ADVANCED BIOSTATICS
ABSTAT18
Principal Component Analysis

Carina Silva
(*carina.silva@estesl.ipl.pt*)
Higher School of Technologies and Health of Lisbon &
Center of Statistics and Applications, University of Lisbon | CEAUL

ESCOLA SUPERIOR DE TECNOLOGIA DA SAÚDE DE LISBOA
INSTITUTO POLITÉCNICO DE LISBOA

CEAUL
Centro de Estatística e Aplicações
Universidade de Lisboa

IGC, April 3rd - 6th, 2018

Introduction

- The tools used for exploratory analyses are mainly multivariate methods, i.e. statistical methods dealing with many variables at the same time.

Introduction

- The tools used for exploratory analyses are mainly multivariate methods, i.e. statistical methods dealing with many variables at the same time.
- There are three basic tools of exploratory analysis (= exploratory statistics): **Principal Component Analysis (PCA)**; Correspondence Analysis (CA or COA) and Multi-Dimensional Scaling (MDS).

Introduction

- The tools used for exploratory analyses are mainly multivariate methods, i.e. statistical methods dealing with many variables at the same time.
- There are three basic tools of exploratory analysis (= exploratory statistics): **Principal Component Analysis (PCA)**; Correspondence Analysis (CA or COA) and Multi-Dimensional Scaling (MDS).
- These three methods are based on the same fundamental principles. Their role is to visually summarize all analyzed variables, and to reveal their inter-relationships.

Principal Component Analysis

The principles are:

- variables are expressed geometrically as vectors,

Principal Component Analysis

The principles are:

- variables are expressed geometrically as vectors,
- correlations as angles between vectors,

Principal Component Analysis

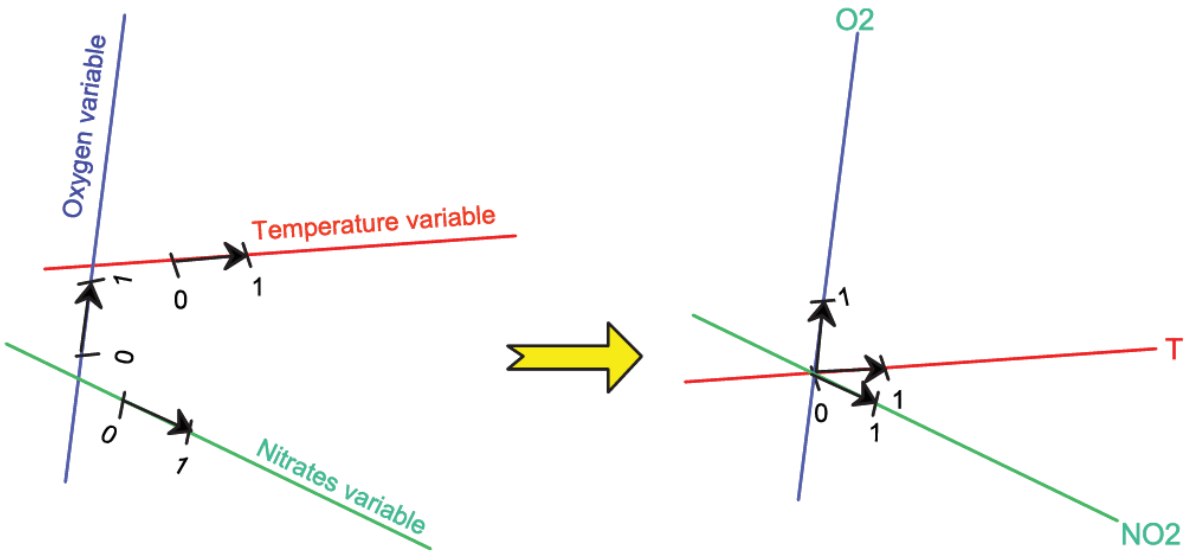
The principles are:

- variables are expressed geometrically as vectors,
- correlations as angles between vectors,
- analyses are seen as projections of these vectors onto planes.

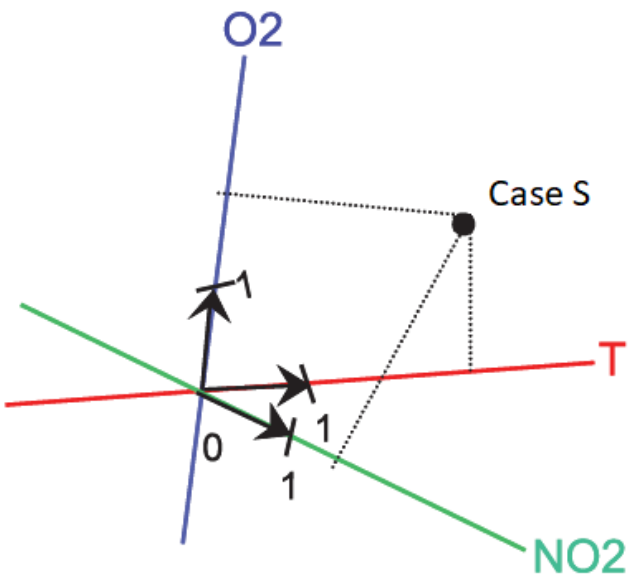
Variables as dimensions

A variable is represented by a straight line (i.e. a vector), having an origin (point 0), and a direction (positive values or negative values).

