Data analysis
Principal component analysis

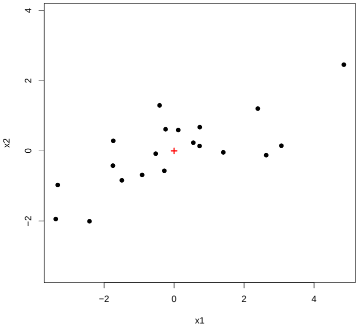
Olivier Roustant

September 19, 2020

Principal Component Analysis (PCA): Outline

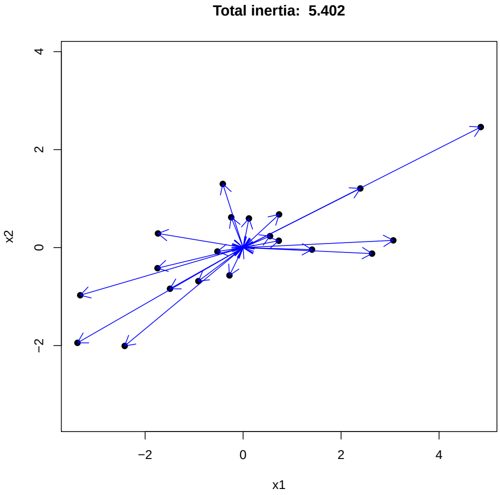
1. **Figures only!**
2. **Theory**
3. **Variations (metric, weights)**
4. **Results interpretation**
5. **Conclusion and further readings**

The aim: To reduce dimension



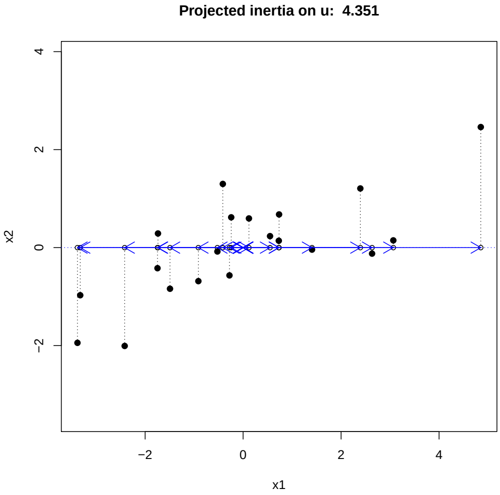
This is a $2D$ cloud of points, centered at 0.
Can you find a 1D axis 'containing' the maximum of information?

Inertia



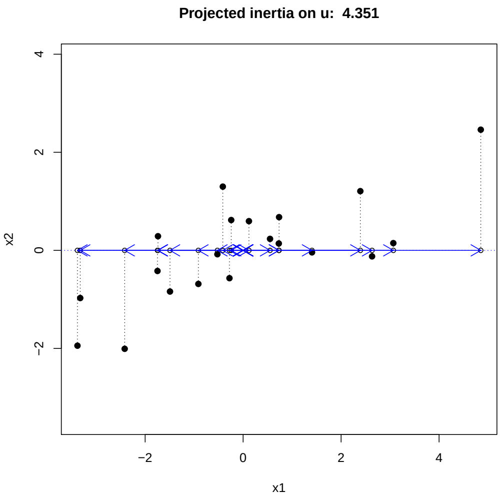
Total inertia: mean square of distances to the center.

Inertia



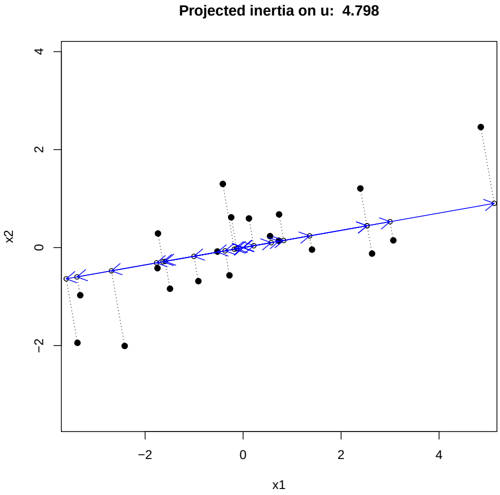
Projected inertia: inertia of projections. How much do we lose?

Maximizing the projected inertia



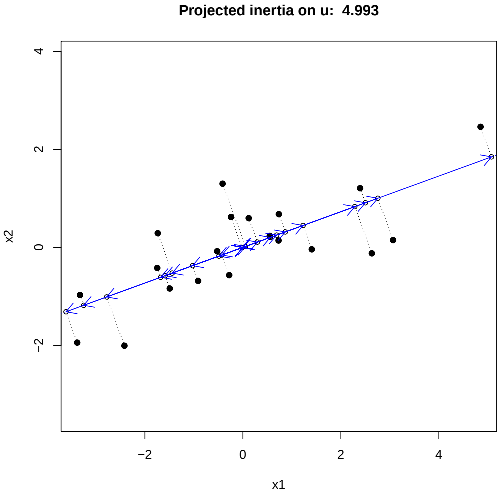
Projected inertia: For what axis is it maximal?

