_j \mathbf{b}_j^T$

Note that if $\mathbf{b} = (\beta_1, \ldots, \beta_n)$ one gets, $\mathbf{Ab} = \mathbf{A}\begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} = \sum_{j=1}^n \beta_j \mathbf{a}_j$

Example

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

Linear combinations and matrix products

Let $\mathbf{X} = [\,\mathbf{x}_1 \ \cdots \ \mathbf{x}_p\,]$ be matrix of type $N \times p$.

A linear combination of the p columns $\mathbf{x}_1, \ldots, \mathbf{x}_p$ of X (considered as p vectors of $\mathbb{R}^N$) is an expression of 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 (also called loadings in the PCA context).

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 $\mathbf{A}$ if there is $\lambda \in \mathbb{R}$ such that

$$\mathbf{A}\mathbf{v} = \lambda \mathbf{v}.$$

λ is called the corresponding eigenvalue.

Example

If we consider $\mathbf{A} = \begin{bmatrix} 3 & 0 & 2 \\ 0 & -1 & 1 \\ 2 & 0 & 0 \end{bmatrix}$ and $\mathbf{v} = \begin{bmatrix} 10 \\ 1 \\ 5 \end{bmatrix}$, we obtain

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

Hence $\mathbf{v}$ is an eigenvector of $\mathbf{A}$ associated with 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 $\mathbf{A}$, denoted $\det(A)$ (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 code

```
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

- *We say that a vector* $v \in \mathbb{R}^p$ *is unitary if* v *has length one, that is,* $\|v\|^2 = v^T v = 1$.
- *We say that* u *and* v *belonging to* $\mathbb{R}^p$ *are orthogonal if* $u^T v = 0$.
- *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* $v_1, \ldots, v_q$ *are unitary and pairwise orthogonal vectors, that is,*

$$v_i^T v_i = 1 \quad \textit{and} \quad v_i^T v_j = 0, \quad \forall i, j, \ i \neq j.$$

Denoting $\mathbf{V}_{p \times q} = [\, \mathbf{v}_1 \ \cdots \ \mathbf{v}_q \,]$ 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 if and only if $\mathbf{V}^T \mathbf{V} = \mathbf{I}_q$.

If moreover $q = p$, we have $\mathbf{V}^T \mathbf{V} = \mathbf{V} \mathbf{V}^T = \mathbf{I}_p$, that is, $\mathbf{V}^{-1} = \mathbf{V}^T$.

Vector decomposition in an orthonormal basis ($q = p$)

If $u \in \mathbb{R}^p$ is unitary and $\{\mathbf{v}_1, \ldots, \mathbf{v}_p\}$ is an orthonormal set of $\mathbb{R}^p$ with p vectors we have the decomposition,

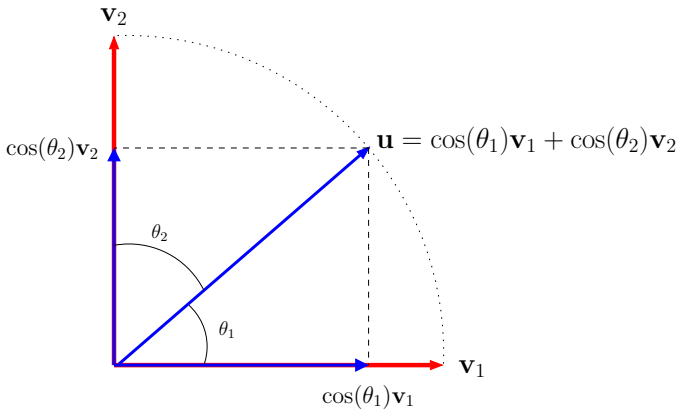
$$\mathbf{u} = (\mathbf{u}^T\mathbf{v}_1)\mathbf{v}_1 + \cdots + (\mathbf{u}^T\mathbf{v}_p)\mathbf{v}_p = \cos(\theta_1)\mathbf{v}_1 + \cdots + \cos(\theta_p)\mathbf{v}_p,$$

with

$$\cos^2(\theta_1) + \cdots + \cos^2(\theta_p) = u^T u = 1,$$

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

The case $p = 2$:



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

$$\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T \tag{1}$$

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 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 4, we can rewrite (1) as,

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

which is called the spectral decomposition of $\mathbf{A}$. Applying this theorem we can prove a decomposition for arbitrary matrices, called singular value decomposition (see next slide), which constitutes one of the most important and useful results in linear algebra.

Singular value decomposition (SVD) of an arbitrary matrix

Theorem (compact SVD)

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

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

where:

- $\mathbf{U} = [\,\mathbf{u}_1 \;\cdots\; \mathbf{u}_r\,]$ *verifies* $\mathbf{U}^T\mathbf{U} = \mathbf{I}_r$*, i.e., the columns of* U *are unit and pairwise orthogonal vectors of* $\mathbb{R}^N$*, which are called left singular vectors of* $\mathbf{A}$*.*
- $\mathbf{V} = [\,\mathbf{v}_1 \;\cdots\; \mathbf{v}_r\,]$ *verifies* $\mathbf{V}^T\mathbf{V} = \mathbf{I}_r$*, i.e., the columns of* V *are unit and pairwise orthogonal vectors of* $\mathbb{R}^p$*, which are called right singular vectors of* $\mathbf{A}$*.*
- $\mathbf{\Delta} = diag(\delta_1, \ldots, \delta_r)$ *with nonzero diagonal elements* $\delta_1 \geq \cdots \geq \delta_r > 0$*, which are called singular values of* $\mathbf{A}$ *and verify*

$$\mathbf{A}\mathbf{v}_i = \delta_i \mathbf{u}_i, \qquad \mathbf{A}^T\mathbf{u}_i = \delta_i \mathbf{v}_i.$$

Using the results of slide 4 we can rewrite (2) 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

(As an exploratory analysis tool)

Summary statistics - univariate case

Given vectors $\mathbf{x} = (x_1, \ldots, x_N)$ and $\mathbf{y} = (y_1, \ldots, y_N) \in \mathbb{R}^N$ containing N observations of two quantitative variables, 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 $\mathbf{y}$:

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

Variable's cloud and individual's cloud

Consider a data matrix $\mathbf{X}_{N \times p} = [\, x_{ij} \,]$ with $x_{ij} \in \mathbb{R}$. We assume that:

- each column $\mathbf{x}_j = (x_{1j}, \ldots, x_{Nj})$ of $\mathbf{X}$ represents the observations of a quantitative variable across N individuals and we write,

$$\mathbf{X}_{N \times p} = [\, \mathbf{x}_1 \quad \cdots \quad \mathbf{x}_p \,].$$

  The p columns of $\mathbf{X}$ define a cloud of p vectors in $\mathbb{R}^N$, which we call the variable's cloud.

- each row $\mathbf{x}^i = (x_{i1}, \ldots, x_{ip})$ of $\mathbf{X}$, i.e., each column of $(\mathbf{X})^T$, represents the observations of p variables for a single individual and we write,

$$\mathbf{X}^T_{p \times N} = [\, \mathbf{x}^1 \quad \cdots \quad \mathbf{x}^N \,].$$

  The N rows of $\mathbf{X}$ define a cloud of N points in $\mathbb{R}^p$, which we call the individuals's cloud.

Summary statistics - multivariate case

- The (sample) mean vector of the dataset $\mathbf{X}$, i.e., the cloud's center of gravity, is

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

that is, $\mathbf{x}^G = (\bar{x}_1, \ldots, \bar{x}_p)$ with $\bar{x}_j = \frac{1}{N}\sum_{i=1}^{N} x_{ij}$.

- The (sample) covariance matrix of data matrix $\mathbf{X}_{N\times p}$ is

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

where the (sample) covariance between variables j and k is equal to

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

- The total variability of $\mathbf{X}$ is

$$\text{tr}(\mathbf{S}) = s^2_{11} + \cdots + s^2_{pp} = \frac{1}{N-1}\sum_{j=1}^{p}\sum_{i=1}^{N}(x_{ij} - \bar{x}_j)^2 = \frac{1}{N-1}\sum_{i=1}^{N}\|\mathbf{x}^i - \mathbf{x}^G\|^2.$$

Centering the data matrix