- If we consider 3 variables, they correspond to 3 distinct lines, hence defining 3 dimensions, i.e. a space:

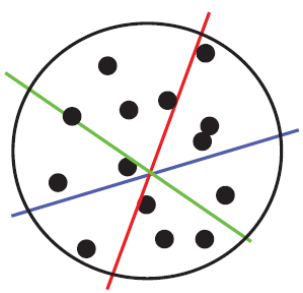


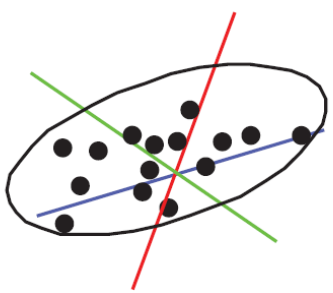
- Let's consider water quality sampling in a river. In one given location, temperature, oxygen rate and nitrates concentration are measured; this corresponds to one data point (i.e. one case, on site S at time t) in which 3 variables are expressed. This point is located somewhere in the space created by the 3 variables, according to the values taken for each variable:



Variables as dimensions

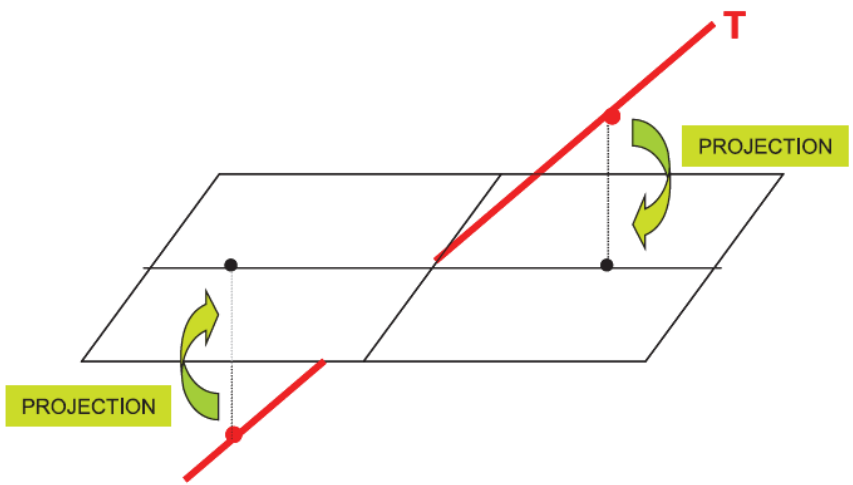
- If we have for example 15 locations (cases) to study 3 variables, then the 15 points constitute a cluster in a space of dimension 3.
- This cluster can have a relatively spherical shape if data points are more or less distributed in the same way for each variable
- On the contrary the cluster will have an elongated shape if the points are more scattered on one variable than on the others.



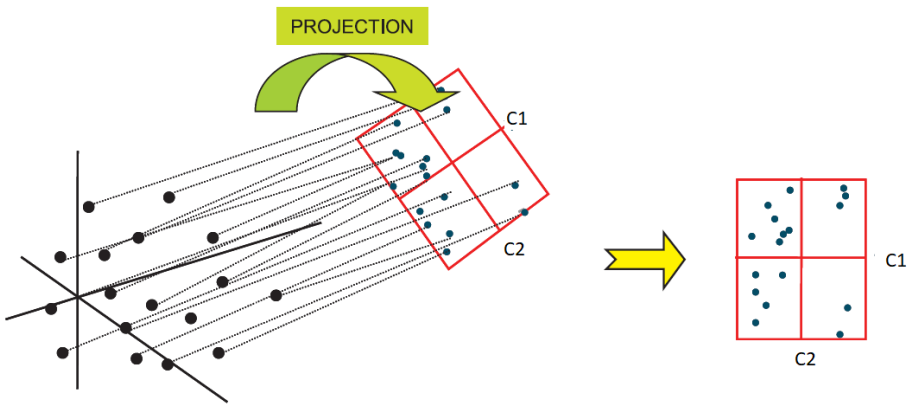


Projection onto a PC map

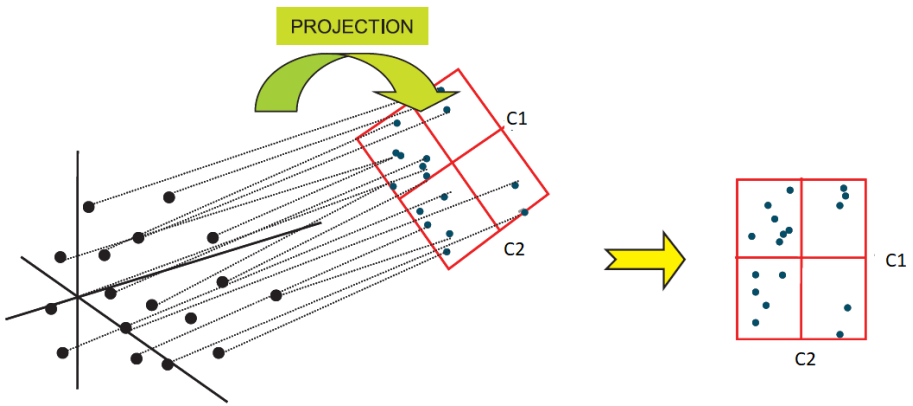
The purpose of the exploratory analysis of data is to summarize at best all the information expressed in a hyperspace (i.e. all the measurement points that it contains), and to project it on a plane in dimension 2.



A constraint is that the cluster should be as dispersed as possible, which in arithmetic terms corresponds to a maximal variance obtained by a calculation called ("matrix diagonalization").