Maximizing the projected inertia



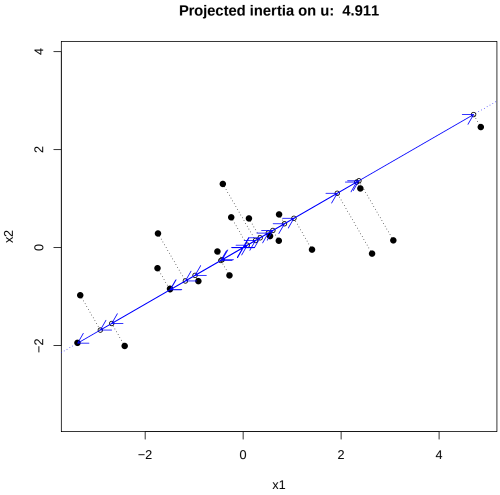
Projected inertia: For what axis is it maximal?

Maximizing the projected inertia



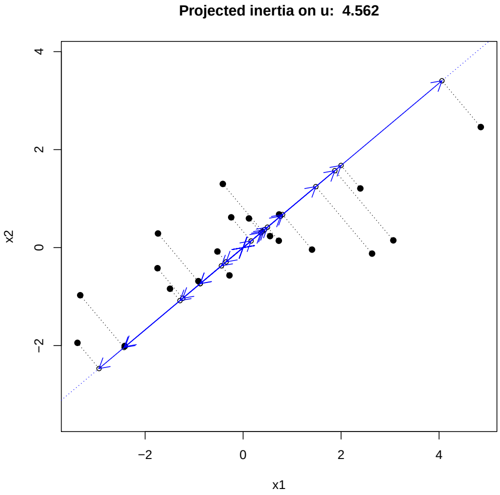
Projected inertia: For what axis is it maximal?

Maximizing the projected inertia



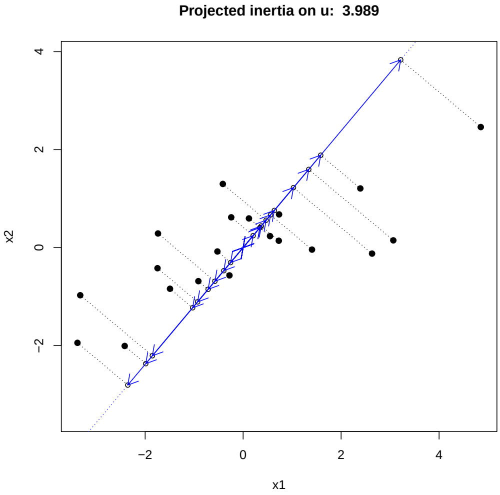
Projected inertia: For what axis is it maximal?

Maximizing the projected inertia



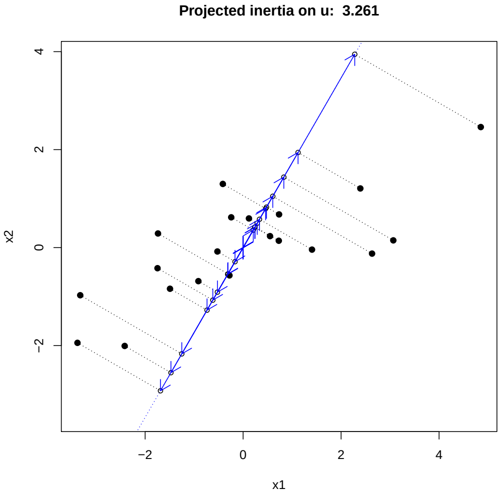
Projected inertia: For what axis is it maximal?

Maximizing the projected inertia



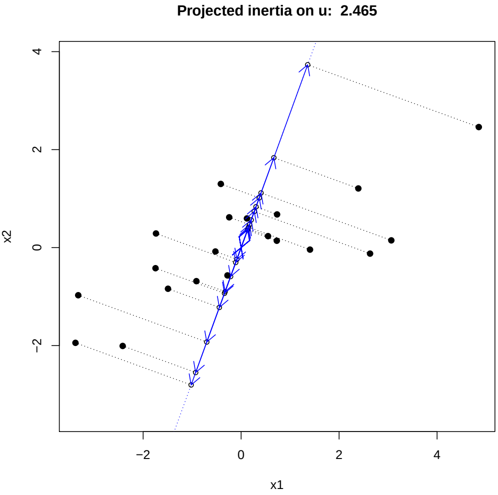
Projected inertia: For what axis is it maximal?

