

# Data Analysis and Information Exploitation (ADEI)

Bachelor Degree in Informatics Engineering  
Information System Track

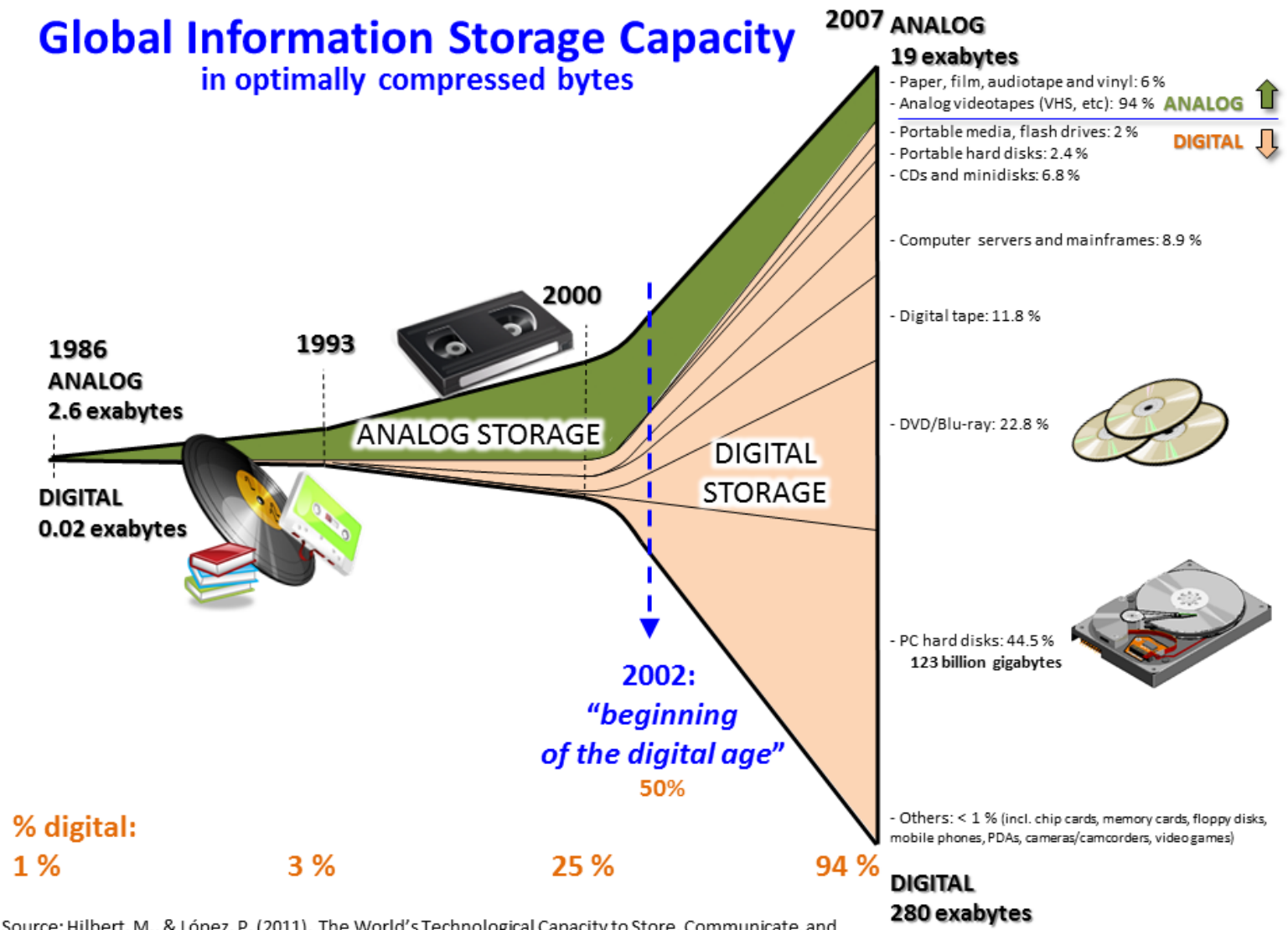
FIB-ADEI – 6 ECTS - Course 2020-21

BarcelonaTech - UPC

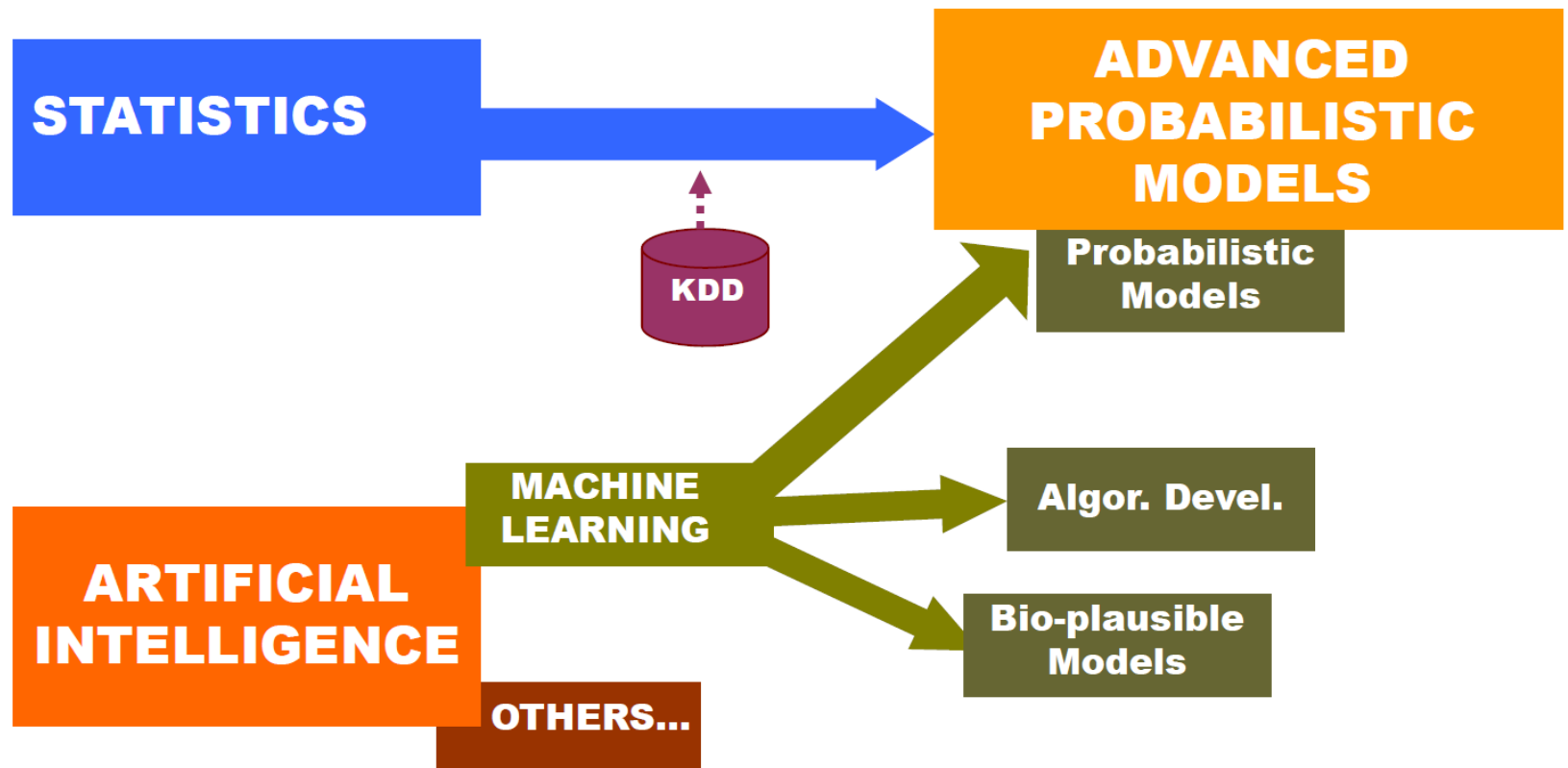
## UNIT 3. KNOWLEGDE DISCOVERY

Lecturer: Lídia Montero (DEIO)

## Global Information Storage Capacity in optimally compressed bytes



Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60 –65. <http://www.martinhilbert.net/WorldInfoCapacity.html>



Source: Lluís Belanche – SOCO -UPC

## Big Data characteristics:

### Volume

The quantity of generated and stored data. The size of the data determines the value and potential insight- and whether it can actually be considered big data or not.

### Variety

The type and nature of the data. This helps people who analyze it to effectively use the resulting insight.

### Velocity

In this context, the speed at which the data is generated and processed to meet the demands and challenges that lie in the path of growth and development.

### Variability

Inconsistency of the data set can hamper processes to handle and manage it.

### Veracity

The quality of captured data can vary greatly, affecting accurate analysis.

Purpose collected data (Surveys) is a fundamental issue

- Today the word "survey" is used most often to describe a method of gathering information from a sample of individuals. This "sample" is usually just a fraction of the population being studied.
- Not only do surveys have a wide variety of purposes, they also can be conducted in many ways
  - including over the telephone, by mail, or in person.
  - Nonetheless, all surveys do have certain characteristics in common.
- Unlike a census, where all members of the population are studied, surveys gather information from only a portion of a population of interest
  - the size of the sample depending on the purpose of the study.

- In a good survey, the sample is not selected haphazardly or only from persons who volunteer to participate.
- It is scientifically chosen so that each person in the population will have a measurable chance of selection.
  - This way, the results can be reliably projected from the sample to the larger population.
- Information is collected by means of standardized procedures so that every individual is asked the same questions in more or less the same way.
  - The use of questionnaires is strongly recommended.
  - All of the survey's results should be presented in completely anonymous summaries, such as statistical tables and charts.