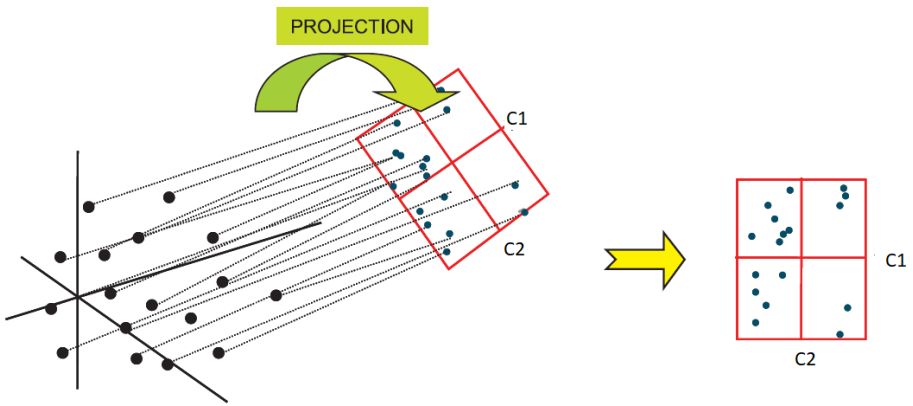


A constraint is that the cluster should be as dispersed as possible, which in arithmetic terms corresponds to a maximal variance obtained by a calculation called ("matrix diagonalization").



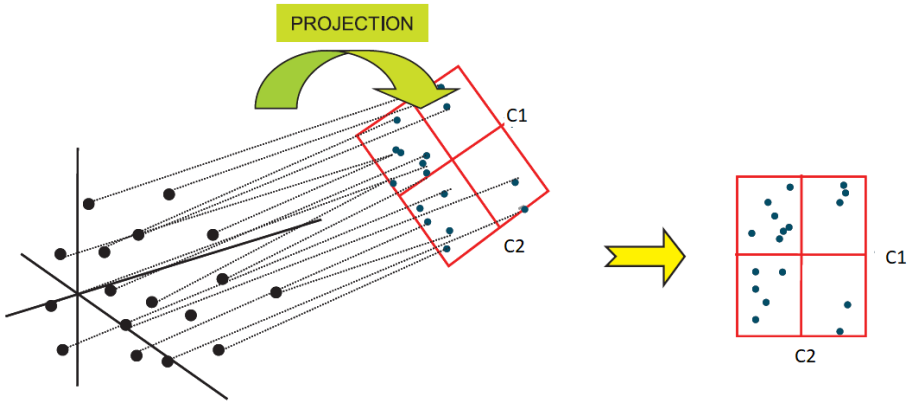
- A PC map is made of PC axes, named PCi (i.e. PC1, PC2, PC3, etc.), the first of which is horizontal by convention.

A constraint is that the cluster should be as dispersed as possible, which in arithmetic terms corresponds to a maximal variance obtained by a calculation called ("matrix diagonalization").



- A PC map is made of PC axes, named PCi (i.e. PC1, PC2, PC3, etc.), the first of which is horizontal by convention.
- PC axes have no unit or fixed direction.

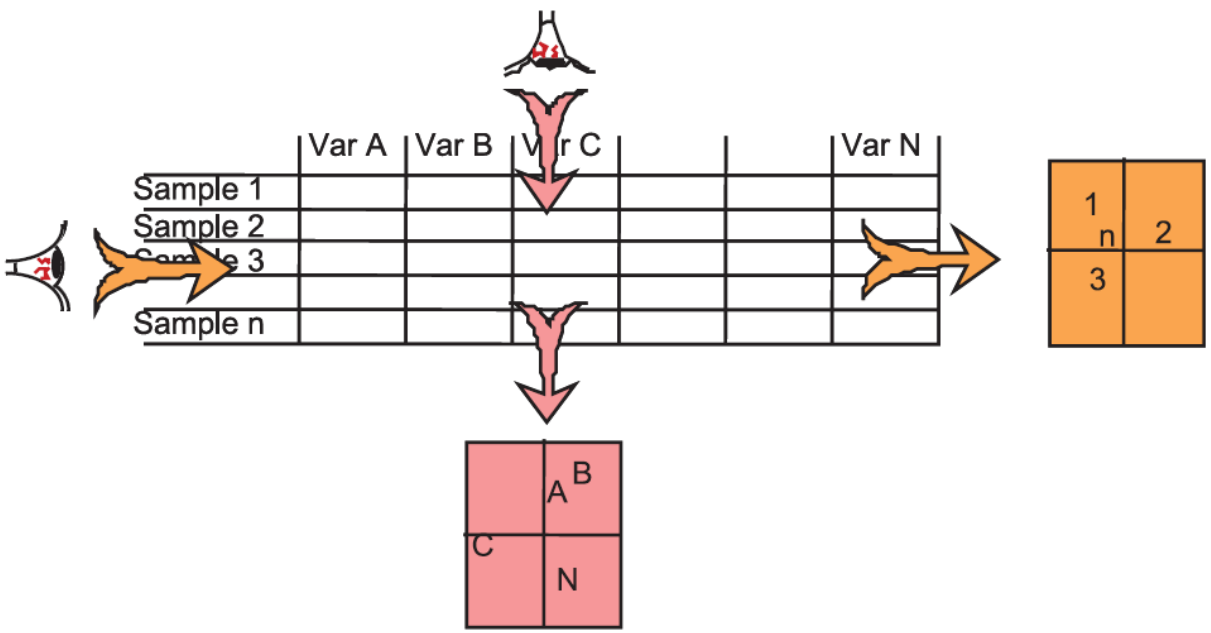
A constraint is that the cluster should be as dispersed as possible, which in arithmetic terms corresponds to a maximal variance obtained by a calculation called ("matrix diagonalization").