Maximizing the projected inertia



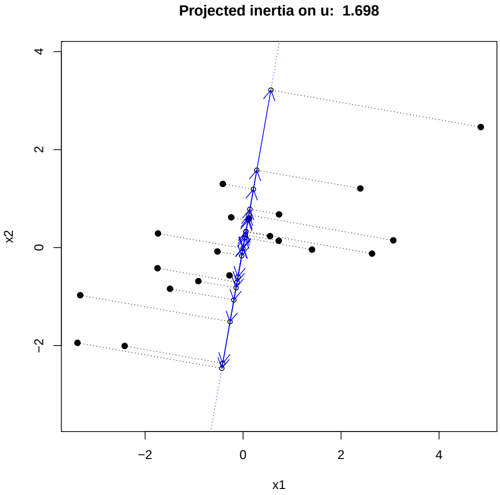
Projected inertia: For what axis is it maximal?

Maximizing the projected inertia



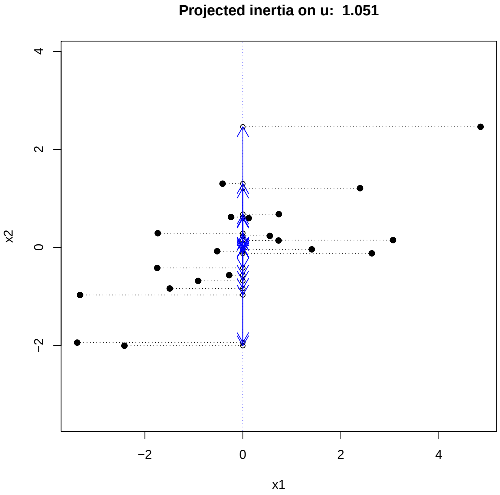
Projected inertia: For what axis is it maximal?

Maximizing the projected inertia



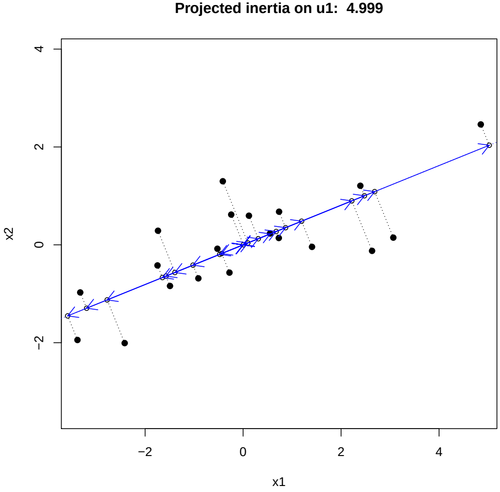
Projected inertia: For what axis is it maximal?

Maximizing the projected inertia



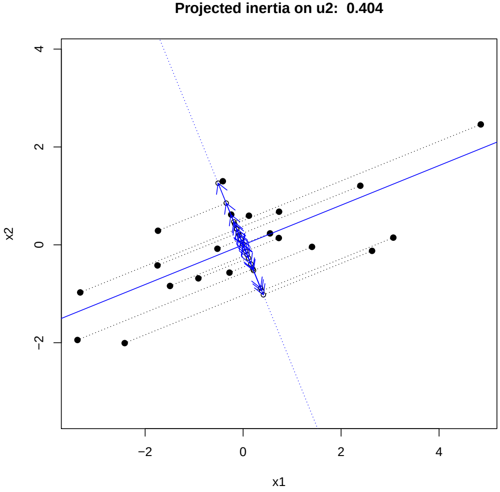
Projected inertia: For what axis is it maximal?

Maximizing the projected inertia



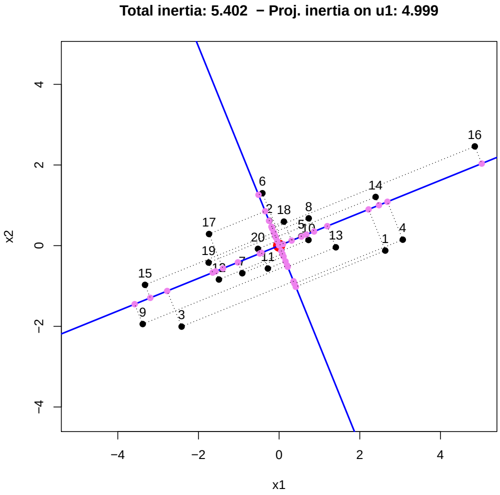
Projected inertia: Maximal for the largest eigenvalue of the covariance matrix