# Data Analysis and Information Exploitation (ADEI)

Bachelor Degree in Informatics Engineering  
Information System Track

FIB-ADEI – 6 ECTS - Course 2020-21

BarcelonaTech - UPC

## TOPIC 3. KNOWLEGDE DISCOVERY

### 3-1. Component And Factor Analysis

- Data from network sensors collecting traffic data from a rich variety of detection stations and sensor types, based on traditional or advanced ICT sources, as for example:
  - Traffic Detection Stations: inductive loop detectors, radars, magnetometers...
  - Antennas to capture Bluetooth or Wi-Fi equipped devices
  - CCTV image processing (e.g. License Plate Recognition, emulation of Traffic Detection Stations)
  - Capture of digital footprints, e.g. TAGs
  - GPS tracking of fleets
  - *Call Detail Records from mobile phones*
- Socioeconomic data from authorities (CAD support currently)
- Data collected from surveys



- **Facial Recognition**
- **Facial Expression Recognition**
- **Data Compression**
- **Dimension Reduction**

- Want to identify specific person, based on facial image
  - Robust to glasses, lighting,...
- ⇒ Can't just use the given 256 x 256 pixels



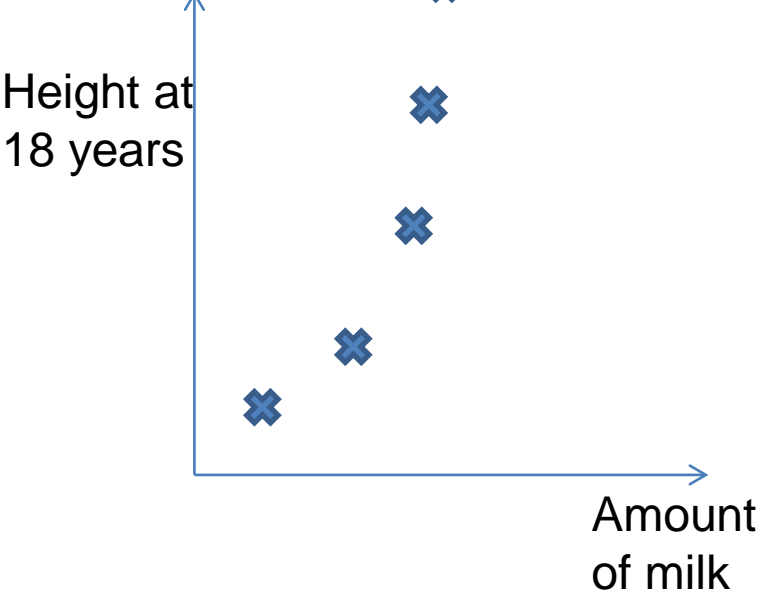
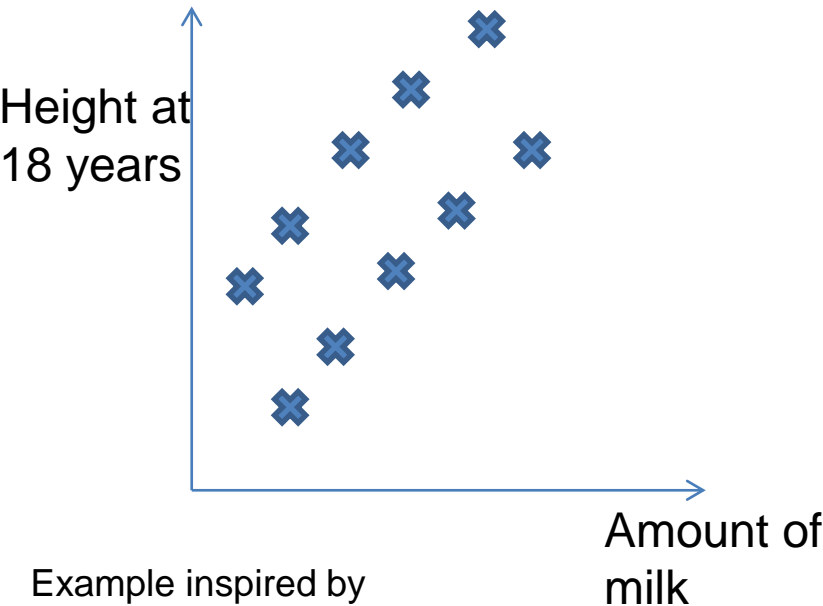
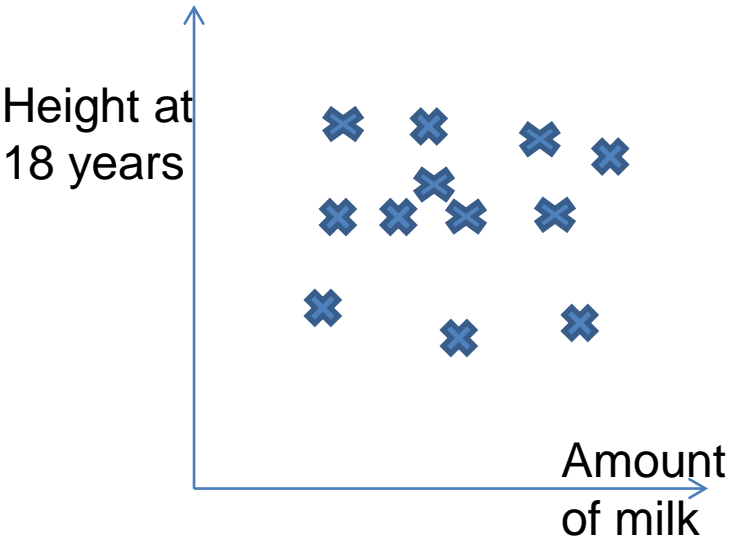
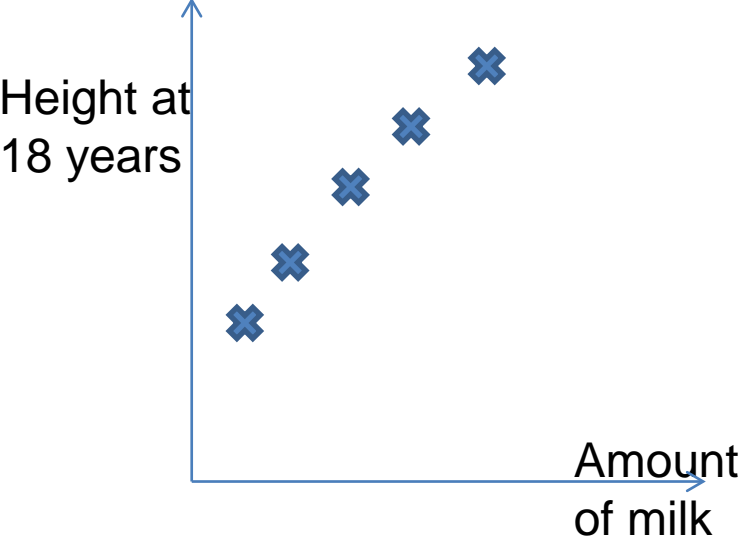
- We study phenomena that can not be directly observed
  - ego, personality, intelligence in psychology
  - Underlying factors that govern the observed data
- We want to identify and operate with underlying latent factors rather than the observed data
  - E.g. topics in news articles
  - Transcription factors in genomics
- We want to discover and exploit hidden relationships
  - “beautiful car” and “gorgeous automobile” are closely related
  - So are “driver” and “automobile”
  - But does your search engine know this?
  - Reduces noise and error in results

- Discover a new set of factors/dimensions/axes against which to represent, describe or evaluate the data
  - For more effective reasoning, insights, or better visualization
  - Reduce noise in the data
  - Typically a smaller set of factors: dimension reduction
  - Better representation of data without losing much information
  - Can build more effective data analyses on the reduced-dimensional space: classification, clustering, pattern recognition
- Factors are combinations of observed variables
  - May be more effective bases for insights, even if physical meaning is obscure
  - Observed data are described in terms of these factors rather than in terms of original variables/dimensions

- Areas of variance in data are where items can be best discriminated and key underlying phenomena observed
  - Areas of greatest “signal” in the data
- If two items or dimensions are highly correlated or dependent
  - They are likely to represent highly related phenomena
  - If they tell us about the same underlying variance in the data, combining them to form a single measure is reasonable
    - Parsimony
    - Reduction in Error
- So the goal is to combine related variables, and focus on uncorrelated or independent ones, especially those along which the observations have high variance
- Obtaining a smaller set of variables that explain most of the variance in the original data, in more compact and insightful form
  - Dimension reduction technique

- What if the dependences and correlations are not so strong or direct?
- And suppose you have 3 variables, or 4, or 5, or 10000?
- Look for the phenomena underlying the observed covariance/co-dependence in a set of variables
  - Once again, phenomena that are uncorrelated or independent, and especially those along which the data show high variance
- These phenomena are called “factors” or “principal components” or “independent components,” depending on the methods used
  - Factor analysis: based on variance/covariance/correlation
  - Independent Component Analysis: based on independence

- Data Visualization
- Data Compression
- Noise Reduction
- Data Classification
- Trend Analysis
- Factor Analysis



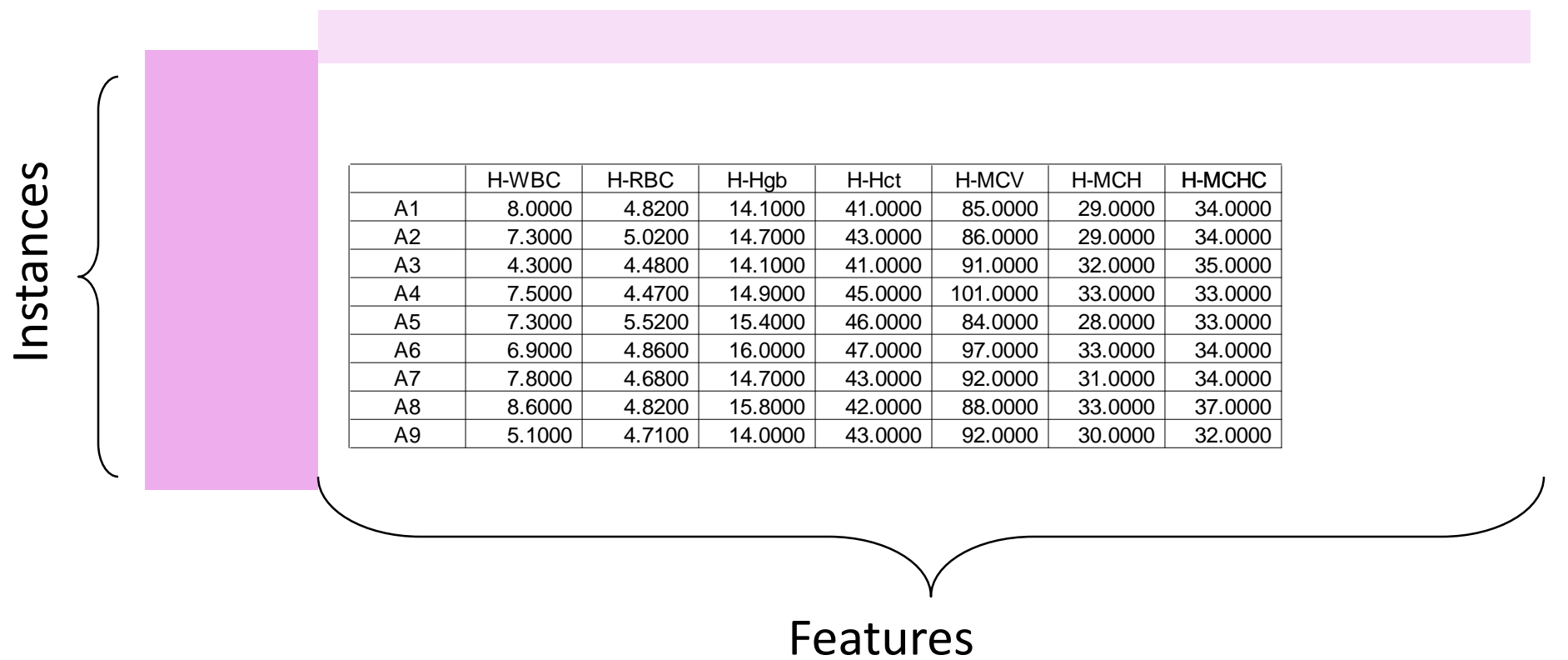
Example inspired by  
Monique Becue notes