- A PC map is made of PC axes, named PCi (i.e. PC1, PC2, PC3, etc.), the first of which is horizontal by convention.
- PC axes have no unit or fixed direction.
- The PC map should not be read as a biplot with graduated axes, but just as a road map in which the proximity of two points expresses a certain correlation.

Projection of variables or of cases

- Data points express both variables and cases.
- Multivariate data analysis can be seen as the analysis of Variables OR of Cases, and the PC map can be that of the projection of Cases or that of the projection of Variables.



Data Organization

Multivariate data are a collection of observations (or measurements) of:

- p variables ($k = 1, ..., p$)
- n cases ($j = 1, ..., n$)

Data Organization

- x_{jk} measurement of the k^{th} variable on the j^{th} case.

	Variable 1	Variable 2	...	Variable k	...	Variable p
1:	x_{11}	x_{12}	...	x_{1k}	...	x_{1p}
2:	x_{21}	x_{22}	...	x_{2k}	...	x_{2p}
	$\vdots$	$\vdots$		$\vdots$		$\vdots$
j:	x_{j1}	x_{j2}	...	x_{jk}	...	x_{jp}
	$\vdots$	$\vdots$		$\vdots$		$\vdots$
n:	x_{n1}	x_{n2}	...	x_{nk}	...	x_{np}

- The first subscript (j) represents the ROW location in the data array.
- The second subscript (k) represents the COLUMN location in the data array.

Descriptive Statistics Review

- When we have a large amount of data, it is often hard to get a manageable description of the nature of the variables under study.
- Such descriptive statistics include:
  - Means
  - Variances
  - Covariances
  - Correlations

Sample Mean

- For the k^{th} variable, the sample mean is: $\bar{x}_k = \frac{1}{n}\sum_{j=1}^{n} x_{jk}$
- An array of the means for all p variables then looks like this (which we will come to know as the mean vector):

$$\bar{\mathbf{x}} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \bar{x}_3 \\ \bar{x}_4 \end{bmatrix}$$

Sample Variance

- For the k^{th} variable, the sample variance is:
  $s_k^2 = s_{kk} = \frac{1}{n}\sum_{j=1}^{n}(x_{jk} - \bar{x}_k)^2$
- Note the "kk" subscript, this will be important because the equation that produces the variance for a single variable is a derivation of the equation of the covariance for a pair of variables.
- For a pair of variables, i and k, the sample covariance is:
  $s_{ik} = \frac{1}{n}\sum_{j=1}^{n}(x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k)$

Sample Covariance Matrix

- Making a matrix of all sample covariances give us:

$$\mathbf{S}_n = \begin{bmatrix} s_{11} & s_{12} & \dots & s_{1p} \\ s_{21} & s_{22} & \dots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \dots & s_{pp} \end{bmatrix}$$

Sample Correlation

- Sample covariances are dependent upon the scale of the variables under study.
- For this reason, the correlation is often used to describe the association between two variables.
- For a pair of variables, i and k, the sample correlation is found by dividing the sample covariance by the product of the standard deviation of the variables:
  $r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}}$
- The sample correlation:
  - Ranges from -1 to 1.
  - Measures linear association.
  - Is invariant under linear transformations of i and k.

Sample Correlation Matrix

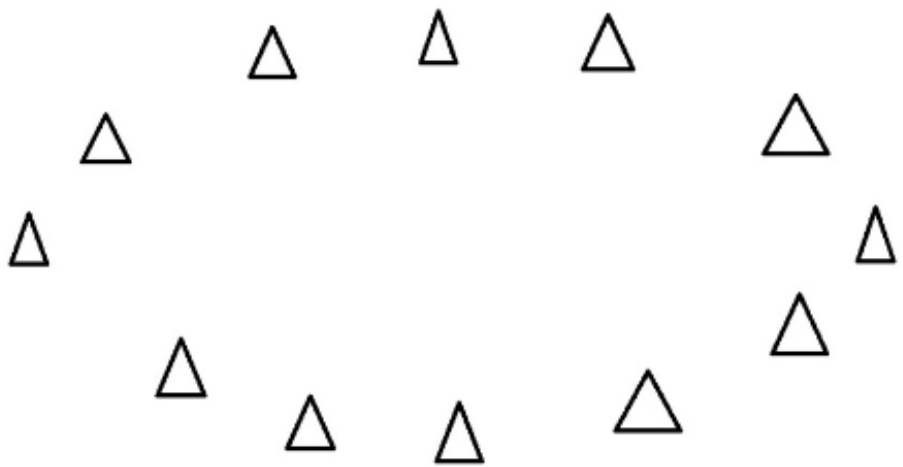
- Making a matrix of all sample correlations give us:

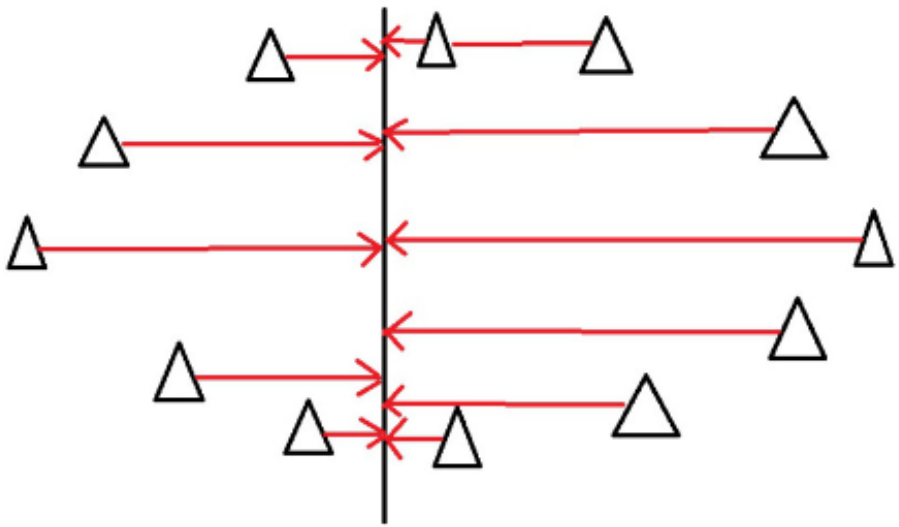
$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \dots & 1 \end{bmatrix}$$

What are principal components?