Maximizing the projected inertia, recursion



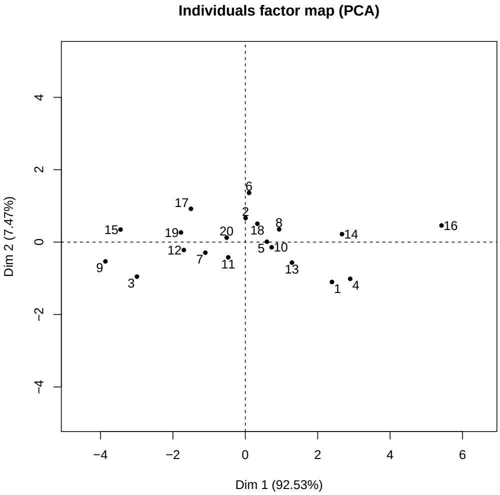
The second largest eigenvalue maximizes the projected inertia in the orthogonal of the first

Maximizing the projected inertia, summary



Projected points on the first two ‘principal components’

Maximizing the projected inertia, summary



Representation with package FactoMineR. Percentages are inertia ratio w.r.t. total inertia

Theory

Notations and assumption

- $\mathbf{X}$: a matrix of size $n \times p$, representing the data:

	$\mathbf{x}^1$	$\dots$	$\mathbf{x}^j$	$\dots$	$\mathbf{x}^p$
$\mathbf{x}_1$	x_1^1	$\dots$	x_1^j	$\dots$	x_1^p
$\vdots$	$\vdots$		$\vdots$		$\vdots$
$\mathbf{x}_i$	x_i^1	$\dots$	x_i^j	$\dots$	x_i^p
$\vdots$	$\vdots$		$\vdots$		$\vdots$
$\mathbf{x}_n$	x_n^1	$\dots$	x_n^j	$\dots$	x_n^p

- $\mathbf{g}$: center of gravity (empirical mean), $\mathbf{g} = \bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \ (\in \mathbb{R}^p)$.

$\mathbf{g}$	$\overline{\mathbf{x}^1}$	$\dots$	$\overline{\mathbf{x}^j}$	$\dots$	$\overline{\mathbf{x}^p}$

We assume that $\mathbf{g} = 0$, i.e. the data have been centered.

Notations and assumption

- The rows of $\mathbf{X}$ lie in $\mathbb{R}^p$, and form the **indivuals space**.
  It is an Euclidean space, equipped with the usual ℓ^2 norm $\|.\|$.

- The columns of $\mathbf{X}$ lie in $\mathbb{R}^n$, and form the **variables space**.
  It is an Euclidean space. Instead of choosing the usual ℓ^2 norm, we rescale it by $1/n$. Indeed, as the data are centered, it corresponds to the empirical covariance:

$$\langle \mathbf{x}^j, \mathbf{x}^k \rangle_{\mathbb{R}^n} := \frac{1}{n} \sum_{i=1}^{n} x_i^j x_i^k = \widehat{\mathrm{cov}}(\mathbf{x}^j, \mathbf{x}^k).$$

  Notice that **orthogonal variables = uncorrelated variables**.
  Γ denotes the $p \times p$ empirical covariance matrix:

$$\Gamma = \left(\widehat{\mathrm{cov}}(\mathbf{x}^j, \mathbf{x}^k) \right)_{1 \le j,k \le p} = \frac{1}{n} \mathbf{X}^\top \mathbf{X} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top.$$

Notations and assumption

- **Inertia**: mean squared distance of the data to their center (here $\mathbf{0}$),

$$\mathcal{I} = \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{x}_i\|^2$$

- **Projected inertia** on a subspace $F \subseteq \mathbb{R}^p$. Same definition for the projected points onto F (we denote by Π_F the projection operator):

$$\mathcal{I}_F = \frac{1}{n}\sum_{i=1}^{n} \|\Pi_F(\mathbf{x}_i)\|^2$$

Properties of inertia

Link with variance, and inertia decomposition.

Consider a $1D$ axis spanned by a unit vector $\mathbf{a}$, and denote $\mathcal{I}_{\mathbf{a}} = \mathcal{I}_{\mathbb{R}\mathbf{a}}$. Then:

$$\mathcal{I}_{\mathbf{a}} = \mathbf{a}^\top \Gamma \mathbf{a}, \qquad \text{and} \qquad \mathcal{I} = \mathcal{I}_{\mathbf{a}} + \mathcal{I}_{\mathbf{a}^\perp}$$

Moreover, $\mathcal{I}_{\mathbf{a}}$ and $\mathcal{I}$ are interpreted in terms of variances:

- $\mathcal{I}_{\mathbf{a}}$ is the empirical variance of the projected points onto $\mathbb{R}\mathbf{a}$:
- $\mathcal{I}$ is the sum of the empirical variances of the p variables.

$$\mathcal{I}_{\mathbf{a}} = \frac{1}{n}\sum_{i=1}^{n} \langle \mathbf{x}_i, \mathbf{a} \rangle^2, \qquad \mathcal{I} = \sum_{j=1}^{p} \hat{\sigma}_j^2, \quad \text{with} \quad \hat{\sigma}_j^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i^j)^2$$

Remark: The empirical variances are computed here by dividing by n the sum of squares, contrarily to unbiased statistical estimates, which divide by $n-1$.