Example:

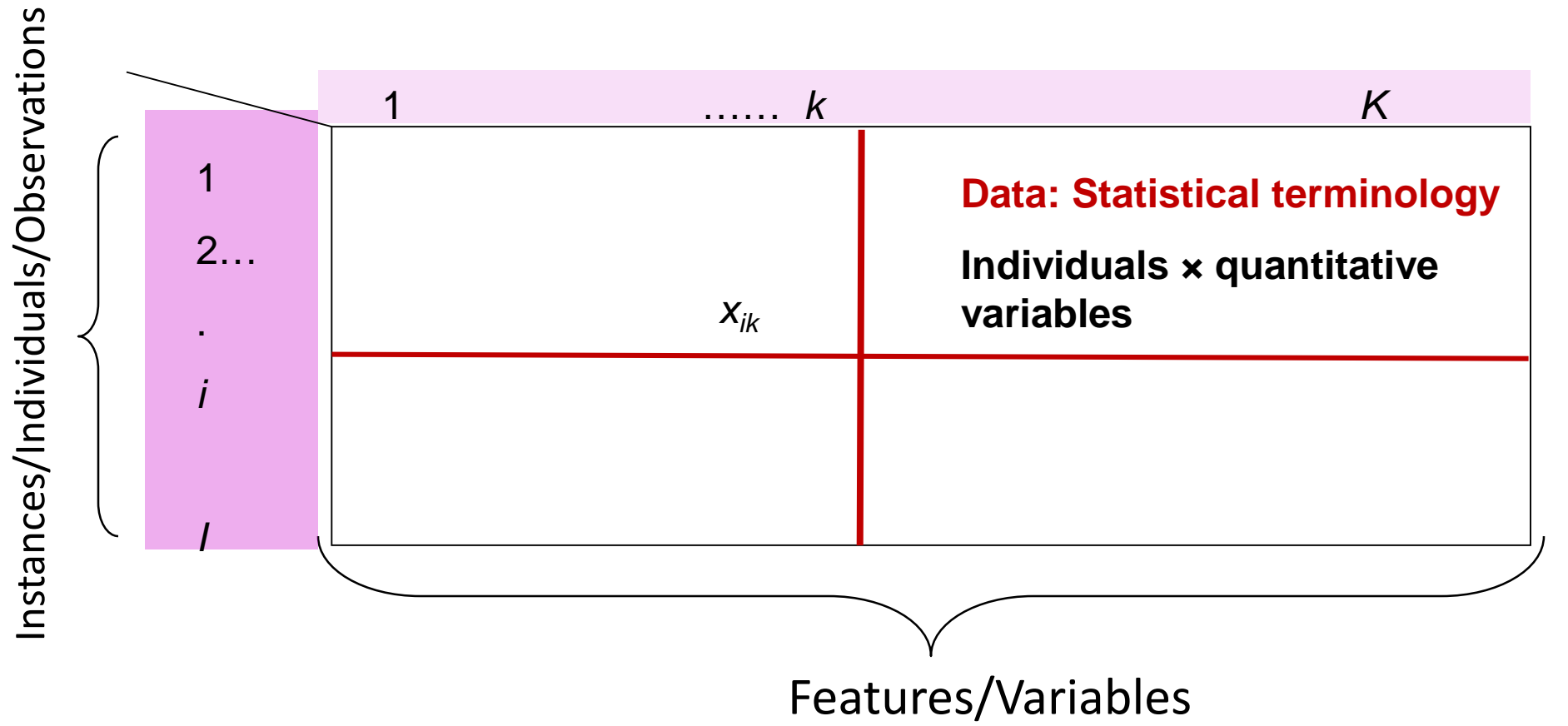
- Given 53 blood and urine samples (features) from 65 people.
- How can we visualize the measurements?

- Matrix format (65x53)



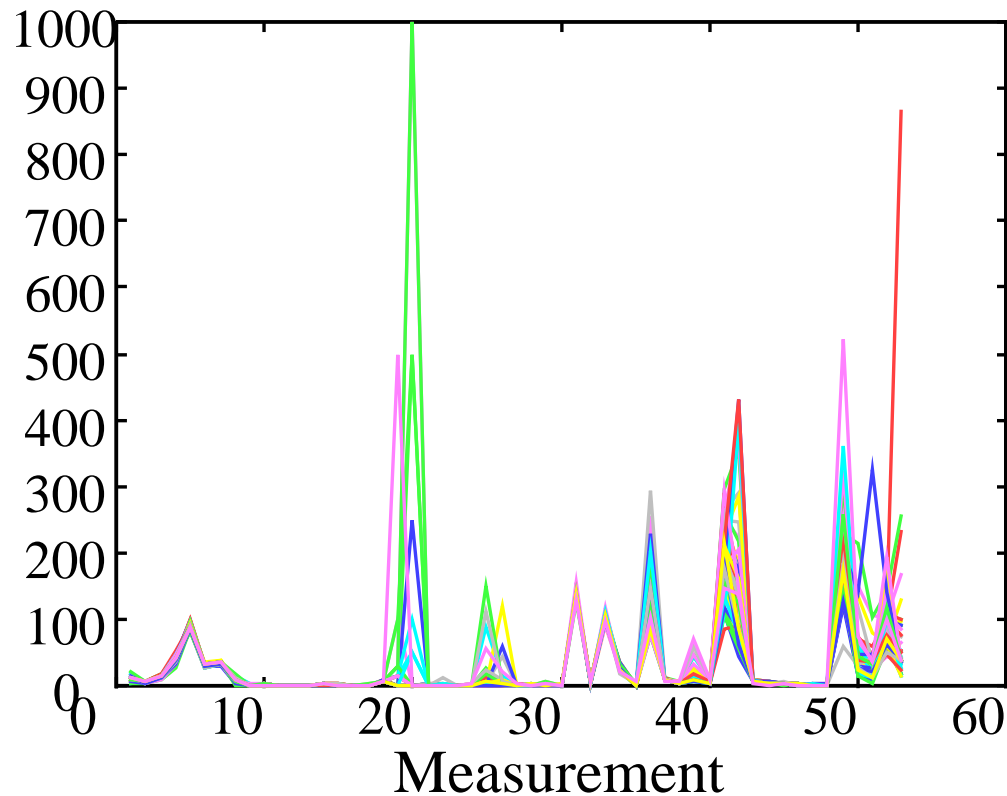
Difficult to see the correlations between the features...

- Matrix format (65x53) :Duality of the table



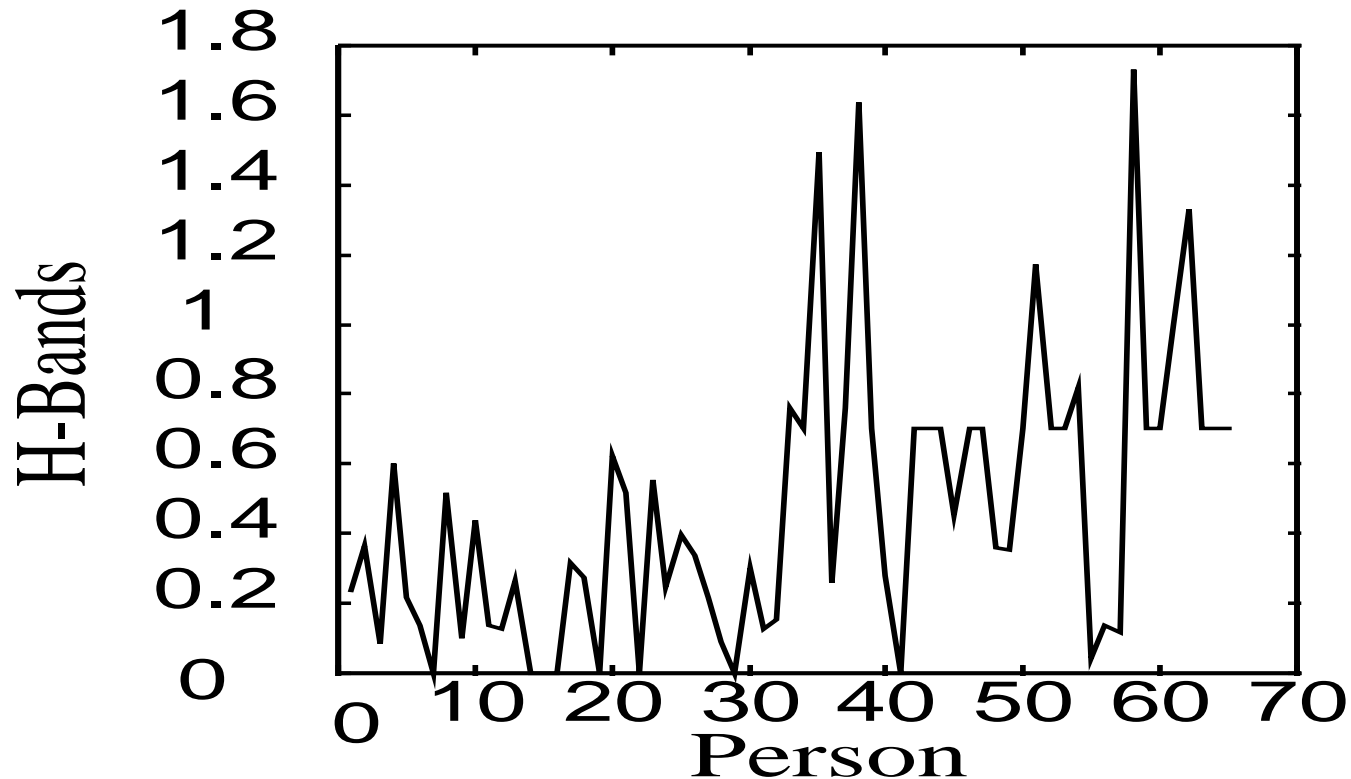
Difficult to see the correlations between the features...