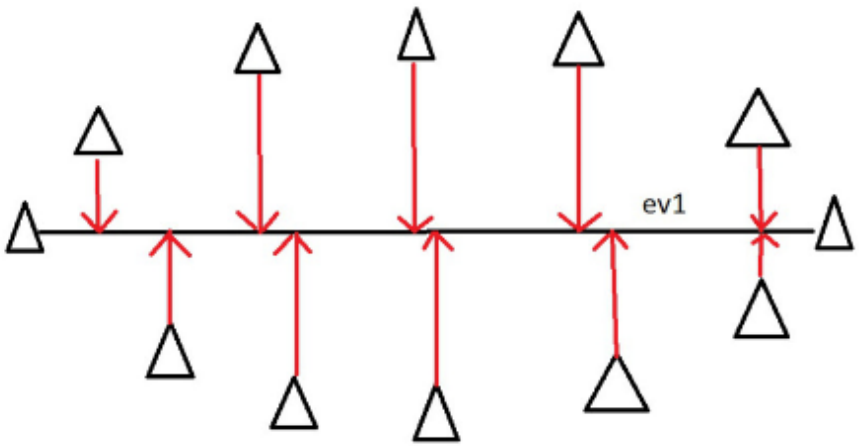
- They're the underlying structure in the data.
- They are the directions where there is the most variance, the directions where the data is most spread out

- Imagine that the triangles are points of data. To find the direction where there is most variance, find the straight line where the data is most spread out when projected onto it.





- The data isn't very spread out here, therefore it doesn't have a large variance. It is probably not the principal component.



- On this line the data is way more spread out, it has a large variance. In fact there isn't a straight line you can draw that has a larger variance than a horizontal one. A horizontal line is therefore the principal component in this example.

The principal component axes that constitutes the PC map are calculated as follows

- the first axis, named C1, goes through the cluster in such a way that the variance, i.e. the distribution of data points on this axis, is maximal;
- the second axis, named C2, must be orthogonal to C1 and goes through the data cluster in such a way that the variance is again maximal;
- the third axis, named C3, must be orthogonal to the two first axes and goes through the data cluster in such a way that the variance is again maximal;
- and so on for C4, C5, etc.

Frequently Ask Questions

- Why maximize the variance on each axis?

Frequently Ask Questions

- Why maximize the variance on each axis?
  - In order to have the widest possible scattering of the data points, i.e. to clarify the information and make it as readable as possible.

Frequently Ask Questions

- Why maximize the variance on each axis?
  - In order to have the widest possible scattering of the data points, i.e. to clarify the information and make it as readable as possible.
- How can the variance be maximized?

Frequently Ask Questions

- Why maximize the variance on each axis?
  - In order to have the widest possible scattering of the data points, i.e. to clarify the information and make it as readable as possible.
- How can the variance be maximized?
  - By creating a principal component axis that goes through the most elongated direction of the cluster

Frequently Ask Questions

- Why maximize the variance on each axis?
  - In order to have the widest possible scattering of the data points, i.e. to clarify the information and make it as readable as possible.
- How can the variance be maximized?
  - By creating a principal component axis that goes through the most elongated direction of the cluster
- Why choose axes successively orthogonal to one another?

Frequently Ask Questions

- Why maximize the variance on each axis?
  - In order to have the widest possible scattering of the data points, i.e. to clarify the information and make it as readable as possible.
- How can the variance be maximized?
  - By creating a principal component axis that goes through the most elongated direction of the cluster
- Why choose axes successively orthogonal to one another?
  - So that the information on each axis is as much as possible independent from the previous axes.

Maths in PCA

Let's say we have a random vector $X_1, X_2, ..., X_p$.

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix}$$

with population variance-covariance matrix:

$$var(\mathbf{X}) = \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_p \end{pmatrix}$$

Variance and Covariance

To understand how PCA works, we need to recall the concepts of variance and covariance.

- Variance of a sample is given by $s^2 = \frac{\sum(x_i - \bar{x})}{n-1}$
- Covariance between two variables, x and y, as $cov(x, y) = \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1}$,
  - when x varies with y, this expression will tend do accumulate positive terms;
  - when they are independent, the covariance will be zero;
  - and will be negative if they are anti-correlated.

Note that the variance of a variable is just the covariance of that variable with itself.

Principal Components will be linear combinations of the original variables:

$$\xi_1 = \phi_{11}X_1 + \phi_{12}X_2 + \ldots + \phi_{1p}X_p$$
$$\xi_2 = \phi_{21}X_1 + \phi_{22}X_2 + \ldots + \phi_{2p}X_p$$
$$\vdots$$
$$\xi_p = \phi_{p1}X_1 + \phi_{p2}X_2 + \ldots + \phi_{pp}X_p$$

where:

- ξ_i are the principal components, where $cov(\xi_i, \xi_j) = 0$, i.e., he principal components are uncorrelated with one another.
- ϕ_{ij} are the coefficients, where they are collected into vectors (eigenvectors):

$$\phi_i = \begin{pmatrix} \phi_{i1} \\ \phi_{i2} \\ \vdots \\ \phi_{ip} \end{pmatrix}$$

- The eigenvectors are normal vectors: $|\phi_i| = 1$

How do we find the coefficients ϕ_{ij} for a principal component?