Properties of inertia (proofs)

- By definition, $\Pi_{\mathbf{a}}(\mathbf{x}_i) = \langle \mathbf{x}_i, \mathbf{a} \rangle = \mathbf{a}^\top \mathbf{x}_i = \mathbf{x}_i^\top \mathbf{a}$. Thus:

$$\mathcal{I}_{\mathbf{a}} = \frac{1}{n}\sum_{i=1}^{n} \langle \mathbf{x}_i, \mathbf{a} \rangle^2 = \frac{1}{n}\sum_{i=1}^{n} (\mathbf{a}^\top \mathbf{x}_i)(\mathbf{x}_i^\top \mathbf{a}) = \mathbf{a}^\top \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top \right) \mathbf{a} = \mathbf{a}^\top \Gamma \mathbf{a}.$$

- The inertia decomposition follows from Pythagore's theorem:

$$\|\mathbf{x}_i\|^2 = \|\Pi_F(\mathbf{x}_i)\|^2 + \|\Pi_{F^\perp}(\mathbf{x}_i)\|^2.$$

- $\mathcal{I}_{\mathbf{a}}$ is equal to the mean square of the real numbers $\langle \mathbf{x}_i, \mathbf{a} \rangle$ $(i = 1, \dots, n)$. This is the empirical variance, as they are centered: $\overline{\langle \mathbf{x}_., \mathbf{a} \rangle} = \langle \overline{\mathbf{x}_.}, \mathbf{a} \rangle = 0$.
- Finally, when $\mathbf{a} = \mathbf{e}_j$, the j$^{\text{th}}$ first vector of the canonical basis of $\mathbb{R}^p$,

$$\mathcal{I}_{\mathbf{e}_j} = \frac{1}{n}\sum_{i=1}^{n} \langle \mathbf{x}_i, \mathbf{e}_j \rangle^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i^j)^2 = \hat{\sigma}_j^2. \qquad \text{(again the } \langle \mathbf{x}_., \mathbf{e}_j \rangle \text{ are centered)}$$

Thus, by Pythagore's theorem,

$$\mathcal{I} = \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{x}_i\|^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p} \langle \mathbf{x}_i, \mathbf{e}_j \rangle^2 = \sum_{j=1}^{p} \mathcal{I}_{\mathbf{e}_j} = \sum_{j=1}^{p} \hat{\sigma}_j^2.$$

Main result

Theorem (principal component analysis)

As the covariance matrix Γ is real symmetric, it admits a spectral decomposition in orthogonal eigenspaces. Denote $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$ the eigenvalues, and $\mathbf{v}_1, \dots, \mathbf{v}_p$ orthogonal eigenvectors. Then:

- $\mathbf{v}_1$ maximizes $\mathcal{I}_{\mathbf{a}}$ over $\mathbf{a}$, which is then equal to λ_1.
- $\mathbf{v}_2$ maximizes $\mathcal{I}_{\mathbf{a}}$ over $\mathbf{a}$ in $(\mathbf{v}_1)^\perp$, which is then equal to λ_2.
- $\mathbf{v}_3$ maximizes $\mathcal{I}_{\mathbf{a}}$ over $\mathbf{a}$ in $(\mathbf{v}_1, \mathbf{v}_2)^\perp$, which is then equal to λ_3.
- ...

Furthermore the inertia (called *total inertia*) is decomposed:

$$\mathcal{I} = \mathcal{I}_{\mathbf{v}_1} + \cdots + \mathcal{I}_{\mathbf{v}_p} = \lambda_1 + \cdots + \lambda_p$$

Main result (proof)

Actually, this is exactly how the spectral decomposition of Γ is obtained. Let us check that it works. Without loss of generality, assume that $\mathbf{a}, \mathbf{v}_1, \dots, \mathbf{v}_p$ are vectors of norm 1. Let us decompose $\mathbf{a}$ in the basis of eigenvectors:

$$\mathbf{a} = a_1\mathbf{v}_1 + \cdots + a_p\mathbf{v}_p.$$

By properties of eigenvectors, $\mathbf{v}_j^\top \Gamma \mathbf{v}_k = \lambda_k \mathbf{v}_j^\top \mathbf{v}_k = \lambda_k \delta_{j,k}$. Now:

$$\mathcal{I}_\mathbf{a} = \mathbf{a}^\top \Gamma \mathbf{a} = \sum_{j,k=1}^{p} a_j a_k \mathbf{v}_j^\top \Gamma \mathbf{v}_k = \sum_{k=1}^{p} \lambda_k a_k^2 \leq \lambda_1 \|\mathbf{a}\|^2 = \lambda_1.$$