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

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

Similarly to slide 15 the (sample) covariance s^2_{jk} between variables j and k can be written in a very convenient way using the centred vectors, $\mathbf{x}^*_j$ and $\mathbf{x}^*_k$, as a simple inner product (in $\mathbb{R}^N$) divided by $N-1$:

$$s^2_{jk} = \frac{1}{N-1}(\mathbf{x}^*_j)^T\mathbf{x}^*_k. \tag{3}$$

The (column) centred data matrix, hereafter denoted $\mathbf{X}^*$, is obtained gathering together the p centred vectors,

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

which amounts to say that,

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

Centred data matrix and covariance

- Using the centred data matrix $\mathbf{X}^*$ we can express and manipulate the covariance matrix $\mathbf{S} = [s_{jk}^2]$ of $\mathbf{X}$ in a very convenient way as,

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

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 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 explains the maximum proportion of the data set total variability.
- The above dimension reduction is (particularly) effective when the original variables are (highly) correlated.
- PCA is one of the most widely used multivariate techniques.

Example: iris flower data set

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

iris

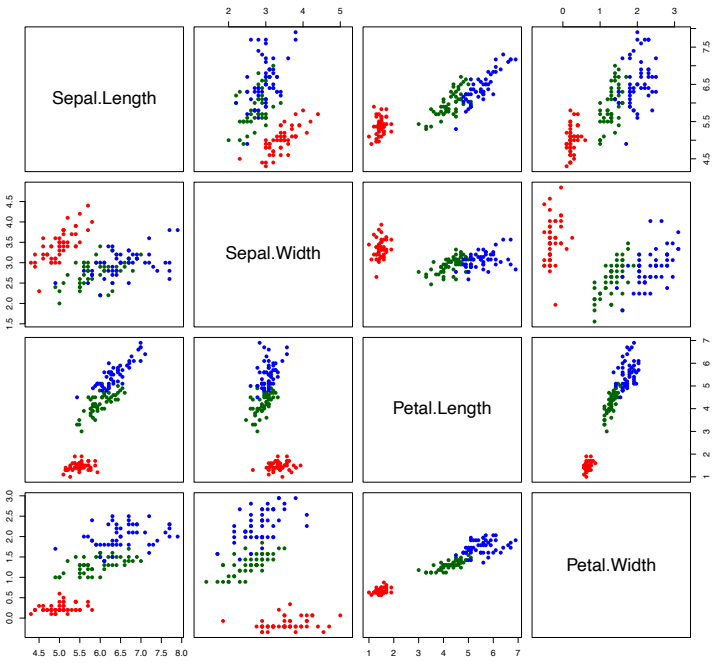
	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5	3.6	1.4	0.2	setosa
...					
51	7	3.2	4.7	1.4	versicolor
52	6.4	3.2	4.5	1.5	versicolor
53	6.9	3.1	4.9	1.5	versicolor
54	5.5	2.3	4	1.3	versicolor
55	6.5	2.8	4.6	1.5	versicolor
...					
101	6.3	3.3	6	2.5	virginica
102	5.8	2.7	5.1	1.9	virginica
103	7.1	3	5.9	2.1	virginica
104	6.3	2.9	5.6	1.8	virginica
105	6.5	3	5.8	2.2	virginica
...					
150	5.9	3	5.1	1.8	virginica

How to visualize the iris flower cloud of points?

- 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 the iris cloud on two dimensional linear spaces (planes), using all six possible combinations of two variables.

The "mosaic" of the 6 plane projections of the iris flowers

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



Best 2-dimensional representation using PCA

- Another approach is to construct the respective principal components (PCs), which are linear combinations of the original iris flowers measurements, and project on a lower dimensional space (using some of the PCs) keeping, as much as possible, the maximum of the dataset total variability.
- Actually, the projection of the cloud of iris flowers on the plane associated with the first two PCs, called principal factorial plane (PFP), retains 98.1% of the iris dataset variability and thus provides an excellent 2-dimensional portray of the original cloud of iris flowers.
- The PFP is in fact 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 total variability.

Best two-dimensional representation of the iris flowers

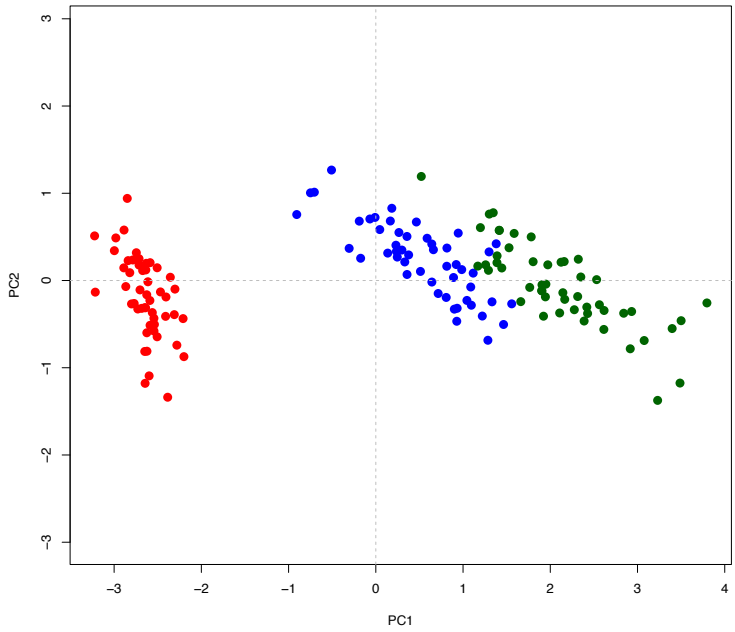


Figure: Projection of the 150 setosa, versicolor and virginica iris flowers.

Setting up PCA: eigen(values/vectors) of 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, that is, $\mathbf{S}^T = \mathbf{S}$.
- 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 non-correlated then the eigenvalues of $\mathbf{S}$ are strictly positive real numbers and $\mathbf{S}$ is invertible.
- We have an orthonormal set of p eigenvectors of $\mathbf{S}$,

  $$\left\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v_p}\right\},$$

  associated with the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_p$.

Setting up PCA: covariance between linear combinations

Given $\mathbf{a}, \mathbf{b} \in \mathbb{R}^p$, the covariance between the linear combinations $\mathbf{Xa}$ and $\mathbf{Xb}$ (see slide 5) is

$$\boxed{cov(\mathbf{Xa}, \mathbf{Xb}) = \mathbf{a}^T \mathbf{S} \mathbf{b}.} \tag{4}$$

Actually, using (3) of slide 18 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,

$$\boxed{var(\mathbf{Xa}) = \mathbf{a}^T \mathbf{S} \mathbf{a}.} \tag{5}$$

Exercise

Prove that centering a linear combination of p vectors $\mathbf{x}_1, \ldots, \mathbf{x}_p$ is equivalent to the linear combination of the p centred vectors $\mathbf{x}_1^*, \ldots, \mathbf{x}_p^*$ 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 $\mathbf{a} = (\alpha_1, \ldots, \alpha_p) \in \mathbb{R}^p$.

First principal component - formulation

To define the first principal component we seek a linear combination of the p variables $\mathbf{x}_1, \ldots, \mathbf{x}_p$ with maximimum variance, which by (5) is equivalent 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

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

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

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

Remark

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 of the form (see slide 12),

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

where $\{\mathbf{v}_1, \ldots, \mathbf{v}_p\}$ is an orthonormal set of $\mathbb{R}^p$ formed by eigenvectors of $\mathbf{S}$ and $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$ are the corresponding (real) eigenvalues.

Since $\|\mathbf{a}\| = 1$ we have, by the results of slide 11, a decomposition

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

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, together with relations $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$, $\|\mathbf{v}_i\|^2 = \mathbf{v}_i^T \mathbf{v}_i = 1$, for all i and $\mathbf{v}_i^T \mathbf{v}_j = 0$, $i \neq j$, we obtain by straightforward computations (all inner products envolving $\mathbf{v}_i$ and $\mathbf{v}_j$, $j \neq i$, cancel out),

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

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

$$\mathbf{a}^T \mathbf{S} \mathbf{a} = \mathbf{a}^T \lambda_1 \mathbf{a} = \lambda_1 \mathbf{a}^T \mathbf{a} = \lambda_1,$$

since $\mathbf{v}_1 = \mathbf{a}$ is a unit eigenvector of $\mathbf{S}$ associated to eigenvalue λ_1.

Thus the maximum variance of a linear combination $\mathbf{X}\mathbf{a}$ of the p columns $\mathbf{x}_1, \ldots, \mathbf{x}_p$, with unit vector of coefficients $\mathbf{a}$ is attained when $\mathbf{a} = \mathbf{v}_1$ is a (unit) eigenvector of $\mathbf{S}$ associated with the largest eigenvalue λ_1. Hence the first principal component is

$$\boxed{\text{PC}_1 : \quad \mathbf{y}_1 = \mathbf{X}\mathbf{v}_1 \quad \text{with maximum variance equal to } \lambda_1.}$$