- Spectral format (65 pictures, one for each person)



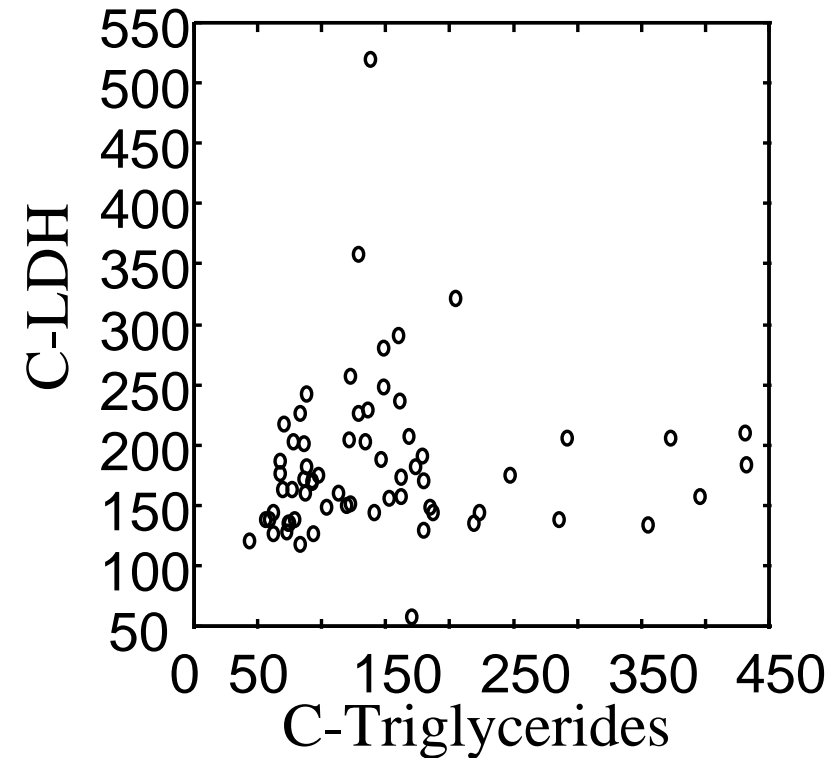
**Difficult to compare the different patients...**

- Spectral format (53 pictures, one for each feature)

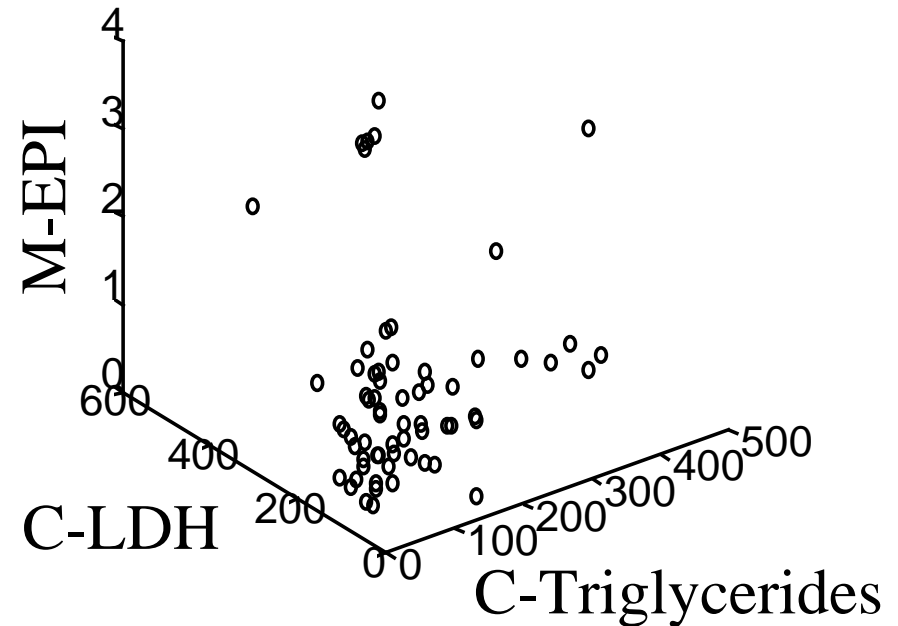


Difficult to see the correlations between the features...

## Bi-variate



## Tri-variate



- **How can we visualize the other variables???**

... difficult to see in 4 or higher dimensional spaces...

- Is there a representation better than the coordinate axes?
- Is it really necessary to show all the 53 dimensions?
  - ... what if there are strong correlations between the features?
- How could we find the *smallest* subspace of the 53-D space that keeps the *most information* about the original data?
- Solution: **Principal Component Analysis**

## 1. Compresses the information

- Finds the directions with most variability
- Projects the information down on these dimensions
- The goal of PCA is to reduce the dimensionality of the data while **retaining as much as possible of the variation present in the dataset.**

## 2. Presents the information in simple plots

- **Scores plot**
  - Projection of data onto subspace
- **Loadings plot**
  - Plot of relation between original variables and subspace dimensions

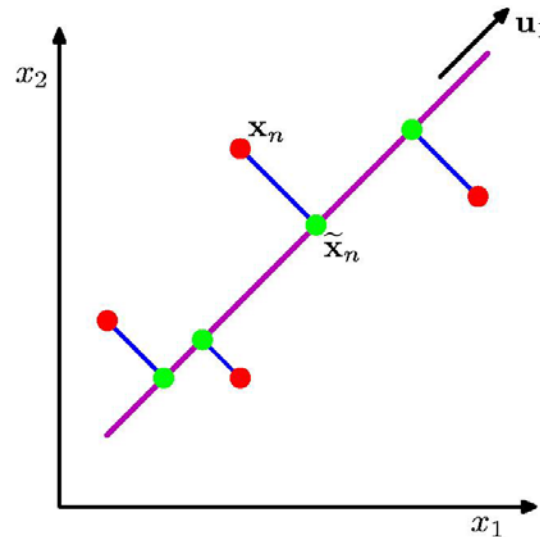


## Goals of PCA: Highlight ...

- The structure of the row-individuals through an Euclidean representation to **detect the individuals that are similar from the point of view of the active variables**
- The structure of the column-variables a representation that **evidences the variables highly correlated**

This method aims at discovering the data structure, the underlying system, the patterns, the general rules but also the clues towards hidden information and explain:

- The variability of the individuals from the variables point of view
- The dispersion of the variables from the individuals point of view



Orthogonal projection of data onto lower-dimension linear space that...

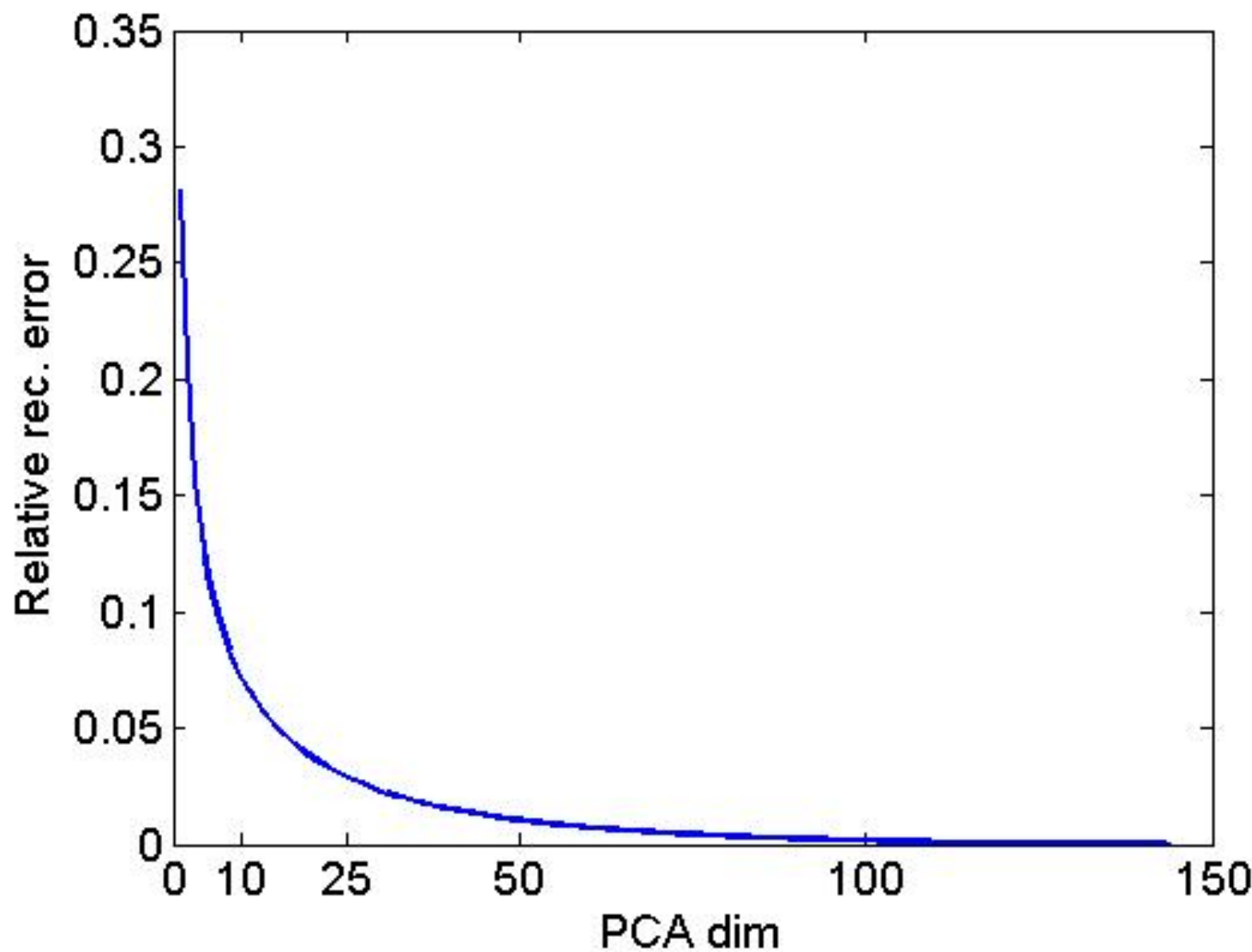
- maximizes variance of projected data (purple line)
- minimizes mean squared distance between
  - data point and
  - projections (sum of blue lines)

## Idea:

- Given data points in a  $d$ -dimensional space, project into **lower dimensional** space while **preserving as much information** as possible
  - Eg, find best planar approximation to 3D data
  - Eg, find best 12-D approximation to  $10^4$ -D data
- In particular, choose projection that ***minimizes squared error*** in reconstructing original data