The inequality above is an equality when $\mathbf{a} = \mathbf{v}_1$.
Similarly, if $\mathbf{a}$ belongs to $\mathbf{v}_1^\perp$, then $a_1 = 0$. Hence,

$$\mathcal{I}_\mathbf{a} = \sum_{k=2}^{p} \lambda_k a_k^2 \leq \lambda_2 \|\mathbf{a}\|^2 = \lambda_2$$

with equality if $\mathbf{a} = \mathbf{v}_2$. And so on.

Finally, the inertia decomposition gives the formula for $\mathcal{I}$.

Principal components

- The eigenvectors $\mathbf{v}_1, \dots, \mathbf{v}_p$ define a new orthonormal basis in $\mathbb{R}^p$.
- The change of variables is defined by:

$$\mathbf{C} = \mathbf{XP}, \qquad \text{with} \quad \mathbf{P} = [\mathbf{v_1}, \dots, \mathbf{v_p}] \, .$$

  The $n \times p$ matrix $\mathbf{C}$ is called **matrix of principal components**. The columns of $\mathbf{C}$ are called **principal variables**. They contain the coordinates of the individuals in the new space.

- Principal variables are centered, uncorrelated and $\widehat{\mathrm{var}}(\mathbf{C}^k) = \lambda_k$:

$$\left(\widehat{\mathrm{cov}}(\mathbf{C}^j, \mathbf{C}^k)\right)_{1 \leq j,k \leq p} = \frac{1}{n}\mathbf{C}^\top \mathbf{C} = \mathbf{P}^\top \Gamma \mathbf{P} = \mathrm{diag}(\lambda_1, \dots, \lambda_p).$$

Remark: singular value / spectral decomposition

PCA can be done with **Singular Value Decomposition (SVD)**, which decomposes a rectangular matrix $n \times m$ or rank r as

$$\mathbf{X} = \mathbf{U}\Lambda^{1/2}\mathbf{V}^\top,$$

where Λ is the diagonal matrix containing the r non-zero eigenvalues of $\mathbf{X}^\top\mathbf{X}$ (or $\mathbf{X}\mathbf{X}^\top$), ranked by decreasing order, and $\mathbf{U}$ (resp. $\mathbf{V}$) is an orthogonal matrix for $\|.\|_{\mathbb{R}^n}$ (resp. for $\|.\|_{\mathbb{R}^m}$) containing the eigenvectors of $\mathbf{X}\mathbf{X}^\top$ (resp. $\mathbf{X}^\top\mathbf{X}$).

In the frequent case when $p = r$ (e.g. $n > p$), we have:

$$\mathbf{V} = \mathbf{P}, \qquad \Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_n).$$

(In the general case, $\mathbf{V}$ contains the r columns of $\mathbf{P}$ corresponding to non-zero eigenvalues.) Further, due to our definition of the scalar product in $\mathbb{R}^n$, we have $\frac{1}{n}\mathbf{U}^\top\mathbf{U} = I_p$. Then, you can recover all the formulas of the textbook, e.g.:

$$\mathbf{C} = \mathbf{X}\mathbf{P} = \mathbf{U}\Lambda^{1/2}\mathbf{P}^\top\mathbf{P} = \mathbf{U}\Lambda^{1/2}.$$

Variations (metric, weights)

Changing the metric in the individuals space

Consider a new norm on $\mathbb{R}^p$, called **metric**, defined by a positive definite matrix $\mathbf{M}$, of size p:

$$\|\mathbf{x}\|_M^2 = \mathbf{x}^\top \mathbf{M}\mathbf{x}.$$

Let $\mathbf{R}$ be an invertible matrix s.t. $\mathbf{R}^\top \mathbf{R} = \mathbf{M}$ (e.g. square root, Choleski decomposition). Then, the map

$$\mathbf{R} : \begin{array}{ccc} (\mathbb{R}^p, \|.\|_M) & \rightarrow & (\mathbb{R}^p, \|.\|) \\ \mathbf{x} & \mapsto & \mathbf{R}\mathbf{x} \end{array}$$

is an isometry, and thus preserves distances and orthogonality.