The larger the value of λ_1, the more the cloud of points is elongated along the PC_1.

Second principal component - formulation

To define the second principal component PC_2, we seek a linear combination of the vectors containing the observations of the p variables, that maximizes the variance and is uncorrelated with PC_1. Since

$$0 = cov(\mathbf{Xa}, \mathbf{Xv}_1) = \mathbf{a}^T\mathbf{S}\mathbf{v}_1 = \mathbf{a}^T\lambda_1\mathbf{v}_1 = \lambda_1(\mathbf{a}^T\mathbf{v}_1) \Leftrightarrow \mathbf{a} \perp \mathbf{v}_1,$$

(assumindo $\lambda_1 > 0$), we can formulate the problem as

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

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,

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

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

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

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

$$\text{PC}_2 : \quad \mathbf{y}_2 = \mathbf{X}\mathbf{v}_2 \quad \text{with maximum variance equal to } \lambda_2$$

General case - formulation

In general, to define the j-th principal component PC_j, $j = 2, \ldots, p$, we seek a linear combination of the variables, $\mathbf{x}_1, \ldots, \mathbf{x}_p$ that maximizes the variance and is uncorrelated with $\text{PC}_1, \ldots, \text{PC}_{j-1}$, which is equivalent to the problem:

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

We construct in this way a collection of p variables, called principal components,

$$\text{PC}_1 : \; \mathbf{y}_1 = \mathbf{X}\mathbf{v}_1, \qquad \text{PC}_2 : \; \mathbf{y}_2 = \mathbf{X}\mathbf{v}_2, \qquad \ldots, \qquad \text{PC}_p : \; \mathbf{y}_p = \mathbf{X}\mathbf{v}_p,$$

with maximum variances,

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

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

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

Matrix of loadings

Each vector $\mathbf{v}_j$ contains the coefficients, also called loadings, of the j-th principal component w.r.t. the original variables $\mathbf{x}_1, \ldots, \mathbf{x}_p$. In other words, writing $\mathbf{v}_j = (\alpha_1, \ldots, \alpha_p)$, the j-th principal component PC_j is defined as the linear combination

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

We shall denote $\mathbf{V} = [\, \mathbf{v}_1 \ \cdots \mathbf{v}_p \,]$, which is called matrix of loadings.

- If the p eigenvalues of the covariance matrix $\mathbf{S}$ are pairwise distinct, i.e., $\lambda_1 > \cdots > \lambda_p \geq 0$, the vector of loadings defining each PC is unique up to sign: if $\mathrm{PC}_j = \mathbf{X}^* \mathbf{v}_j$ is a solution of (11) of slide 32, 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 hence can be defined in infinitely many distinct ways.

Scores matrix

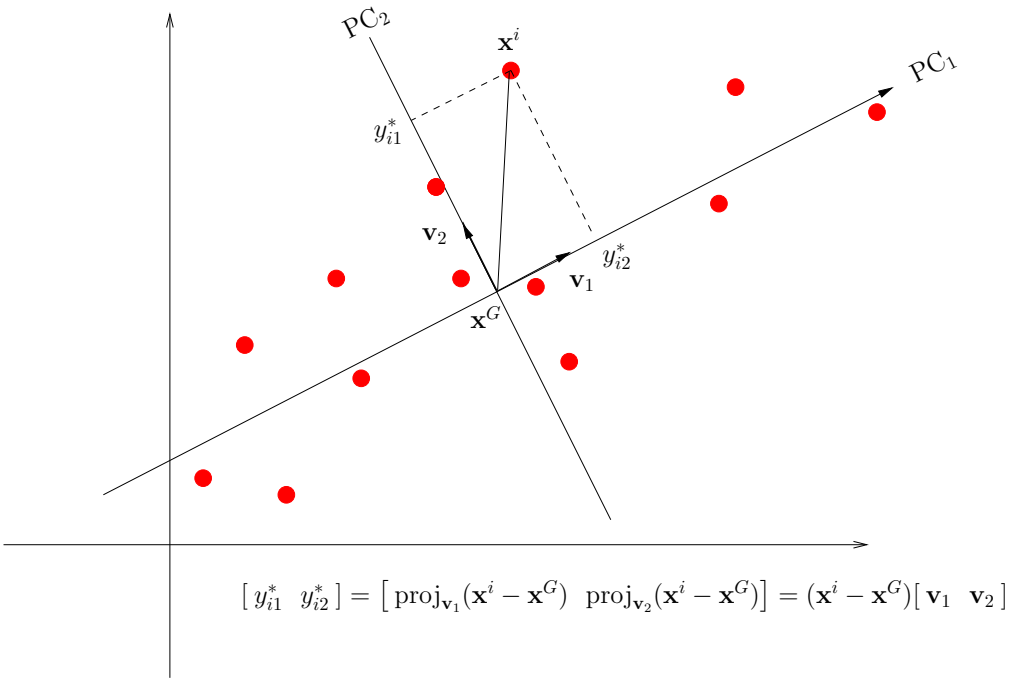
Recall that,

- $\mathbf{X}_{N \times p} = [x_{ij}]$ is the original data matrix containing the values of p variables across N individuals.
- $\mathbf{X}^T = [\, \mathbf{x}^1 \ \cdots \ \mathbf{x}^N \,]$, where $\mathbf{x}^i = (x_{i1}, \ldots, x_{ip})$ is i-th column of $\mathbf{X}^T$, that is, the i-th row of $\mathbf{X}$, corresponding to the coordinates of individual i in the cloud of N points of $\mathbb{R}^p$.
- $\mathbf{x}^G = (\bar{x}_1, \ldots, \bar{x}_p)$ is the center of gravity (also called barycenter) of the cloud of individuals.
- $\mathbf{X}^* = [x^*_{ij}]$ is the centred data matrix, where $x^*_{ij} = x_{ij} - \bar{x}_j$.
- $\mathbf{x}^i - \mathbf{x}^G = (x^*_{i1}, \ldots, x^*_{ip})$ the i-th row of $\mathbf{X}^*$, i.e., the vector of the coordinates of individual i in the centred cloud of N points (see slide 19).
- $\mathbf{V} = [\, \mathbf{v}_1 \ \cdots \mathbf{v}_p \,]$ is the matrix of loadings.

The rows of $\mathbf{Y}^* = \mathbf{X}^*\mathbf{V}$ contain the coordinates, also called scores, of the individuals in the centred data matrix $\mathbf{X}^*$ w.r.t the new coordinate axes defined by $\mathrm{PC}_1, \ldots, \mathrm{PC}_p$.

In other words, each row i of $\mathbf{Y}^*$ contains the coordinates $y^*_{i1}, \ldots, y^*_{ip}$ of the vector $\mathbf{x}^i - \mathbf{x}^G$ w.r.t the coordinate axes defined by the directions $\mathbf{v}_1, \ldots, \mathbf{v}_p$ (see next slide). We call $\mathbf{Y}^* = [y^*_{ij}]$ the scores matrix.

Scores of an individual i when $p = 2$



Covariance of the scores matrix

- The covariance matrix of the scores matrix $\mathbf{Y}^*$ is the diagonal matrix (since the PCs are uncorrelated), containing the variances $\lambda_1 \geq \ldots \geq \lambda_p$. Actually,

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

- The total variability of the scores matrix $\mathbf{Y}^*$ equals the dataset total variability. Actually,

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

- The quality of reduction of the cloud of points projected on the linear space spanned by the first k PCs ($1 \leq k \leq p$) is the proportion of variability explained by these k PCs, that is,

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

Covariance and correlation with the PCs

- Applying (4) from slide 27 we obtain the covariance between each variable $\mathbf{x}_j = \mathbf{X}\mathbf{e}_j$ ($\mathbf{e}_j = (0, 0, ..., 1, ..0)$ is the unitary vector that defines the x_j-axis) and each PC_k defined by $\mathbf{y}_k = \mathbf{X}\mathbf{v}_k$:

$$cov(\mathbf{x}_j, \text{PC}_k) = \mathbf{e}_j^T \mathbf{S}\mathbf{v}_k = \mathbf{e}_j^T \lambda_k \mathbf{v}_k = \lambda_k \mathbf{e}_j^T \mathbf{v}_k = \lambda_k v_{jk},$$

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

- Therefore the correlation between the variable $\mathbf{x}_j = \mathbf{X}\mathbf{e}_j$ and the k-th principal component PC_k, defined by $\mathbf{y}_k = \mathbf{X}\mathbf{v}_k$, is

$$cor(\mathbf{x}_j, \text{PC}_k) = \frac{cov(\mathbf{x}_j, \mathbf{y}_k)}{\sqrt{var(\mathbf{x}_j)}\sqrt{var(\mathbf{y}_k)}} = \frac{\lambda_k \, v_{jk}}{s_j \sqrt{\lambda_k}} = \frac{\sqrt{\lambda_k} \, v_{jk}}{s_j}.$$

- Variables more correlated (positively or negatively) with a given PC_k are usually considered more important to help to interpret the component. Those variables can however have lower absolute loadings since their variances also come into play. And vice-versa.

PCA via the R command prcomp

The following R code performs PCA on the covariance matrix of the iris flowers dataset using the command prcomp (to be detailed in next slides):

R code