- The solution involves the eigenvalues and eigenvectors of the new variance-covariance matrix Λ.

How do we find the coefficients ϕ_{ij} for a principal component?

- The solution involves the eigenvalues and eigenvectors of the new variance-covariance matrix Λ.
- The variance for the ith principal component is equal to the ith eigenvalue:

$$var(\xi_i) = var(\phi_{i1}X_1 + \phi_{i2}X_2 + \ldots + \phi_{ip}X_p) = \lambda_i$$

How do we find the coefficients ϕ_{ij} for a principal component?

- The solution involves the eigenvalues and eigenvectors of the new variance-covariance matrix Λ.
- The variance for the ith principal component is equal to the ith eigenvalue:

$$var(\xi_i) = var(\phi_{i1}X_1 + \phi_{i2}X_2 + \ldots + \phi_{ip}X_p) = \lambda_i$$

- Where the eigenvalues are ordered so that λ_1 is the largest value and λ_p the smallest: $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$

How do we find the coefficients ϕ_{ij} for a principal component?

- The solution involves the eigenvalues and eigenvectors of the new variance-covariance matrix Λ.
- The variance for the *i*th principal component is equal to the *i*th eigenvalue:

$$var(\xi_i) = var(\phi_{i1}X_1 + \phi_{i2}X_2 + \ldots + \phi_{ip}X_p) = \lambda_i$$

- Where the eigenvalues are ordered so that λ_1 is the largest value and λ_p the smallest: $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$
- $trace(\Sigma) = \sigma_1^2 + \sigma_2^2 + ... + \sigma_p^2 = \lambda_1 + \lambda_2 + ... + \lambda_p = trace(\Lambda)$
  This will give us an interpretation of the components in terms of the amount of the full variation explained by each component.

- The data covariance matrix Σ ($p \times p$ matrix) is transformed into a diagonal matrix Λ ($p \times p$ matrix), whose diagonal elements λ_{ii} are the variances of the new variables $\xi = (\xi_1, ..., \xi_p)$.

- The data covariance matrix Σ ($p \times p$ matrix) is transformed into a diagonal matrix Λ ($p \times p$ matrix), whose diagonal elements λ_{ii} are the variances of the new variables $\xi = (\xi_1, ..., \xi_p)$.
- This diagonal matrix is constructed such that the diagonal elements are in descending order.

- The data covariance matrix Σ ($p \times p$ matrix) is transformed into a diagonal matrix Λ ($p \times p$ matrix), whose diagonal elements λ_{ii} are the variances of the new variables $\xi = (\xi_1, ..., \xi_p)$.
- This diagonal matrix is constructed such that the diagonal elements are in descending order.
- Hence $\lambda_{11} (= \lambda_1)$ represents the largest amount of variation in the data and λ_{pp} the least amount.

- The data covariance matrix Σ ($p \times p$ matrix) is transformed into a diagonal matrix Λ ($p \times p$ matrix), whose diagonal elements λ_{ii} are the variances of the new variables $\xi = (\xi_1, ..., \xi_p)$.
- This diagonal matrix is constructed such that the diagonal elements are in descending order.
- Hence $\lambda_{11} (= \lambda_1)$ represents the largest amount of variation in the data and λ_{pp} the least amount.
- The proportion of the total variation contributed by each component ξ_i is given by $\frac{\lambda_{ii}}{\sum \lambda_{ii}}$.

- The data covariance matrix Σ ($p \times p$ matrix) is transformed into a diagonal matrix Λ ($p \times p$ matrix), whose diagonal elements λ_{ii} are the variances of the new variables $\xi = (\xi_1, ..., \xi_p)$.
- This diagonal matrix is constructed such that the diagonal elements are in descending order.
- Hence $\lambda_{11} (= \lambda_1)$ represents the largest amount of variation in the data and λ_{pp} the least amount.
- The proportion of the total variation contributed by each component ξ_i is given by $\frac{\lambda_{ii}}{\sum \lambda_{ii}}$.
- We retain only the components which together represent a certain percentage of the total variation, say, 90%.

Eigenvalues - Resume

- Each principal component axis contains a fraction of the total variance (also called the eigenvalue) of the data cluster.

Eigenvalues - Resume

- Each principal component axis contains a fraction of the total variance (also called the eigenvalue) of the data cluster.
- The percentage of total variance expressed by a given PC axis is called “explained variance”.

Eigenvalues - Resume

- Each principal component axis contains a fraction of the total variance (also called the eigenvalue) of the data cluster.
- The percentage of total variance expressed by a given PC axis is called “explained variance”.
- The sum of all percentages of explained variances equals 100%.

Eigenvalues - Resume

- Each principal component axis contains a fraction of the total variance (also called the eigenvalue) of the data cluster.
- The percentage of total variance expressed by a given PC axis is called “explained variance”.
- The sum of all percentages of explained variances equals 100%.
- If PC1 represents 33% of the total variance, it means that it summarizes 33% of the total information contained in the data.

Eigenvectors and Eigenvalues - Resume

- Eigenvectors and values exist in pairs: every eigenvector has a corresponding eigenvalue.
- The number of eigenvectors/values that exist equals the number of dimensions the data set has (p).
- An eigenvector is a direction, (vertical, horizontal, 45 degrees etc.)