Indeed: $\|\mathbf{R}\mathbf{x}\|^2 = (\mathbf{R}\mathbf{x})^\top(\mathbf{R}\mathbf{x}) = \mathbf{x}^\top\mathbf{M}\mathbf{x} = \|\mathbf{x}\|_M^2$.

Changing the metric in the individuals space

Due to the isometry property, we deduce immediately:

PCA with / without metric

PCA on original data $\mathbf{x}_1, \dots, \mathbf{x}_n$ with metric $\|.\|_M$

$$\Leftrightarrow$$

PCA on transformed data $\mathbf{R}\mathbf{x}_1, \dots, \mathbf{R}\mathbf{x}_n$ with $\|.\|$

$$\Leftrightarrow$$

Spectral decomposition of $\Gamma = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{R}\mathbf{x}_i)(\mathbf{R}\mathbf{x}_i)^\top = \mathbf{R}\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)\mathbf{R}^\top$

$$\Leftrightarrow$$

Spectral decomposition of $\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)\mathbf{M}$

Changing the metric in the individuals space

Recall that the data are assumed to be centered.

Example. Standardize (centered) data.

$$\mathbf{M} = \mathrm{diag}\left(\frac{1}{\hat{\sigma}_1^2}, \dots, \frac{1}{\hat{\sigma}_p^2}\right)$$

Then we can choose $\mathbf{R} = \mathrm{diag}\left(\frac{1}{\hat{\sigma}_1}, \dots, \frac{1}{\hat{\sigma}_p}\right)$. Thus doing PCA with the metric **M** is equivalent to doing usual PCA on the standardized data.

Changing the weights in the variable space

In the standard formulation, each individual $\mathbf{x}_1, \dots, \mathbf{x}_n$ has weight $\frac{1}{n}$.

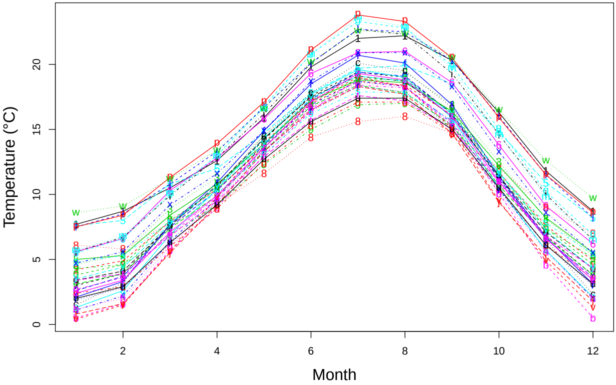
Obviously, one can use positive weights $\omega_1, \dots, \omega_n$ that sum to one. It can be useful if some individuals have more importance.

This can be viewed as an isometric transformation in the space $\mathbb{R}^n$ by the diagonal matrix containing the square roots of ω_i.
The theory is immediately adapted, by modifying the definitions, e.g.:

$$\mathcal{I} = \sum_{i=1}^{n} \omega_i \|\mathbf{x}_i\|^2, \qquad \Gamma = \sum_{i=1}^{n} \omega_i \mathbf{x}_i \mathbf{x}_i^\top.$$

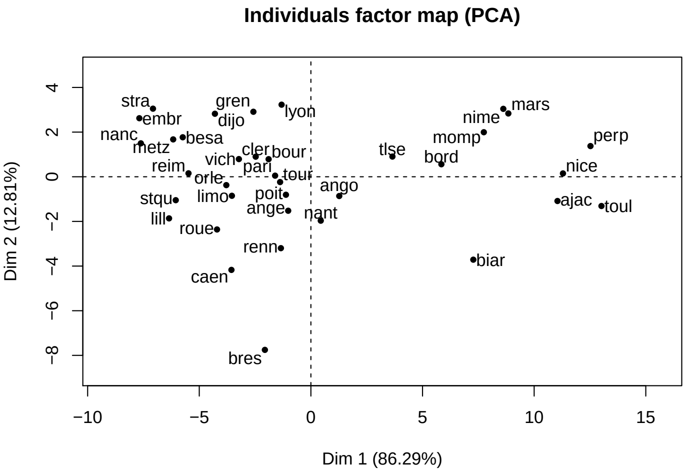
Results interpretation

Example on a temperature dataset



Dataset: Temperature at $n = 36$ cities (individuals) for $p = 12$ months (variables).

Graphics for individuals



PCA: Projection on the first 2 principal axis. They explain more than 95% of the total inertia. Thus, the 12-dimensional data can be well approximated in 2-dimensions only.