```
X=iris[-5] # non standardized iris flowers data matrix
head(X) # shows the first 6 rows of X

iris.pca<-prcomp(X) # performs PCA on covariance matrix
iris.pca # consists of list with several components
summary(iris.pca) # std dev.  and % explained variances
iris.pca$sdev # standard deviations of the PCs
sum(iris.pca$sdev[1]^2) # total variance
V=iris.pca$rotation # matrix of loadings
Yc=iris.pca$x # matrix of scores
plot(iris.pca$x[,1:2],asp=TRUE,pch=16,col=c(rep("red",50),
rep("darkgreen",50),rep("blue",50))) # displays the
projection
```

More on R code - importance of principal components

The R command `summary(iris.pca)` of the previous slide gives, for each $j = 1, \ldots, 4$, the standard deviation $\sqrt{\lambda_j}$ associated with PC_j, the proportion of total variance explained by PC_j, $\frac{\lambda_j}{\sum_k \lambda_k}$, and the cumulative variance explained by the first j principal components:

	PC_1	PC_2	PC_3	PC_4
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

For the iris flower dataset we have that:

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

The R instruction `iris.pca$sdev` also returns the PCs standard deviations, $\sqrt{\lambda_1} = 2.0563$, $\sqrt{\lambda_2} = 0.49263$, $\sqrt{\lambda_3} = 0.2797$ and, $\sqrt{\lambda_4} = 0.15439$. Summing up the squares of these standard deviations we get the dataset total variability:

$$\texttt{sum(iris.pca\$sdev[1]\^{}2)} = 4.572957.$$

More on R code - loadings matrix

The R instruction `Yc <- iris.pca$rotation` returns the matrix of loadings below:

	PC_1	PC_2	PC_3	PC_4
Sepal.Length (SL)	0.3614	-0.6566	0.5820	0.3155
Sepal.Width (SW)	-0.0845	-0.7302	-0.5979	-0.3197
Petal.Length (PL)	0.8567	0.1734	-0.0762	-0.4798
Petal.Width (PW)	0.3583	0.0755	-0.5458	0.7537

Each column j of the loadings matrix `Yc`, say $\mathbf{v}_j = (\alpha_1, \ldots, \alpha_4)$, contains the coefficients of the linear combination defining the j-th principal component PC_j:

$$\mathbf{y}_j = \alpha_1\,\text{SL} + \alpha_2\,\text{SW} + \alpha_3\,\text{PL} + \alpha_4\,\text{PW}.$$

For instance, the first principal component PC_1 is defined by the linear combination of iris flowers measurements,

$$\begin{aligned} \mathbf{y}_1 &= 0.3614\,\text{SL} - 0.0845\,\text{SW} + 0.8567\,\text{PL} + 0.3583\,\text{PW} \\ &\approx 0.3614\,\text{SL} + 0.8567\,\text{PL} + 0.3583\,\text{PW}. \end{aligned}$$

Hence PC_1 is a new synthetic variable that represents a kind of weighted average of the iris flowers measurements, explaining a large amount ($\geq 90\%$) of the iris variability.

More on R code - properties of loadings matrix

The matrix of loadings $\mathbf{V} = [\, \mathbf{v}_1 \quad \cdots \quad \mathbf{v}_4 \,]$ of slide 38 verify the following properties:

- The columns $\mathbf{V}$ are unit and pairwise orthogonal vectors, that is, $\|\mathbf{v}_i\| = 1$ and $\mathbf{v}_i^T \mathbf{v}_j = 0$, $\forall i, j$, $i \neq j$, which amounts to say that $\mathbf{V}^T\mathbf{V}$ is the identity matrix, as can be check executing the following R instruction:

```
round(t(V)%*% V,10) # should return the identity matrix!
```

- The first column of $\mathbf{V}$, $\mathbf{v}_1 = (0.3614, -0.0845, 0.8567, 0.3583)$, is a (unit) eigenvector of the covariance matrix $\mathbf{S}$ of $\mathbf{X}$ associated with the largest eigenvalue λ_1 of $\mathbf{S}$, i.e., verifies $\mathbf{S}\mathbf{v}_1 = \lambda_1 \mathbf{v}_1$, where $\lambda_1 = 4.22837 = 2.0563^2$ is the variance explained by the first principal component:

```
v1 <- V[,1] # first column of V, i.e., vector of loadings for PC1
lambda1 <- iris.pca$sdev[1]^2 # variance of PC1
S <- cov(X) # covariance matrix of X
S %*% v1 ; lambda1*v1 # should be equal!
```

- The second column $\mathbf{V}$ is a (unit) eigenvector of $\mathbf{S}$ associated with the second largest eigenvalue $\lambda_2 = 0.49262^2 = 0.2426745$ of $\mathbf{S}$.
- and so on...

More on R code - scores matrix

The R instruction of slide 38, `scores <- iris.pca$x` returns the scores matrix below,

	PC1	PC2	PC3	PC4
1	-2.68413	-0.31940	0.02791	0.00226
2	-2.71414	0.17700	0.21046	0.09903
3	-2.88899	0.14495	-0.01790	0.01997
4	-2.74534	0.31830	-0.03156	-0.07558
⋮	⋮	⋮	⋮	⋮
23	-3.215939	-0.133468	-0.292397	0.004482
⋮	⋮	⋮	⋮	⋮

Each row i of this matrix contains the coordinates of the i-th iris flower w.r.t. the set of axes defined by the PCs, i.e., w.r.t. the new synthetic variables $\mathbf{y}_1^*, \ldots, \mathbf{y}_4^*$ (after centring). For instance, the 23-th iris flower has (approx.) coordinates $(-3.22, -0.13)$ in the PFP defined by PC_1 and PC_2 (see next slide).

As explained in slide 34, the scores matrix can also be defined as $\mathbf{Y}^* = \mathbf{X}^*\mathbf{V}$, which can be checked using the following R instructions:

```
Yc <- Xc %*% V # alternative definition of scores matrix
head(Yc) ; head(iris.pca$x) # should be equal!
```

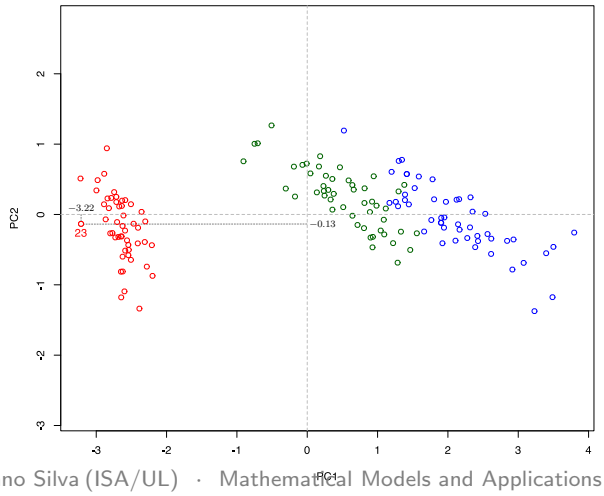
More on R code - projection of iris flowers on the PFP

The R code,