How to start

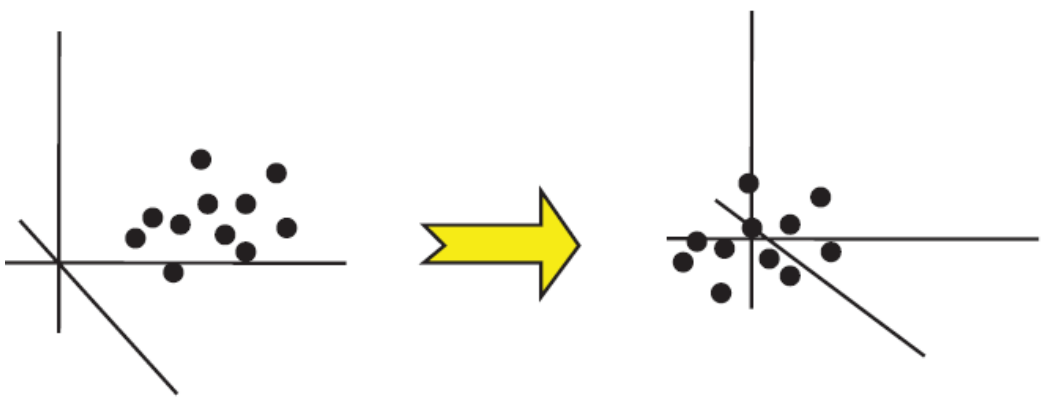
We assume that the multi-dimensional data have been collected in a data matrix, in which the rows are associated with the cases and the columns with the variables.

Centering

- Centering a dataset consists of substracting a value from each cell; this value is calculated from a block of rows or columns.
- In a table where columns are variables, centering by column corresponds to deducing from each value of the variable the mean of this variable.

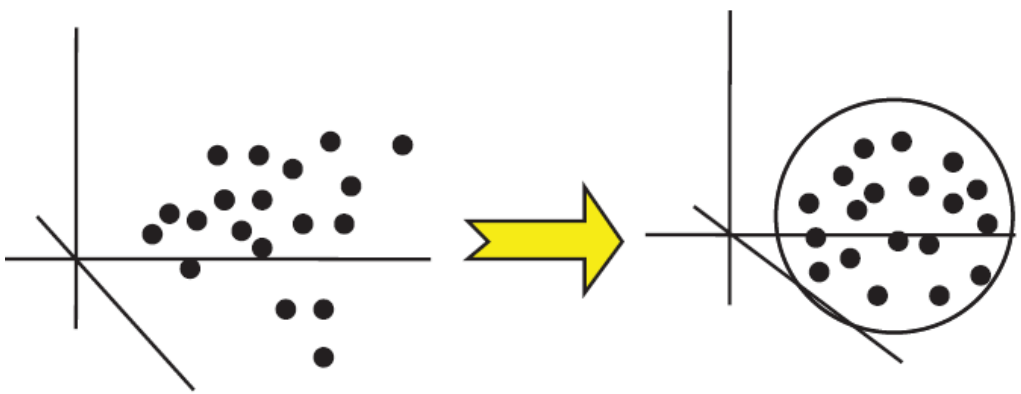
Why center?

- After centering, in a PC map all the points are translated toward the origin of the axes.
- Centering increases the expression of the variability, and therefore the readability of the PC map (structures appear more clearly).



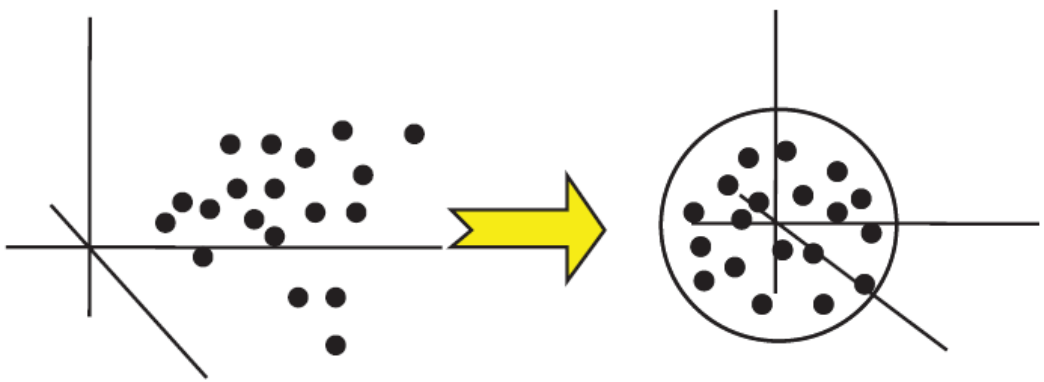
Reducing

- Reducing a table consists of dividing each column of the table by its standard deviation.
- Data must be reduced when they are expressed in different units or orders of magnitude (which requires standardization before analysis and plotting).



Normalization (standardization)

- Normalizing consists of centering and reducing data.
- The reasons for normalizing are the same as those for centering and reducing, i.e. respectively removing the mean to better express specific patterns and homogenizing data expressed in different units.



Preparing Data Analysis

- Start From Correlation Matrix or Covariance Matrix
  - The correlation matrix is simply the covariance matrix, standardized by setting all variances equal to one.
  - When scales of variables are similar, the covariance matrix is always preferred, as the correlation matrix will lose information when standardizing the variance.
  - The correlation matrix is recommended when variables are measured in different scales. (if most of the correlation coefficients are smaller than 0.3, PCA will not help.)