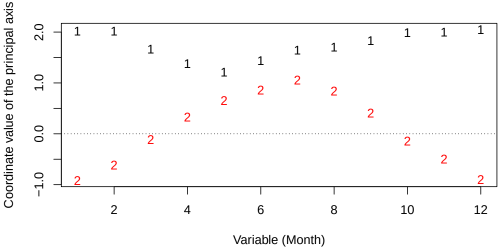
Interpretation of principal components

- Remember that the principal variables $\mathbf{C}^1, \ldots, \mathbf{C}^{12}$ are linear combinations of the original ones (here: the months).
- To get an intuition about their meaning, look at the individuals located at the extremes on each axis.
- Very often, for unscaled data, axis 1 represents a global amount, the other ones contrasts (differences) between variables. Here:
  - Axis 1 ranges cities according to their annual temperature
  - Axis 2 ranges cities according to the contrast summer/winter

Interpretation of principal components

Let us check this by looking at the coordinates of $\mathbf{C}_1, \mathbf{C}_2$ in $\mathbb{R}^{12}$.
Here we can plot them. This confirm our guess:

- $\mathbf{C}_1 \approx 2(x^1 + \cdots + x^{12})$, proportional to the annual temperature
- $\mathbf{C}_2 \approx (x^5 + ... + x^8) - (x^1 + x^2 + x^{11} + x^{12})$, contrast summer/winter



Coordinates of the first 2 principal axis in the 12-dimensional space of individuals.

Graphics for variables

- The components variables $\mathbf{C}^k$ are orthogonal with variance λ_k. Thus, they define an orthonormal basis $\tilde{\mathbf{C}}_k = \mathbf{C}^k / \sqrt{\lambda_k}$.
- Consider the coordinates $a_{j,k}$ of the original variables in this basis

$$a_{j,k} = \mathrm{cov}(\mathbf{X}^j, \tilde{\mathbf{C}}_k).$$

  We thus have, $\|\mathbf{x}^j\|^2_{\mathbb{R}^n} = \hat{\sigma}_j^2 = \sum_k a_{j,k}^2$.
- The idea is to plot these coordinates for two principal components.

Graphics for variables, case of unit variance

- When the variables have been normalized (unit variance),

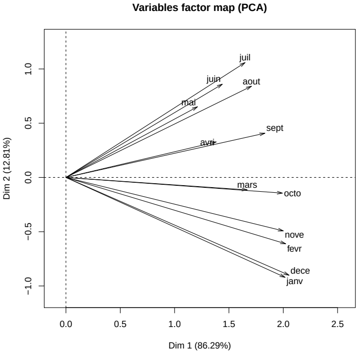
$$a_{j,k} = \mathrm{cor}(\mathbf{X}^j, \tilde{\mathbf{C}}_k) = \cos(\widehat{\mathbf{X}^j, \tilde{\mathbf{C}}_k})$$

  and $\sum_{k=1}^{p} a_{j,k}^2 = 1$.

- Thus the coordinates $(a_{j,k})_k$ belong to a p-dimensional sphere.

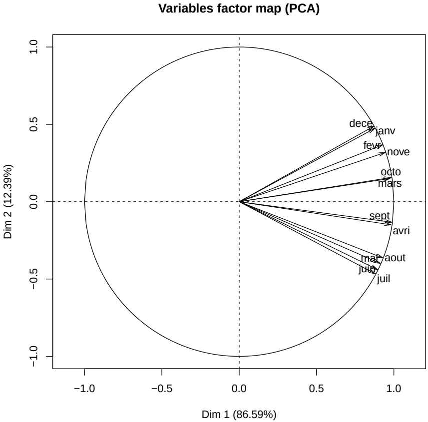
- Further $(a_{j,1}, a_{j,2})$ belongs to the unit disk: $a_{j,1}^2 + a_{j,2}^2 \leq 1$.
  It is closed to the unit circle if $a_{j,3}, \dots, a_{j,p}$ are nearly zero.
  In that case, $\mathbf{X}^j$ is well-represented by $\mathbf{C}^1, \mathbf{C}^2$.
  This is the **circle of correlations** for components (1, 2).

Interpretation of principal components



Coordinates of the variables in the orthonormal basis of component variables. We see again that Axis 1 weigths all months nearly equally, whereas Axis 2 exhibits a contrast summer / winter.

Interpretation of principal components



Circle of correlation (normalized variables). Here all variables are well-represented by the first 2 principal components.

Conclusion and further readings

- PCA is a **dimension reduction technique** which finds **uncorrelated variables, called component variables**, that are linear combination of the original ones, which approximate the best the data in the **mean-square** sense.
- **PCA = spectral decomposition of the covariance matrix**
  - Up to isometric transformations (metric, weights)
- Several graphs can be used to interpret principal components: projection of individuals, circle of correlation (normalized case).
  - Mind that **what you visualize is only a projection.** Several tools quantify the **quality of the representation.**
    $\rightarrow$ See textbook page 29, 30.