```
plot(iris.pca$x[,1:2],pch=16,
     col=rep(c("red","blue","darkgreen"),c(50,50,50)))

text(iris.pca$x[23,1],iris.pca$x[23,2],labels=23,pos=1,col=2)
abline(h=0,lty=2)
abline(v=0,lty=2)
```

projects the set of 150 iris flowers setosa, versicolor, virginica, on the PFP, and labels the dot representing iris flower # 23, whose coordinates (scores) on the PC_1 and PC_2 axes are approximately equal to -3.2 and -0.13 resp. (see also the previous slide).



Additional interpretation elements: contributions and squared cosines

- The contribution of an individual i to the construction of a principal component PC_k is the proportion of the variance of PC_k due to the individual i:

$$\text{ctr}_{i,k} = \frac{(y^*_{i,k})^2}{\sum_{j=1}^{N}(y^*_{j,k})^2} = \frac{(y^*_{i,k})^2}{\lambda_k}.$$

  Individuals with contributions above the average contribution $1/N$ are usually more important to interpret the PC.

- A related notion is the squared 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:

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

  where $d_i^2 = \sum_{j=1}^{p}(y^*_{i,j})^2$ is the squared distance of individual i to the origin (in the centred cloud of individuals). Squared 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 an 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 distances between well represented individuals should be interpreted!

More on R code - correlations, contrib. and squared cosines

R code

```
X <- iris[-5] ; iris.pca<-prcomp(X,scale=FALSE)

N <- nrow(X) # number of individuals (iris flowers)

Xc <- scale(X,scale=FALSE) # centred iris data matrix

Yc <- Xc %*% V # scores matrix

head(Yc) ; head(iris.pca$x) # should be equal!