Interpretation of a PC plot

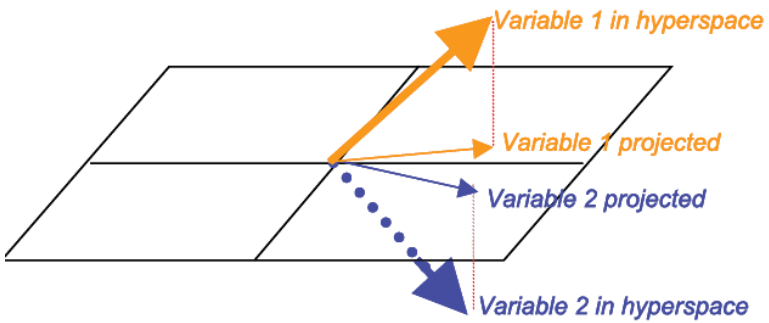
When the multivariate analysis starts, the analyst must decide if the PC map will be that of variables or that of cases. Actually these two PC maps will answer two different questions:

- what are the correlations between variables? map of variables
- what are the correlations or similarities between cases? map of cases

- The PC map of variables can be superimposed on the PC map of cases (and vice versa). One condition for proper interpretation is that the two maps are drawn at the same scale.
- From this superimposition can be deduced correspondences between repetitions and variables.
- The cases or variables close to the origin of axes do not exhibit particular features and thus are not very meaningful in the interpretation.

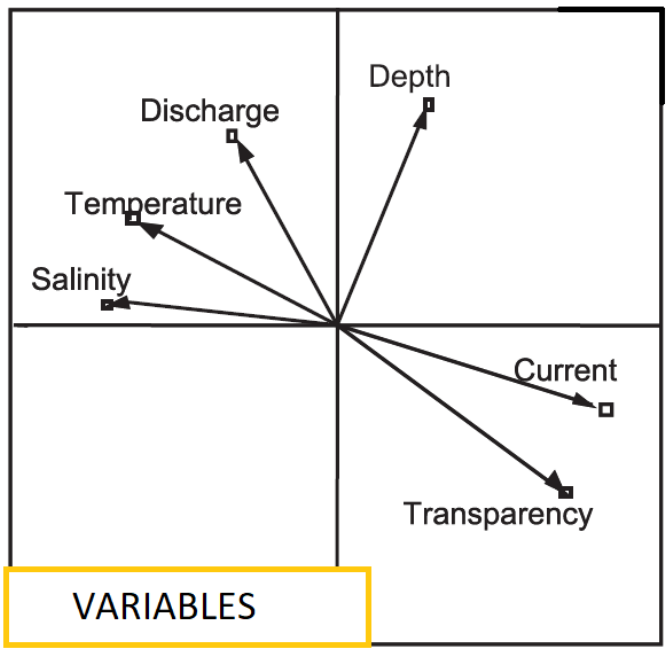
Proximity on the PC map vs. correlation

- Variables are originally in a hyperspace (dimension n),
- They are projected onto a plane (dimension 2),
- In this process of dimensions reduction, it can happen that variables that are distant in the hyperspace are projected close to one another on the factorial map



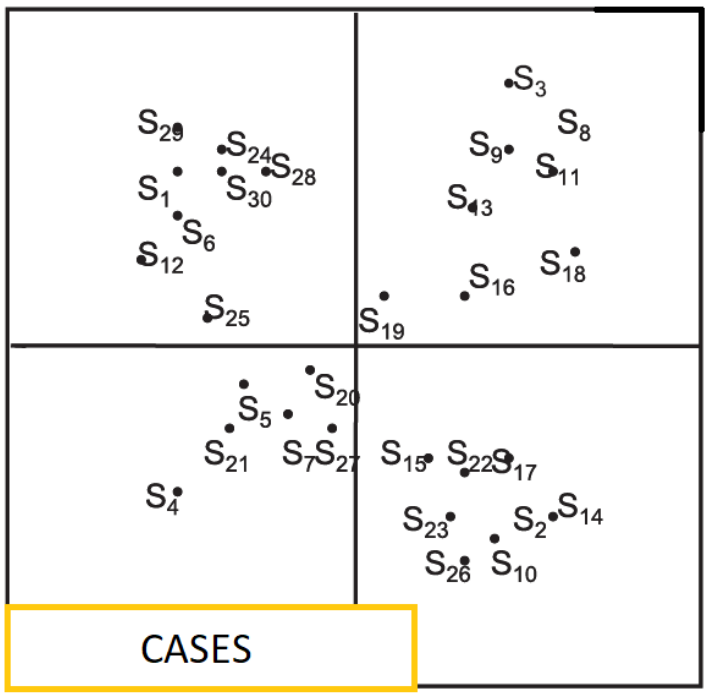
The geographical proximity between two variables on the map will be meaningful only if their correlation is strong, i.e. if the angle between the two variables is small.

Example



- In the PC map of variables are generally represented by vectors originating from the center of the graph;
- Variables close together (e.g. Discharge and Temperature) are correlated in data;
- variables opposed to one another (e.g. Current velocity and Salinity) are anticorrelated (i.e. one has high values when the other has low values);
- variables orthogonal one to another are neither correlated nor anticorrelated, but independent (e.g. Depth and Temperature).

Example



- Sites close together in a particular area of the map are similar in terms of variables measured (e.g. sites S1, S29, S24 have similar values of Temperature and Discharge);
- Sites opposed one to another on the map have opposed values for their variables (e.g. high values of Temperature and Discharge in sites S1 or S29 but low values in sites S10 or S14);
- sites located on a direction orthogonal to others do not have correlated variables (e.g. sites S11, S3, S8 do not have common features with sites S14, S2 or S10).

Examples

See toy example from pca.R file.

Exercises

Task: Exercise from the pca.md file.