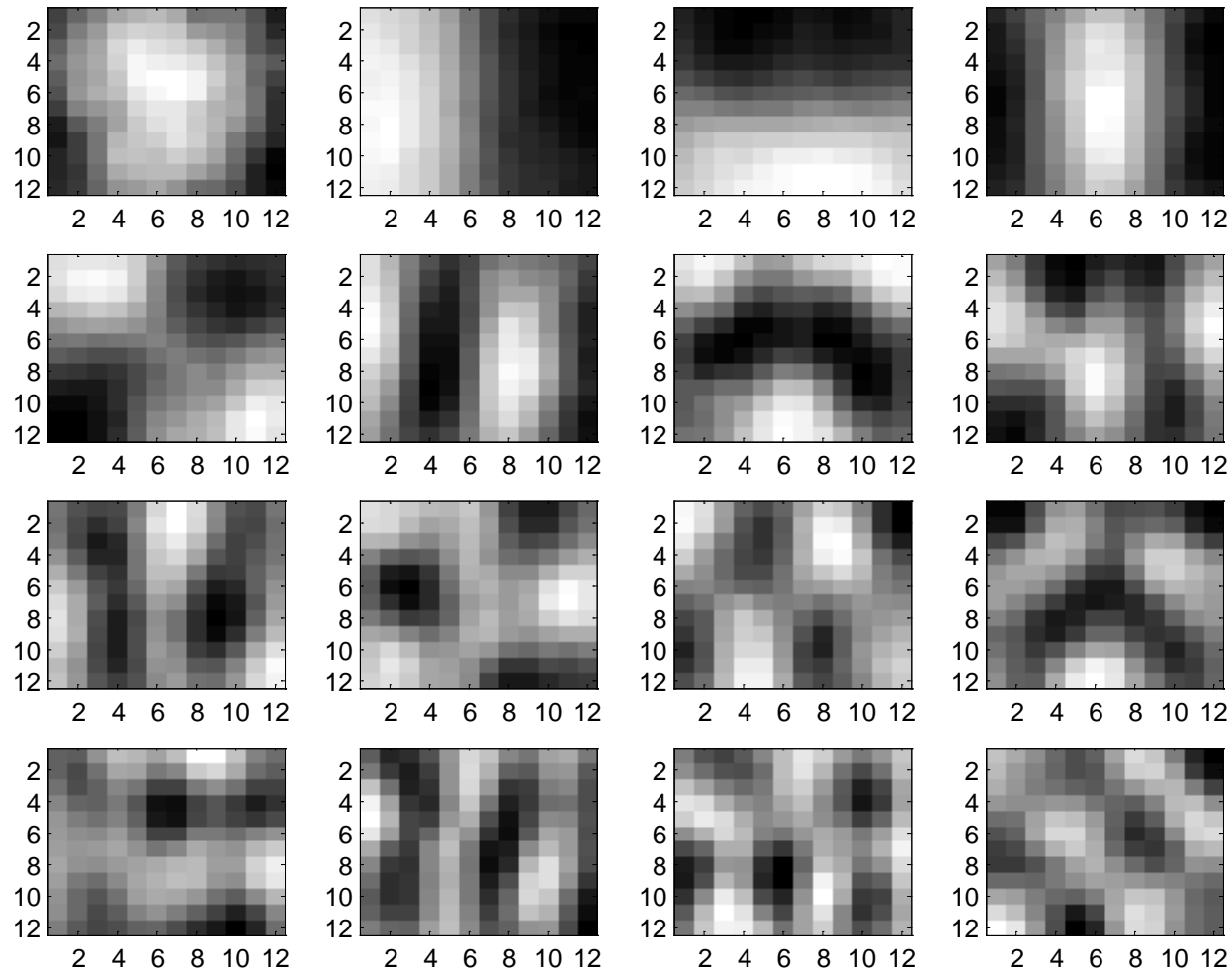


- Divide the original 372x492 image into patches
- Each patch is an instance that contains 12x12 pixels on a grid
- View each as a 144-D vector



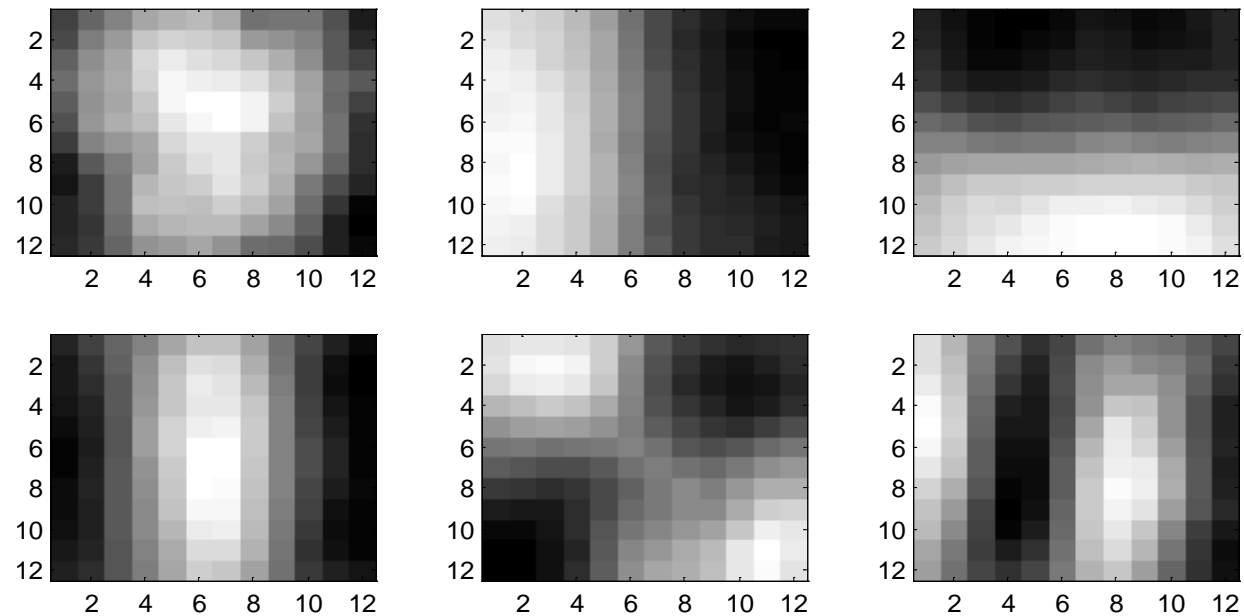






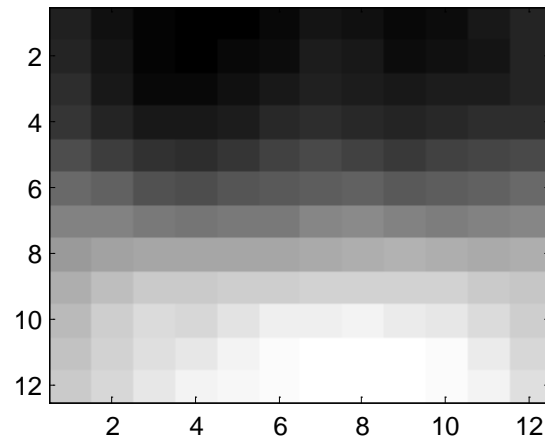
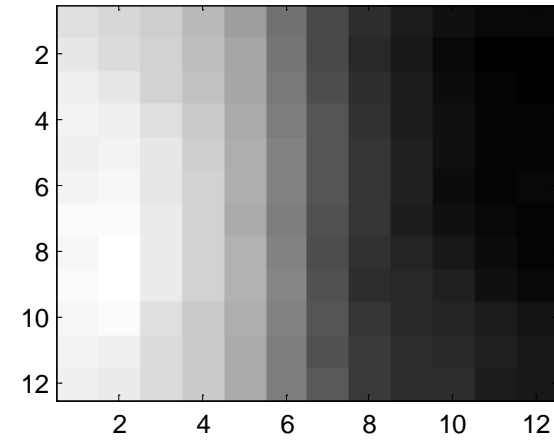
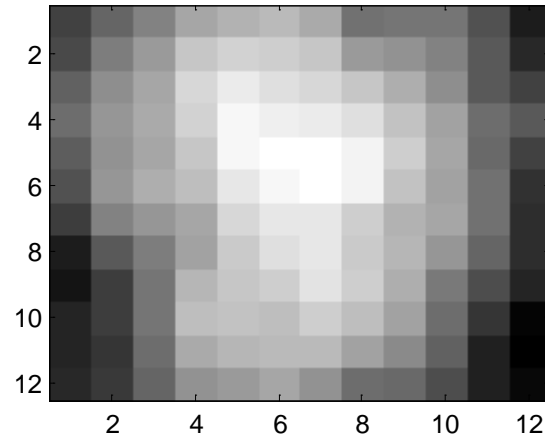