round(cor(iris[-5],iris.pca$x) # cor between each variable and each PC

# total variance
sum(iris.pca$x[,1]^2)/(N-1) ; iris.pca$sdev[1]^2 # should be equal!

# contributions of each individual to the 1st PC
Yc[,1]*Yc[,1]/sum(Yc[,1]*Yc[,1])

# individuals with contribution above the average
Yc[,1]*Yc[,1]/sum(Yc[,1]*Yc[,1]>1/N

# vector of quality of representation (cos2) of individual i in each PC
Yc[1,]*Yc[1,]/sum(Yc[1,]*Yc[1,])

# quality of representation of individual #1 in the PFP
(Yc[1,1]*Yc[1,1]+Yc[1,2]*Yc[1,2])/sum(Yc[1,]*Yc[1,])
```


Drawbacks of the PCA on the covariance matrix

- The first PC tends to be dominated by the variable(s) with higher variance(s) since the PCA seeks the linear combinations $\sum_i \alpha_i \, \mathbf{x_i}$, with $\sum_i \alpha_i^2 = 1$, that maximizes,

$$\text{var}\left(\sum_i \alpha_i \, \mathbf{x_i}\right) = \sum_i \alpha_i^2 \, var(\mathbf{x_i}) + 2 \sum_{i<j} \alpha_i \, \alpha_j \, cov(\mathbf{x_i}, \mathbf{x_j}).$$

- The PCs are invariant under orthogonal transformations of the variables (e.g. rotations), but not under differentiated change of scales in the variables. 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?

Remark

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

Standardized data matrix and correlation

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

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

Gathering these p vectors together we obtain the 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 $\mathbf{Z}$, i.e.,

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

- and the total variability of $\mathbf{Z}$ equals

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

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 defined by variables $\mathbf{y}_j = \mathbf{Z}\mathbf{v}_j$ where $\mathbf{v_1}, \ldots, \mathbf{v_p}$ are unit and pairwise orthogonal eigenvectors with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p$.
- The total variance is now the number of variables,

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

  and the correlation coefficient between $\mathbf{z}_j$ and $\mathbf{y}_k$ reduces to,

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

- Variables j with higher absolute loadings $|\mathbf{v}_{jk}|$ w.r.t. PC$_k$, are more correlated with PC$_k$ and are usually more important to interpret the PC. The presence of variables with loadings of different signs with respect to the same PC may indicate that the component opposes individuals according to the values of those variables.

Interpretation of the results in the 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 $\sqrt{N-1}$:

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

- The length of PC_k is proportional to its standard deviation:

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

- 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 $\mathbf{y}_k$ (note that $\mathbf{y}_k$ and $\mathbf{z}_j$ are centred vectors):

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

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

PCA on the correlation matrix - summary

- The variables have the same variance and therefore their importance is equalized.
- The cloud of individuals tends to have a more rounded shape.
- The PCA tends to reflect existing correlation patterns among variables.
- The first PC tends to be dominated by a group of highly correlated variables since the PCA seeks the linear combinations $\sum_i \alpha_i \, \mathbf{z_i}$, with $\sum_i \alpha_i^2 = 1$, maximizing $var(\sum_i \alpha_i \, \mathbf{z_i})$, that is, maximizing

$$\sum_i \alpha_i^2 \, var(\mathbf{z_i}) + 2 \sum_{i<j} \alpha_i \, \alpha_j \, cov(\mathbf{z_i}, \mathbf{z_j}) = p + 2 \sum_{i<j} \alpha_i \, \alpha_j \, cor(\mathbf{z_i}, \mathbf{z_j}).$$

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

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 variance (usually between 70% and 90%) and to choose the smallest number m of PCs such that the % of explained variance by the first m PCs exceeds the chosen %.
- Scree test: to look for an 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 variable 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 first k PCs while $\lambda_k > \ell_k^*$.

Preamble to **biplots**: a very useful decomposition...

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

$$\mathbf{C} = \mathbf{A}\,\mathbf{B}^T, \tag{12}$$

with $\mathbf{A}_{N\times r}$ and $\mathbf{B}_{p\times r}$, which are called left and right matrices, respectively. In other words, each element of $\mathbf{C}$ can be written as the inner product, in $\mathbb{R}^r$, of a row of a matrix $\mathbf{A}$ by a row of a matrix $\mathbf{B}$. The decomposition can be accomplished using SVD.

In the particular case that target matrix $\mathbf{C}$ has rank one, i.e., has proportional rows and proportional columns, we obtain a decomposition as

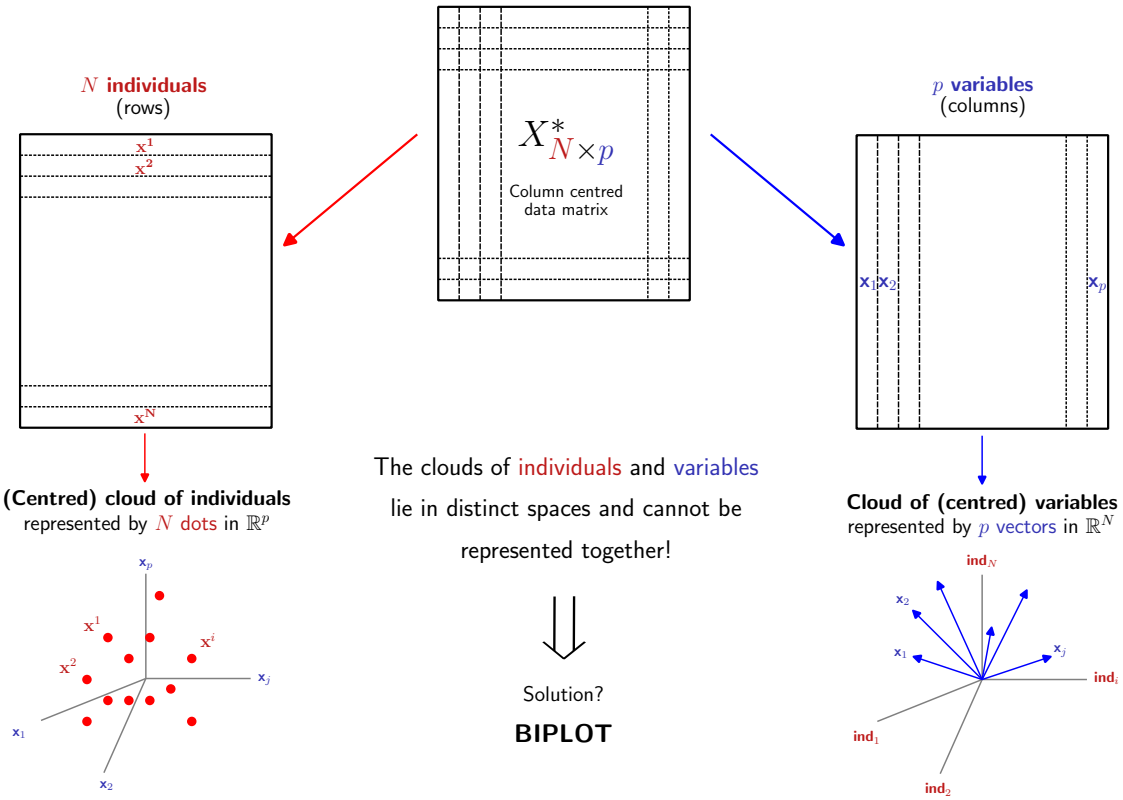
$$\mathbf{C} = \mathbf{a}\,\mathbf{b}^T = \begin{bmatrix} a_1 \\ \vdots \\ a_N \end{bmatrix} \begin{bmatrix} b_1 & \cdots & b_p \end{bmatrix}, \quad \text{with}$$

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

The decomposition (12) is not unique. For instance, in the previous case we can write,

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

Motivation to **biplots**: simultaneous representation of individuals and variables



Biplots

Biplots provide simultaneous representations of individuals and variables of a column centred data matrix $\mathbf{X}^*_{N\times p}$ (of rank r) in a low dimensional space, usually of dimension two or three, by decomposing the data matrix $\mathbf{X}^*$ as a product similar to the one described in slide 52:

$$\mathbf{X}^*_{N\times p} = \mathbf{G}_{N\times r}\,\mathbf{H}^T_{r\times p}.$$

We shall denote, as in slide 53, the p columns of $\mathbf{X}^*$ (variables) by $\mathbf{x}_1,\ldots,\mathbf{x}_p$ and the N rows of $\mathbf{X}^*$ (individuals) by $\mathbf{x}^1,\ldots,\mathbf{x}^N$, that is,

$$\mathbf{X}^* = [\,\mathbf{x}_1 \;\cdots\; \mathbf{x}_p\,] \quad\text{and}\quad (\mathbf{X}^*)^T = [\,\mathbf{x}^1 \;\cdots\; \mathbf{x}^N\,].$$

Likewise, denoting the N the rows of $\mathbf{G}$, that is, the N columns of $\mathbf{G}^T$, by $\mathbf{g}^1,\ldots,\mathbf{g}^N \in \mathbb{R}^r$, we can write

$$\mathbf{G}^T_{r\times N} = [\,\mathbf{g}^1 \;\cdots\; \mathbf{g}^N\,].$$

Similarly, denoting the p rows of $\mathbf{H}$, that is, the p columns of $\mathbf{H}^T$, by $\mathbf{h}^1,\ldots,\mathbf{h}^p \in \mathbb{R}^r$, we can write

$$\mathbf{H}^T_{r\times p} = [\,\mathbf{h}^1 \;\cdots\; \mathbf{h}^p\,].$$

Markers of variables and markers of individuals

- With the notations of the previous slide, relation $\mathbf{X}^* = \mathbf{G}\,\mathbf{H}^T$ can be written as,

$$[\,\mathbf{x}_1 \;\; \cdots \;\; \mathbf{x}_p\,] = \mathbf{G}[\,\mathbf{h}^1 \;\; \cdots \;\; \mathbf{h}^p\,] = [\,\mathbf{G}\,\mathbf{h}^1 \;\; \cdots \;\; \mathbf{G}\,\mathbf{h}^p\,],$$

that is, each variable $\mathbf{x}_j$ can be written as,

$$\mathbf{x}_j = \mathbf{G}\,\mathbf{h}^j, \qquad j = 1, \ldots, p,$$

and we call the row $\mathbf{h}^j$ of $\mathbf{H}$ the marker of variable $\mathbf{x}_j$.

- Similarly, $(\mathbf{X}^*)^T = (\mathbf{G}\,\mathbf{H}^T)^T = \mathbf{H}\mathbf{G}^T$, which can be written as,

$$[\,\mathbf{x}^1 \;\; \cdots \;\; \mathbf{x}^N\,] = \mathbf{H}[\,\mathbf{g}^1 \;\; \cdots \;\; \mathbf{g}^N\,] = [\,\mathbf{H}\,\mathbf{g}^1 \;\; \cdots \;\; \mathbf{H}\,\mathbf{g}^N\,],$$

that is, each individual $\mathbf{x}^i$ can be written as,

$$\mathbf{x}^i = \mathbf{H}\,\mathbf{g}^i, \qquad i = 1, \ldots, N,$$

and we call the row $\mathbf{g}^i$ of $\mathbf{G}$ the marker of individual $\mathbf{x}^i$.

Remarks

- Each element $x_{i,j}$ of X^*, i.e.,, each value of an individual i w.r.t a variable j, is the inner product of their corresponding individual marker $\mathbf{g}^i$ and variable marker $\mathbf{h}^j$.

- Both markers of variables and markers of individuals lie in the same space $\mathbb{R}^r$, with $r = \text{rank}(\mathbf{X}^*)$, and can be meaningfully represented together, giving rise to the so-called biplot.

- There is a close relationship between biplots and PCA when the decomposition is obtained applying SVD to the target matrix, as we shall see in the next slides.

Biplots via SVD

To simplify the discussion we will assume hereafter that $\mathbf{X}^*$ has rank p (we are assuming $N \geq p$). Applying the singular value decomposition (SVD) to $\mathbf{X}^*$ we obtain a decomposition,

$$\mathbf{X}^* = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}^T \tag{13}$$

where,

- $\mathbf{U}_{N\times p}$ is the matrix of left singular vectors and verifies $\mathbf{U}^T\mathbf{U} = \mathbf{I}_p$;
- $\mathbf{V}_{p\times p}$ is the matrix of right singular vectors and verifies $\mathbf{V}^T\mathbf{V} = \mathbf{I}_p$;
- $\boldsymbol{\Delta}_{p\times p} = \text{diag}(\delta_1, \ldots, \delta_p)$ is the diagonal matrix containing the singular values of $\mathbf{X}^*$ and verifies $\delta_1 \geq \cdots \geq \delta_p > 0$.

Using the decomposition (13) above we can write $\mathbf{X}^* = \mathbf{G}\mathbf{H}^T$ in many different ways. We will refer here two of them;

I. $\mathbf{G} = \mathbf{U}\boldsymbol{\Delta}$ and $\mathbf{H} = \mathbf{V}$ - focuses on distances between individuals.

II. $\mathbf{G} = \mathbf{U}$ and $\mathbf{H} = \mathbf{V}\boldsymbol{\Delta}$ - focuses on covar./correl. between variables.

I. Biplots focusing on distances

In the first case where $\mathbf{G} = \mathbf{U}\boldsymbol{\Delta}$ and $\mathbf{H} = \mathbf{V}$, the matrix $\mathbf{G}$ contains the left singular vectors scaled by their respective singular values, which gives rise to the factor scores (coordinates) of the individuals.

Actually, the right singular vectors of the column centred data matrix $\mathbf{X}^*$ are eigenvectors of the covariance matrix $\mathbf{S}$, i.e, vectors of loadings of $\mathbf{X}^*$ and thus the scores matrix is (see slide 77),

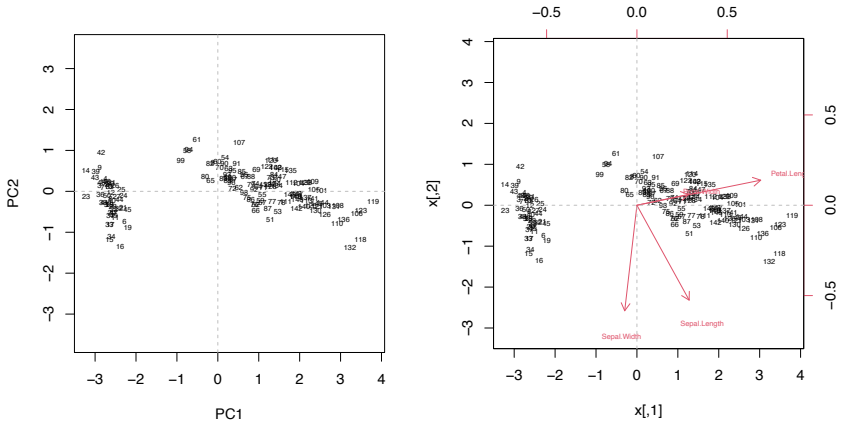
$$\mathbf{Y}^* = \mathbf{X}^*\mathbf{V} = (\mathbf{U}\boldsymbol{\Delta}\mathbf{V}^T)\mathbf{V} = \mathbf{U}\boldsymbol{\Delta}(\mathbf{V}^T\mathbf{V}) = \mathbf{U}\boldsymbol{\Delta}$$

(since $\mathbf{V}^T\mathbf{V}$ is the identity matrix). The matrix $\mathbf{H} = \mathbf{V}$ corresponds to the matrix of right singular vectors, i.e., to the matrix of loadings.

Projection of the iris flowers on the PFP and biplot - I

The projections of iris flowers on the principal factorial plane (defined by PC_1 and PC_2) via PCA (on the left) and using biplot-I (on the right) coincide. The higher the % of variance explained by PC_1 and PC_2, the better distances between projected flowers approximate true distances.

The loadings (coefficients) of each variable w.r.t. PC1 and PC2 can be read out from the positions of the variables labels in the biplot, w.r.t. up and right scales, respectively. It is clear from the biplot that among the four variables Petal.Length has the highest loading w.r.t. PC1 (≈ 0.86).



R code for the image of the previous slide

The `biplot` function allows to display the biplot of a dataset either specifying the $\mathbf{U}$, $\mathbf{V}$ matrices or using the default option, which displays the biplot - II:

R code

```
Xc <- scale(iris[-5],scale=FALSE) # centred iris flower dataset
iris.svd <- svd(Xc) # compute the svd UDV^T of the centred iris dataset
U <- iris.svd$u # left singular vector
V <- iris.svd$v # right singular vectors = loadings matrix
rownames(V)<-colnames(Xc) # to name the variables of loadings matrix
Delta <- diag(iris.svd$d) # diagonal matrix with the singular values
par(mfrow=c(1,2)) # 2 side-by-side windows
plot(iris.pca$x[,1:2],asp=TRUE,pch=16,cex=.01) # small points in PFP
text(iris.pca$x[,1],iris.pca$x[,2],labels=1:150,cex=.5) # labels
abline(h=0,lty=2,col="gray")
abline(h=0,lty=2,col="gray")
biplot(U%*%Delta,V,asp=TRUE,cex=.5)#G=U Delta (scores);H=V (loadings)
abline(h=0,lty=2,col="gray")
abline(h=0,lty=2,col="gray")
```

II. Biplots focusing on covariances/correlations

Considering $\mathbf{G} = \mathbf{U}$ and $\mathbf{H} = \mathbf{V}\mathbf{\Delta}$ we obtain,

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

Hence

$$(N-1)s_{jk}^2 = (\mathbf{h}^j)^T\mathbf{h}^k, \qquad \forall i,j,$$

and therefore the inner product between two variable markers $\mathbf{h}^j$ and $\mathbf{h}^k$ is proportional to the covariance s_{jk}^2 between the corresponding variables $\mathbf{x}_j = \mathbf{G}\mathbf{h}^j$ and $\mathbf{x}_k = \mathbf{G}\mathbf{h}^k$.

In particular, $\sqrt{N-1}\,s_j = \sqrt{(\mathbf{h}^j)^T\mathbf{h}^j} = \|\mathbf{h}^j\|$, and therefore the length of a variable marker is proportional to the standard deviation of the corresponding variable.

Moreover, denoting by θ_{jk} the angle between the variable markers $\mathbf{h}^j$ and $\mathbf{h}^k$, the correlation between the corresponding variables $\mathbf{x}_j = \mathbf{G}\mathbf{h}^j$ and $\mathbf{x}_k = \mathbf{G}\mathbf{h}^k$ is

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

It can be proved that a similar conclusion holds replacing one of the variables by a principal component.

Euclidean and Mahalanobis distances

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

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

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 individuals $\mathbf{x}^i, \mathbf{x}^\ell \in \mathbb{R}^p$ is

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

The importance of Mahalanobis distance in the biplot context resides in the fact that the (squared) Mahalanobis distance between individuals,

$$\mathbf{x}^i = \mathbf{H}\mathbf{g}^i \quad \text{and} \quad \mathbf{x}^\ell = \mathbf{H}\mathbf{g}^\ell,$$

is proportional to the (squared) euclidean distance between the corresponding individual markers,

$$\mathbf{g}^i \quad \text{and} \quad \mathbf{g}^\ell.$$

Euclidean and Mahalanobis distances (cont.)

Actually, from relation (15) of slide 77, we obtain,

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

and therefore the (squared) euclidean distance (multiplied by $N-1$) between the individual markers $\mathbf{g}^i$ and $\mathbf{g}^\ell$ is given by,

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

The last expression is precisely the (squared) Mahalanobis distance between the individuals, $\mathbf{x}^i = \mathbf{H}\mathbf{g}^i$ and $\mathbf{x}^\ell = \mathbf{H}\mathbf{g}^\ell$.

Summary of "exact" interpretations for biplot - II

- The standard deviation of a variable is proportional to the length of the corresponding variable marker.
- The correlation coefficient of two variables is the cosine of the angle between the corresponding variable markers in the biplot-II.
- The correlation of a variable and a principal component is the cosine of the angle between the corresponding variable marker and the PC axis in the biplot-II.
- The Mahalanobis distance between two individuals is proportional to the euclidean distance between the corresponding individual markers.
- The value of a (centred) variable for an individual assumes is proportional to the coordinate of the orthogonal projection of the individual marker onto the line defined by the variable marker, divided by the variable standard deviation.

The last property follows from relation $\mathbf{X}^* = \mathbf{G}\mathbf{H}^T$, i.e., $x^*_{ij} = (\mathbf{g}^i)^T\mathbf{h}^j,\ \forall i,j$:

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

"Approximated" interpretations for biplot - II

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 $\mathbf{H}^T$, respectively, and denote

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

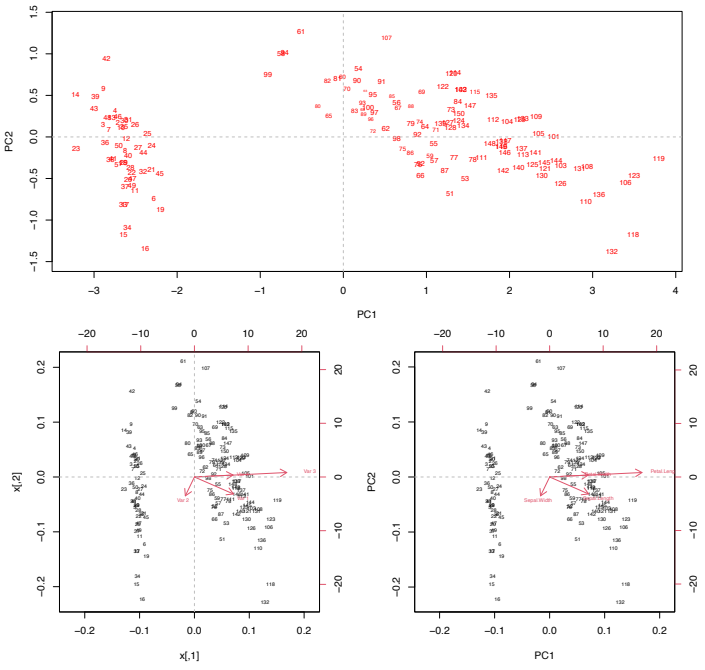
The rows of $G(m)$ and $H(m)$ are approximations of the markers of individuals and variables, respectively. Hence we have:

- The length of a variable marker is approximately proportional to the variable standard deviation.
- The cosine of the angle between two variable markers is approximately equal to the correlation between the corresponding variables.
- The cosine of the angle between a variable marker and an axis of a principal component is approximately equal to the correlation between the corresponding variable and the corresponding principal component.
- 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 proportional to the value of the individual on that variable (divided by the variable standard deviation).

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

PCA vs biplot - II

The top row image contains the projection on the PFP of the iris flowers using PCA, with flowers label sizes proportional to their quality representations (squared cosines). The bottom row images contain two identical biplots - II, one obtained specifying the $\mathbf{U}$, $\mathbf{V}$ matrices (on the left) and the other via the `biplot` function (on the right).



Some interpretations on the images of slide 66

Jointly PC_1 and PC_2 explain about 98% of the dataset total variability (see slide 39) allowing for the following examples of "almost exact" interpretations:

- Iris flower $\#63$ is the worst represented flower in PFP (smallest label), with a quality of representation (squared cosine) ≈ 0.434, meaning that PC_3 and PC_4 may have important contributions for the distance of flower $\#63$ to the origin. Hence distances between flower $\#63$ and the remaining flowers should be interpreted with caution. Most of the other flowers are well represented in the PFP.
- PL has the highest variance among the 4 variables since the corresponding variable marker has the highest length (approx. doubling the others variables lengths).
- The correlation of PL and PW is almost one since the angle between the corresponding variable markers is almost zero. On contrary, the correlation between PL and SW is small since the corresponding variable markers are nearly pairwise orthogonal.
- Likewise, the correlations of PL and PW with PC_1 are very close to one and between PL and PW with PC_2 close to zero.

More interpretations on the images of slide 66

- Almost all iris flowers from versicolor and virginica species, i.e., numbered from 51 to 150, have PL and PW values above the mean since the orthogonal projections of those flowers onto the line defined by the PL and PW variable markers lie in the "positive side", with $\#119$ flower attaining one of the largest PL and PW values.
- In the opposite direction, all setosa flowers, i.e., numbered from 1 to 50, have PL and PW values below the mean since the projections of these flowers onto the line defined by the PL and PW variables clearly lie in the negative side, with $\#23$ attaining one of the smallest PL and PW values.
- Since the correlations between PL and PW with PC1 are very close to one, PC1 contrasts flowers with smaller PL and PW values against flowers with larger PL and PW values and thus can be regarded as a petal measurements axis.
- Distances between (well represented) iris flowers markers in the top row image (obtained using PCA) correspond to euclidean distances between the flowers, while the distances in the bottom row biplots-II between the flowers markers correspond to Mahalanobis distances between the iris flowers. Comparing the biplots in the above and bottom rows we can observe a compression of the distances along PC1 axis, which reflects the fact Mahalanobis distances are smaller along the cloud directions of greater variability (see slide 70).

R code for the images of slide 66

R code