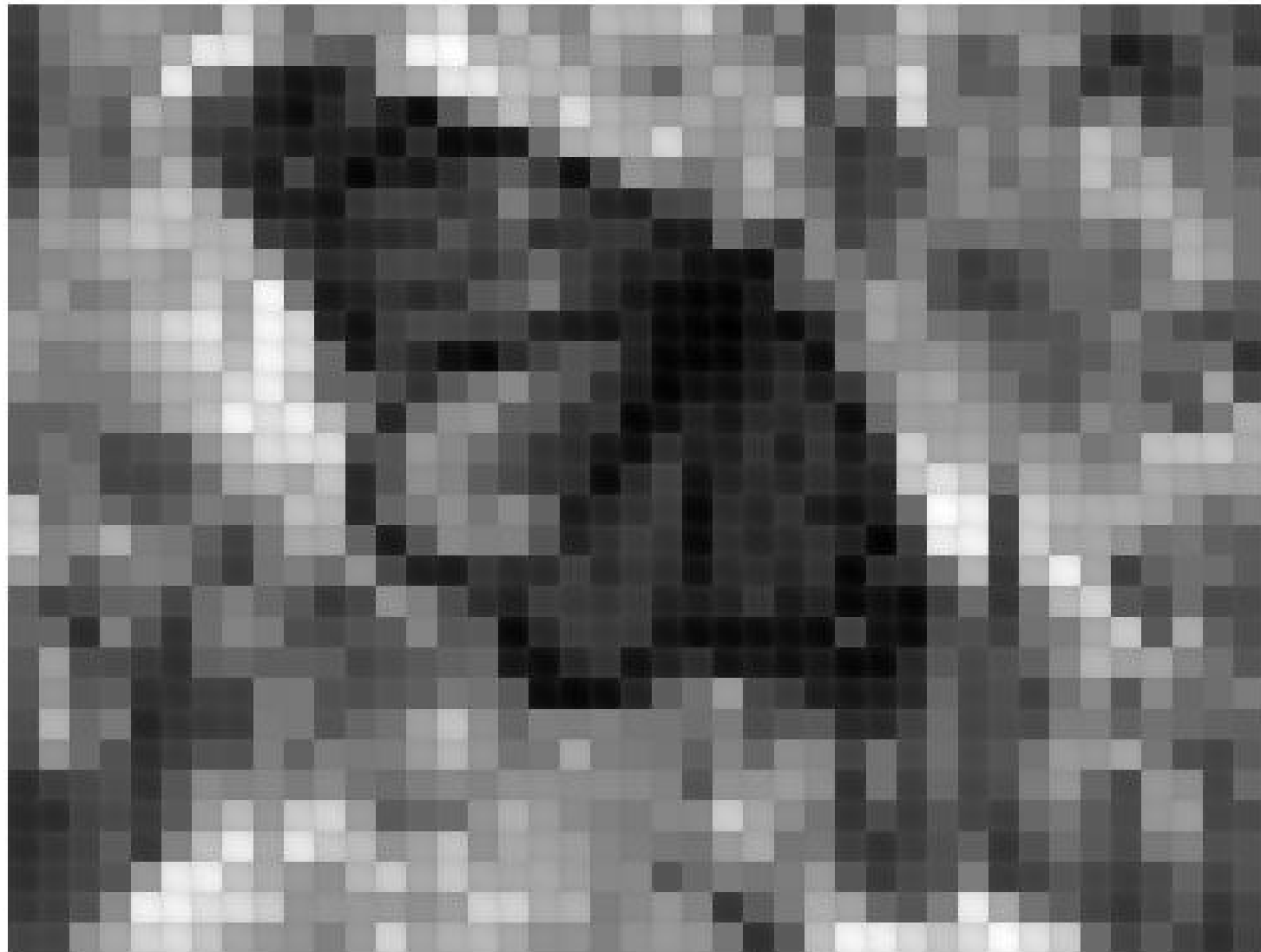


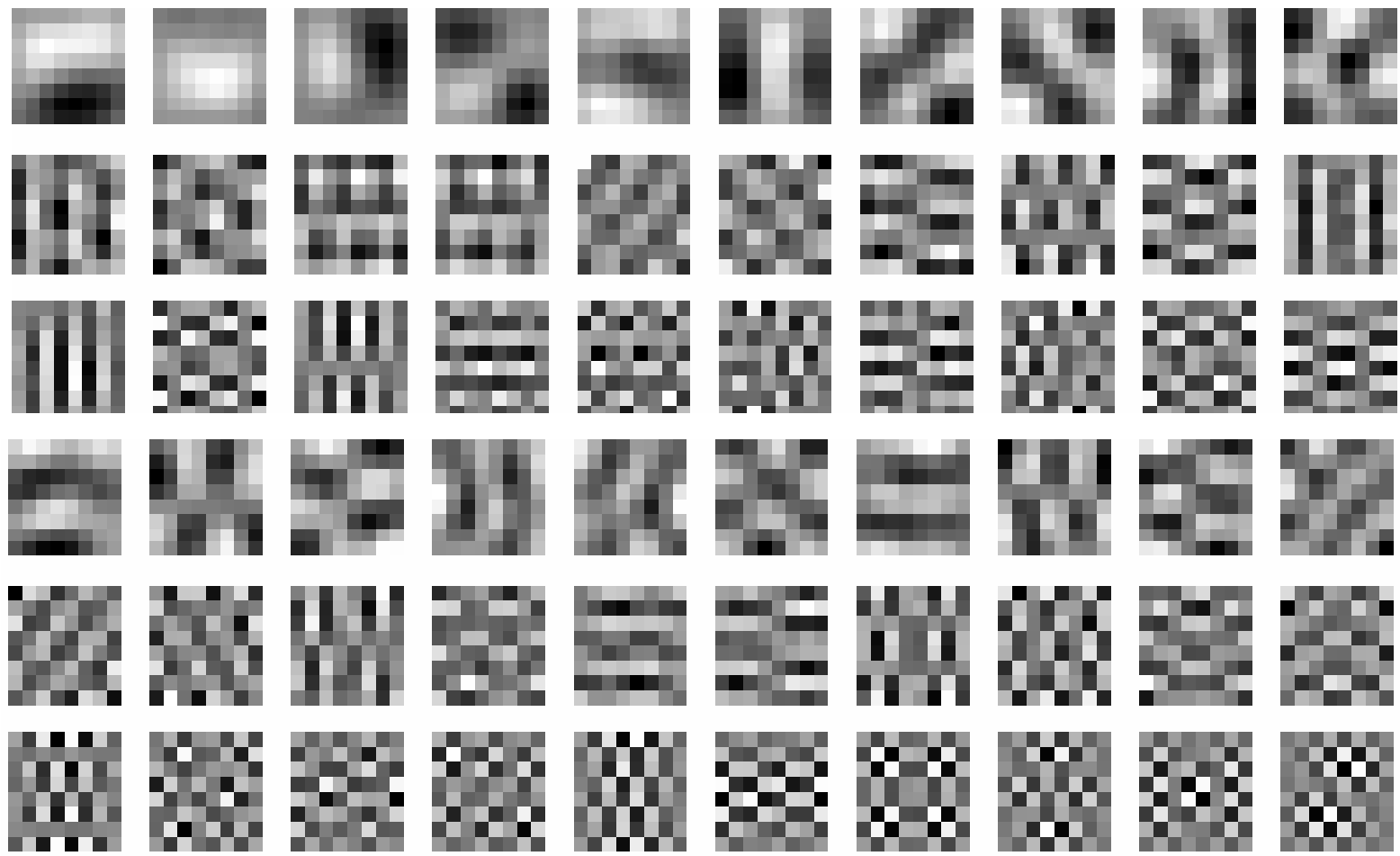




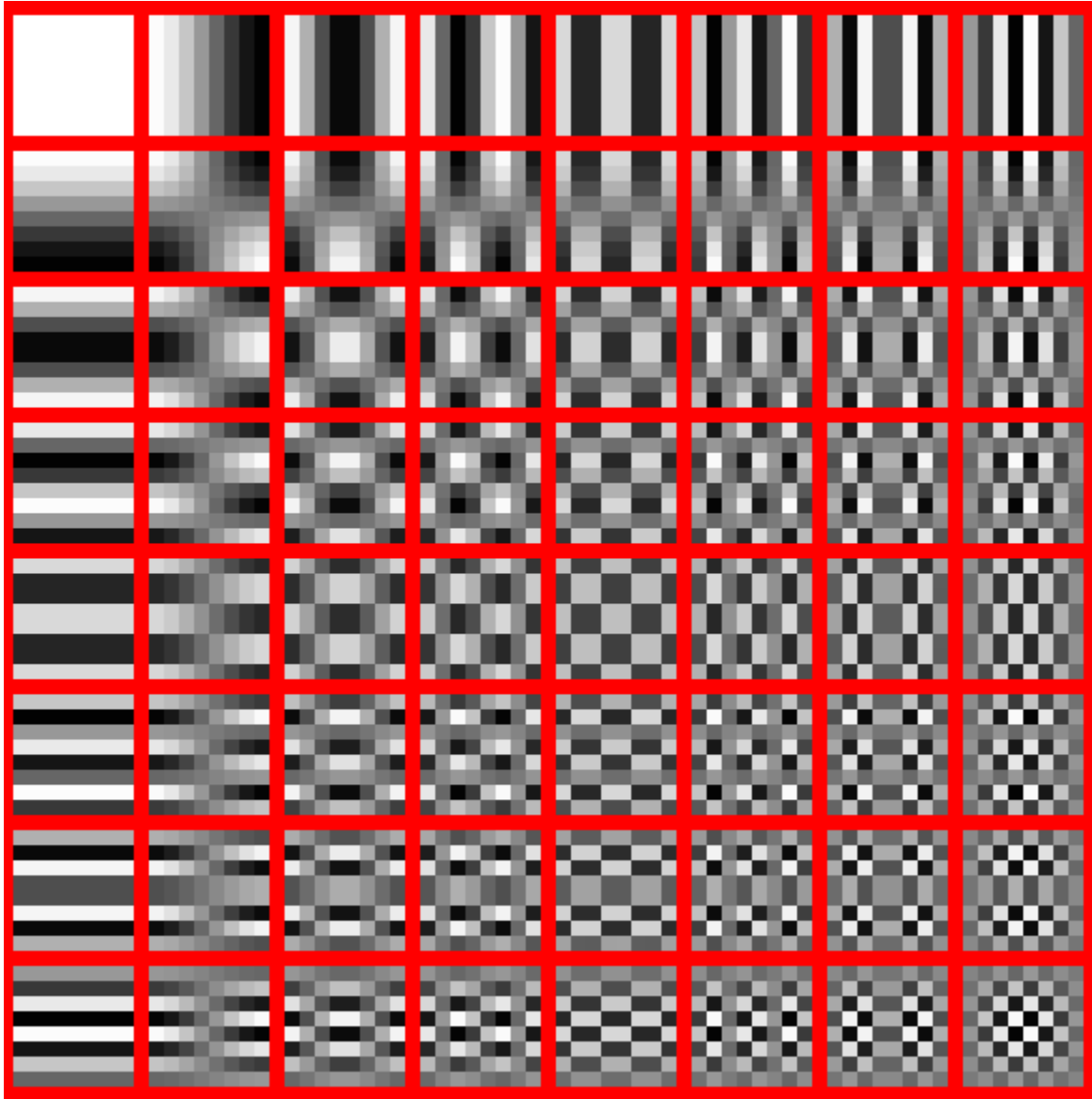
# 3 most important eigenvectors







Looks like the discrete cosine bases of JPG!...



- Noisy picture





- Denoised image using 15 PCA components



- **Vectors** originating from the center of mass, in PCA, either matrix  $Y$  or matrix  $Z$  are used with general terms:

$$y_{ik} = (x_{ik} - \bar{x}_k)$$

$$z_{ik} = \frac{(x_{ik} - \bar{x}_k)}{s_k}$$

**Normalized**

- Principal component #1 points in the direction of the **largest variance (inertia)**.
- Each subsequent principal component...
  - is **orthogonal** to the previous ones, and
  - points in the directions of the **largest variance of the residual subspace**

- Degree to which the variables are linearly correlated is represented by their **covariances**.

$$y_{ik} = (x_{ik} - \bar{x}_k)$$
$$C_{ij} = \frac{1}{I-1} \sum_{m=1}^I (x_{mi} - \bar{x}_i)(x_{mj} - \bar{x}_j)$$

Diagram illustrating the formula for covariance  $C_{ij}$  with annotations:

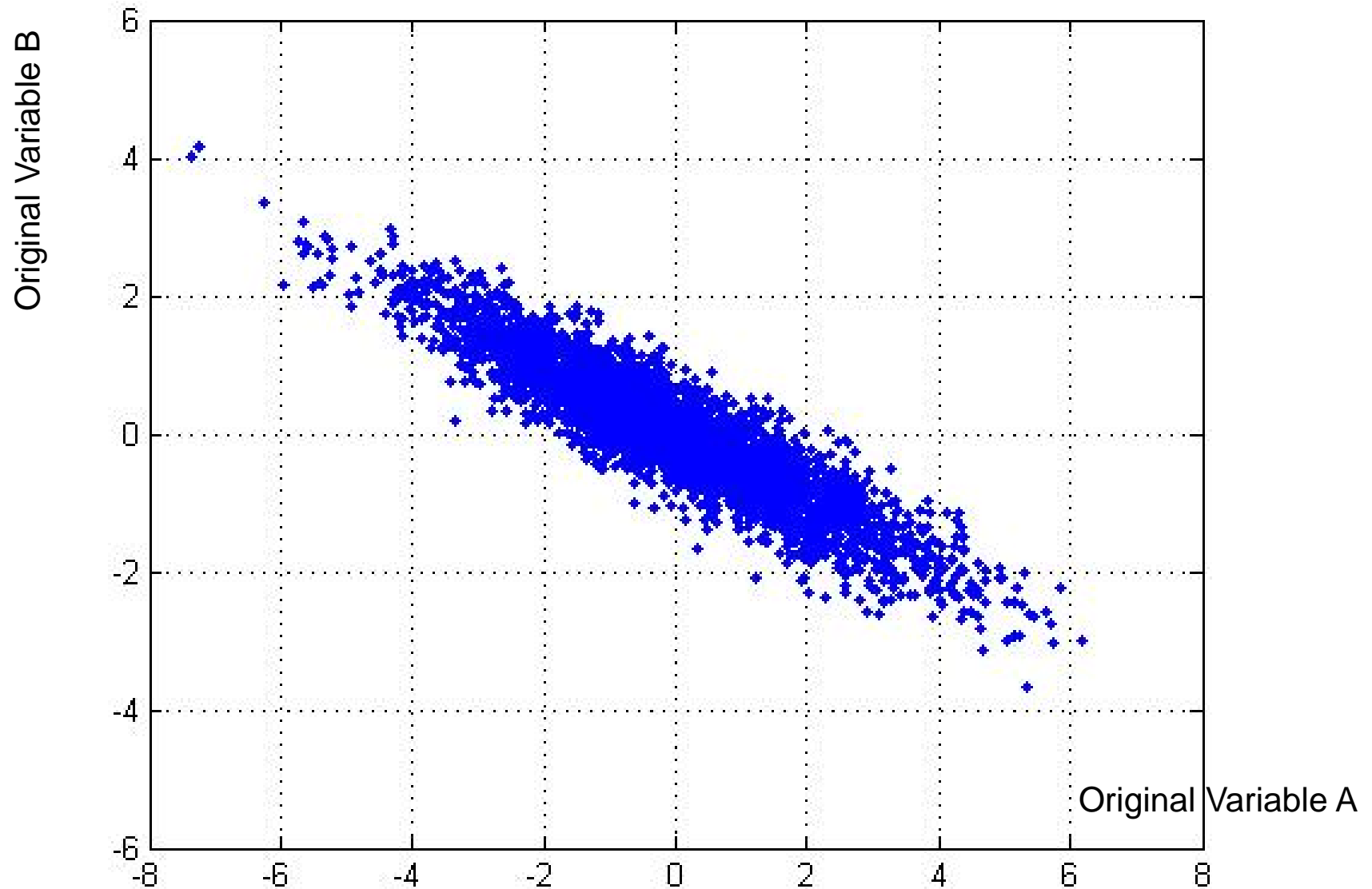
- $C_{ij}$ : Covariance of variables  $i$  and  $j$
- $\sum_{m=1}^I$ : Sum over all  $I$  objects
- $x_{mi}$ : Value of variable  $i$
- $\bar{x}_i$ : Mean of variable  $i$
- $x_{mj}$ : Value of variable  $j$  in object  $m$
- $\bar{x}_j$ : Mean of variable  $j$

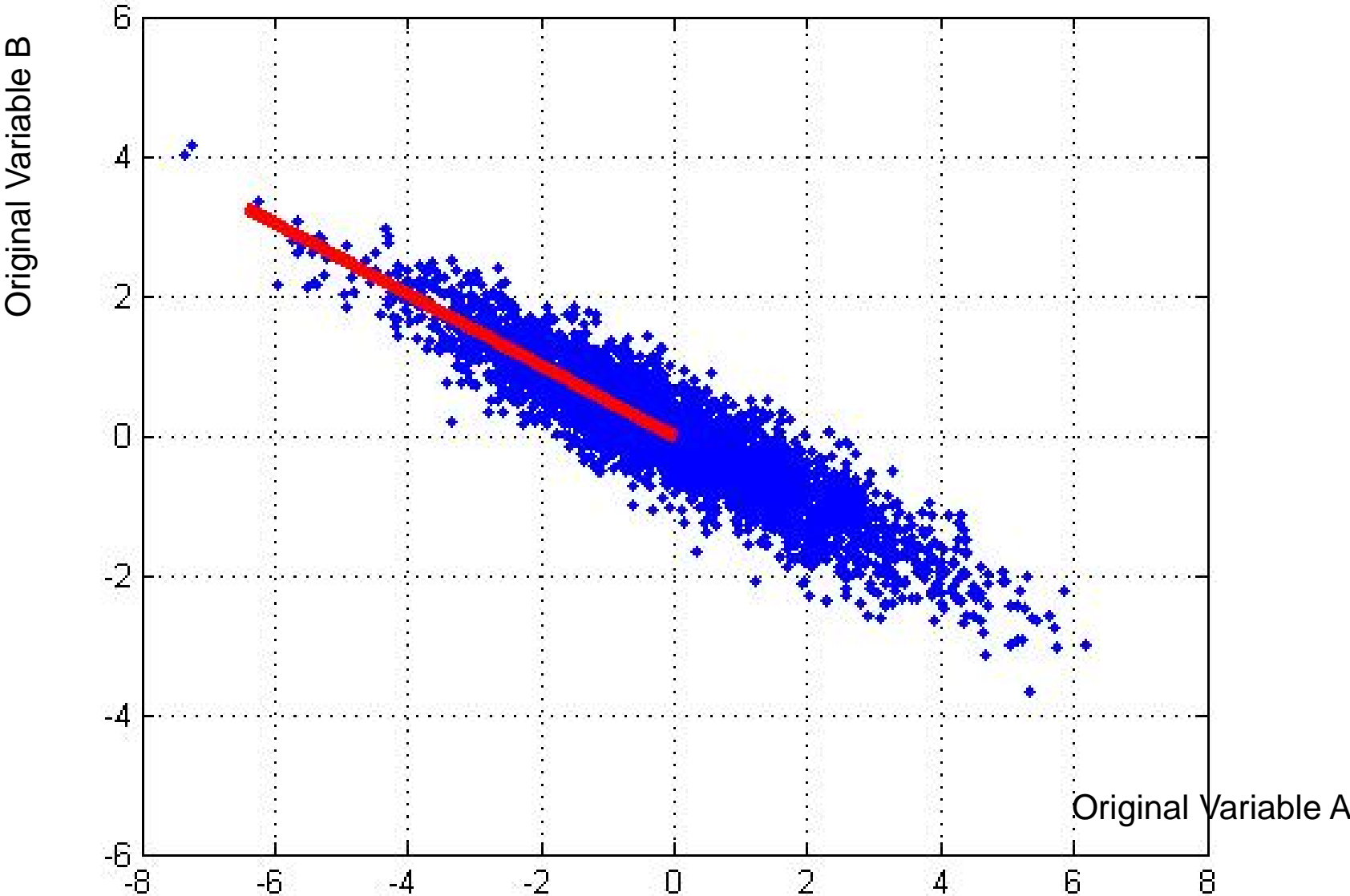
- Normalized PCA: Degree to which the variables are linearly correlated is represented by their **correlations**.

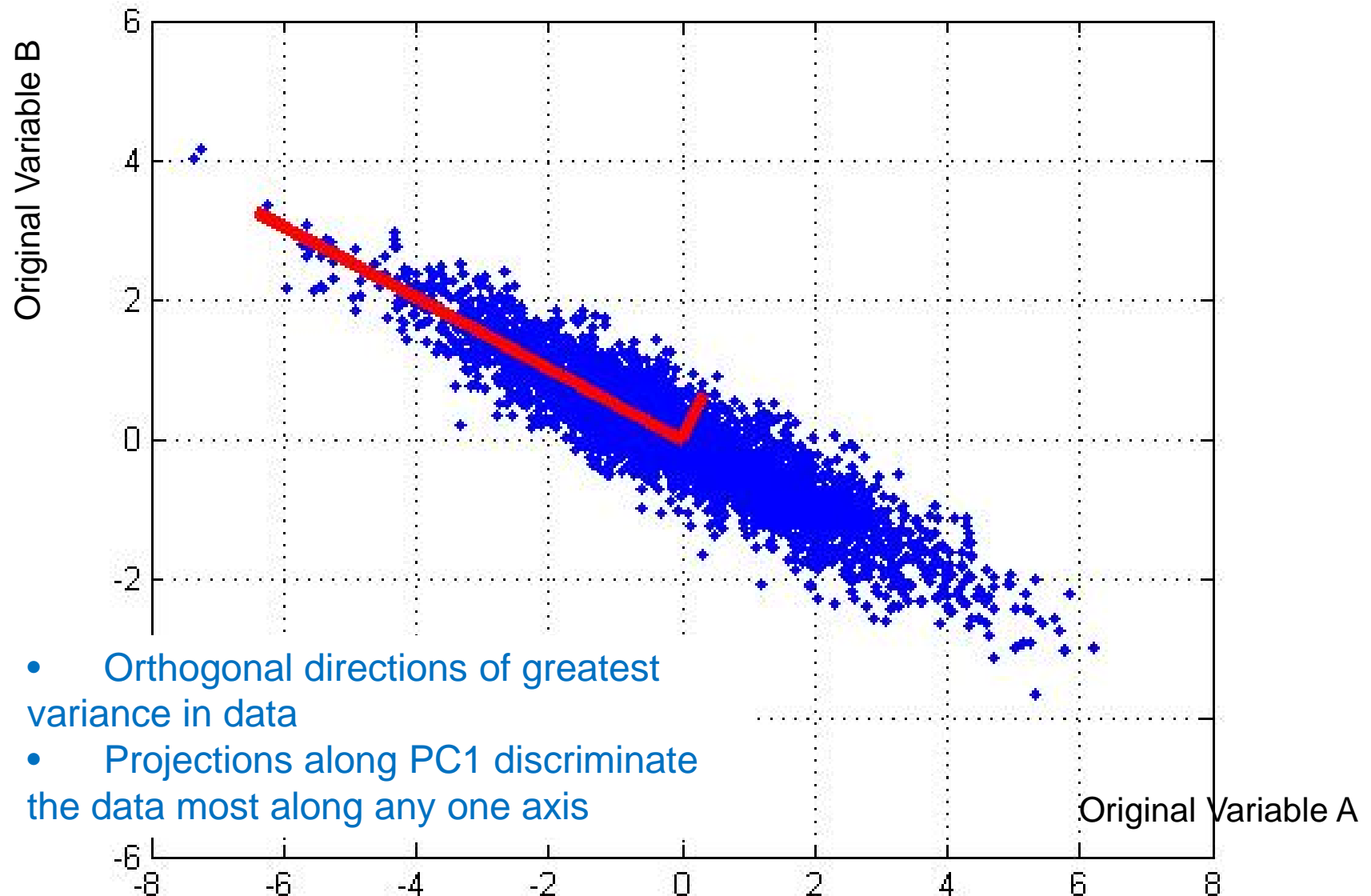
$$z_{ik} = \frac{(x_{ik} - \bar{x}_k)}{s_k}$$

$$C_{ij} = \frac{1}{I-1} \sum_{m=1}^I \frac{(X_{mi} - \bar{X}_i)(X_{mj} - \bar{X}_j)}{S_i S_j}$$