```
layout(matrix(c(1,1,2,3), 2, 2, byrow = TRUE)) #layout of the 3 displays
Yc <-iris.pca$x # scores matrix
qualRepr <- rep(NA,150) #
for (i in 1:150)
qualRepr[i] <- (Yc[i,1]*Yc[i,1]+Yc[i,2]*Yc[i,2])/sum(Yc[i,]*Yc[i,])
plot(iris.pca$x[,1:2],asp=TRUE,pch=16,cex=.01) # projection in the PFP
text(iris.pca$x[,1],iris.pca$x[,2],labels=1:150,cex=.75*qualRepr) #
iris.svd <- svd(Xc) # decomposition Xc=UDV^T using svd
U <- iris.svd$u # left singular vector
V <- iris.svd$v # righ singular vector
Delta <- diag(iris.svd$d) # diagonal matrix with singular values
abline(h=0,lty=2,col="gray") # PC1 axis
abline(v=0,lty=2,col="gray") # PC2 axis
biplot(U, V% * %Delta, asp=TRUE,cex=.5) # biplot - II: G=U and H=V Delta
abline(h=0,lty=2,col="gray")
abline(v=0,lty=2,col="gray")
biplot(iris.pca, asp=TRUE,cex=.5) # by default computes biplot - II
```


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 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 $\mathbf{y}$. In particular, if 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.
- Mahalanobis distances between individuals or between an individual and the cloud's center of gravity are "smaller" along the cloud directions of greater variability.
- Mahalanobis distance is the multivariate generalization of how many standard deviations is an individual far away from the cloud's mean, used, for instance, to detect outliers.

PCA interpretation - summary

- Proportion of variance explained by each PC
- Correlations and respective signs between variables and PCs
- Contributions of individuals to PCs above the average
- Well represented individuals, i.e, with high square cosines
- Biplots

Hors-d'oeuvre

A more geometrical approach to PCA

using the SVD of the centred data matrix

The best rank k linear approximation $\mathbf{X}(k)$ of $\mathbf{X}^*$

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

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

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 $k = 1, \ldots, p$ we obtain the rank k linear approximation of $\mathbf{X}^*$,

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

where $\mathbf{U}(k)$ is the submatrix of $\mathbf{U}$ containing the first k columns, and so on...

Rank one and rank two linear approximations of $\mathbf{X}^*$

When $k = 1, 2$ we obtain the following rank one and rank two linear approximations,

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

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, by $d_{i'}$ the distance between i' and the origin and setting $e_i = d(i, i')$, we have a decomposition

$$d_i^2 = d_{i'}^2 + e_i^2. \tag{14}$$

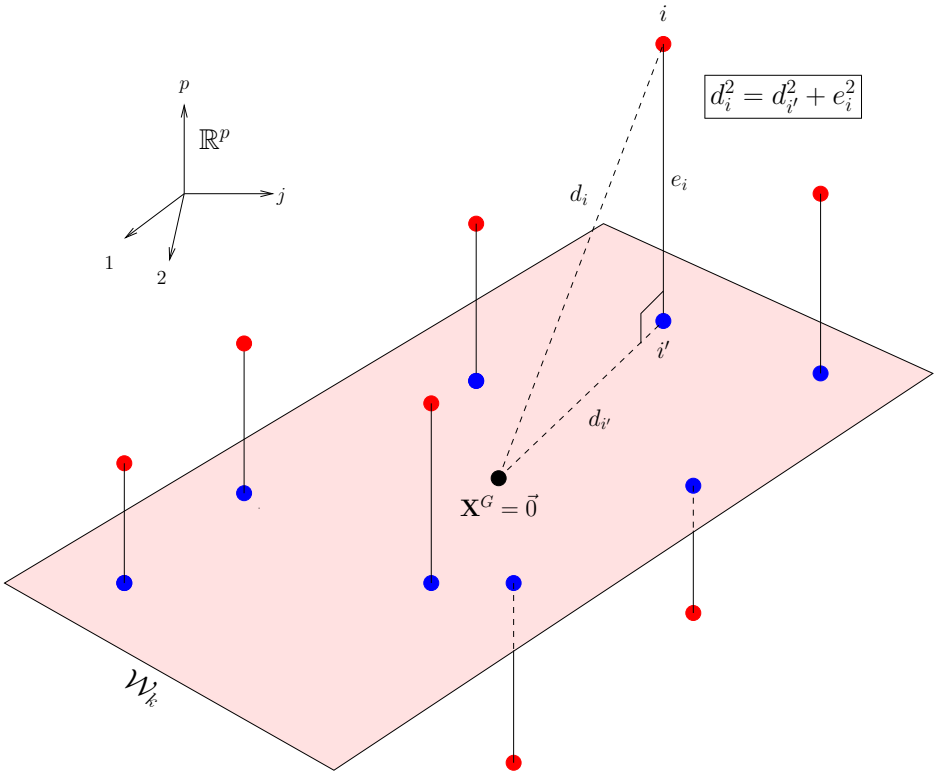
Best fitting k-dimensional linear space

- The cloud of (blue) points $\mathbf{X}(k)$ gives the best rank k linear approximation of $\mathbf{X}^*$, corresponding to the best fitting of a 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 (14) of the slide 74 we obtain (see also next slide),

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

- 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), which is the main goal of PCA!

Best fitting k-dimensional linear space (cont.)



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):

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

which is equivalent to say that,

$$\mathbf{S} = \mathbf{V}\left(\frac{1}{\sqrt{N-1}}\boldsymbol{\Delta}\right)^2\mathbf{V}^T \tag{15}$$

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 $\sqrt{\lambda_j}$, the singular values δ_j of $\mathbf{X}^*$ divided by $\sqrt{N-1}$. The PC factor scores are given by

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

and the left singular vectors verify

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

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

Remark

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

Equivalence between PCA via EVD and via SVD

R code

```
# 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
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) # should be equal!
# (eigenvalues of S = squared singular values of Xc divided by N-1)

head(loadings); head(right.sing) # should be equal!
# (loadings = right sing vectors)

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

Main bibliography

- J. Cadima, *Introduction to Multivariate Statistics*, Slides for the MMA course 2021/22.
- IT Jolliffe (2002) *Principal component analysis*, 2nd edn. New York, NY: Springer-Verlag.
- IT Jolliffe, J Cadima (2016) *Principal component analysis: a review and recent developments*, Phil. Trans. R. Soc. A 374: 20150202 (http://dx.doi.org/10.1098/rsta.2015.0202).
- M Greenacre, T Hastie, P Groenen, A Iodice d'Enza, A Markos and E Tuzhilina (2022) *Principal Component analysis*, Nature Reviews Methods Primers volume 2, Article number: 100 (2022) (http://dx.doi.org/10.1038/s43586-022-00184-w).
- M. Greenacre (2010), *Biplots in Pratice*, Fundación BBVA (ISBN: 978-84-923846-8-6)