Correlation of variables  $i$  and  $j$  (points to  $C_{ij}$ )  
 Sum over all  $I$  objects (points to  $\sum_{m=1}^I$ )  
 Value of variable  $i$  (points to  $X_{mi}$ )  
 St deviation of variable  $i$  (points to  $S_i$ )  
 Mean of variable  $i$  (points to  $\bar{X}_i$ )  
 Value of variable  $j$  in object  $m$  (points to  $X_{mj}$ )  
 St deviation of variable  $j$  (points to  $S_j$ )  
 Mean of variable  $j$  (points to  $\bar{X}_j$ )





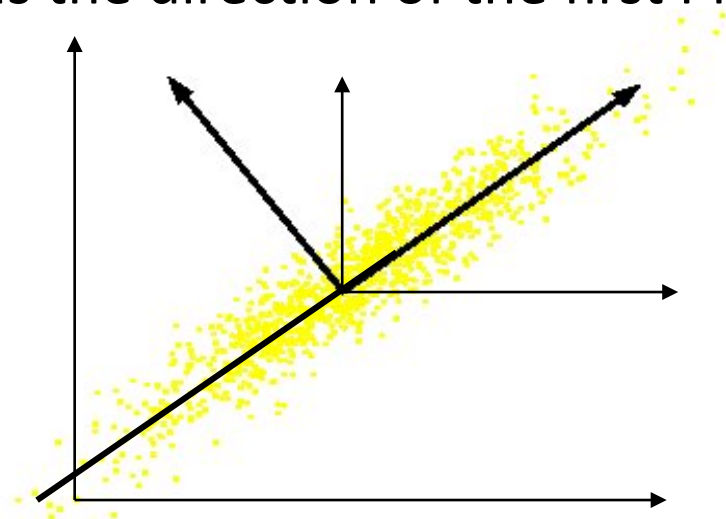


- First principal component is the direction of greatest variability (covariance) in the data
- Second is the next orthogonal (uncorrelated) direction of greatest variability
  - So first remove all the variability along the first component, and then find the next direction of greatest variability
- And so on ...



- Principle
  - Linear projection method to reduce the number of parameters
  - Transfer a set of correlated variables into a new set of uncorrelated variables
  - Map the data into a space of lower dimensionality
  - Form of unsupervised learning
- Properties
  - It can be viewed as a rotation of the existing axes to new positions in the space defined by original variables
  - New axes are orthogonal and represent the directions with maximum variability

- Data points are vectors in a multidimensional space
- Projection of vector  $\mathbf{x}$  onto an axis (dimension)  $\mathbf{u}$  is  $\mathbf{x} \cdot \mathbf{u}$
- Direction of greatest variability is that in which the average square of the projection is greatest
  - I.e.  $\mathbf{u}$  such that  $E((\mathbf{x} \cdot \mathbf{u})^2)$  over all  $\mathbf{x}$  is maximized
  - Usually, just subtract the mean along each dimension, thus center the original axis system at the centroid of all data points, for simplicity.
  - This direction of  $\mathbf{u}$  is the direction of the first Principal Component

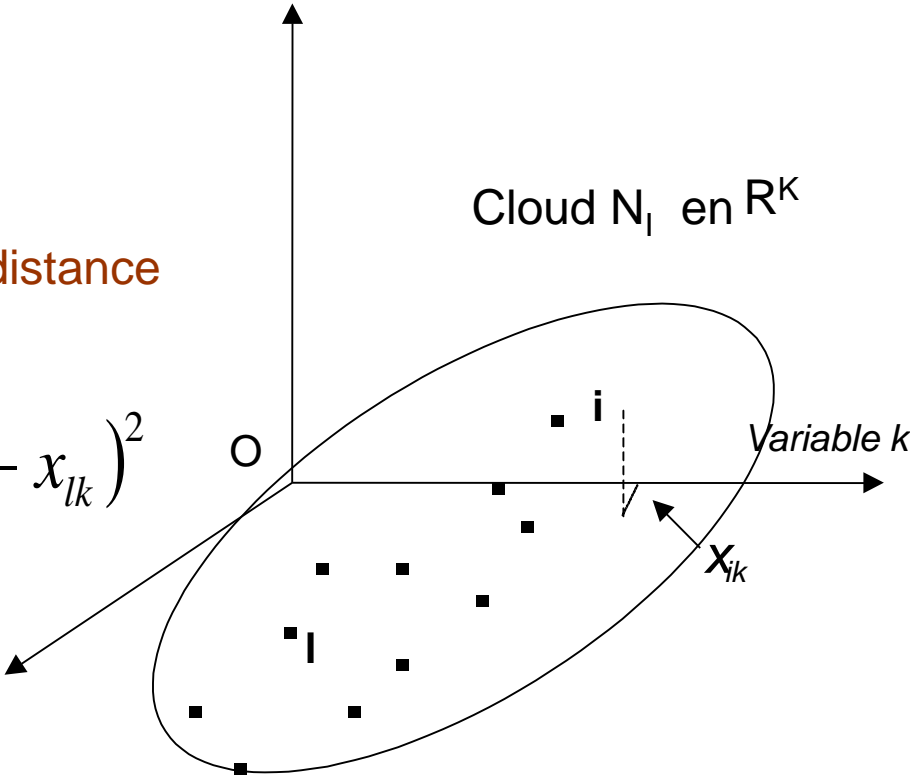


## Principal component analysis

Cloud of individuals

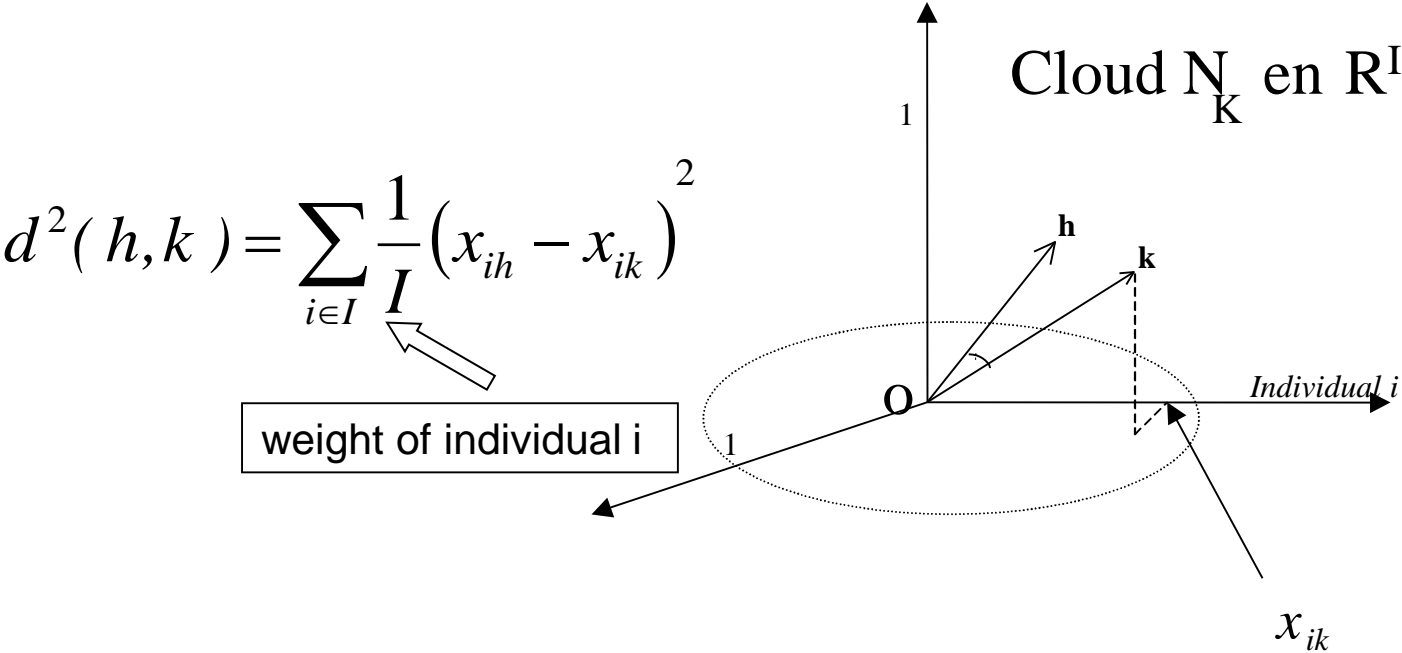
Classical Euclidean distance

$$d^2(i,l) = \sum_{k \in K} (x_{ik} - x_{lk})^2$$



## Principal component analysis

Cloud of variables

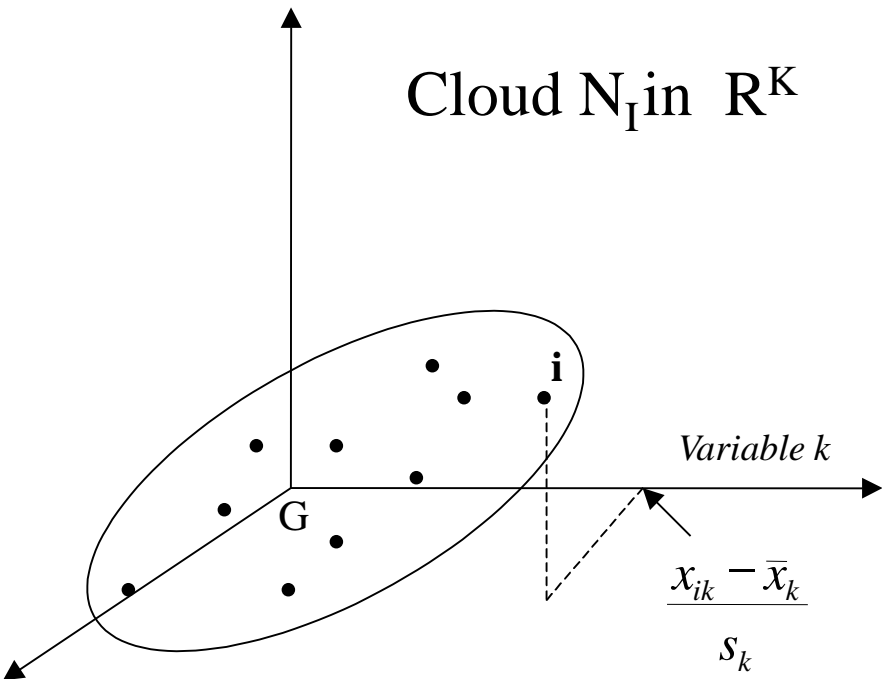
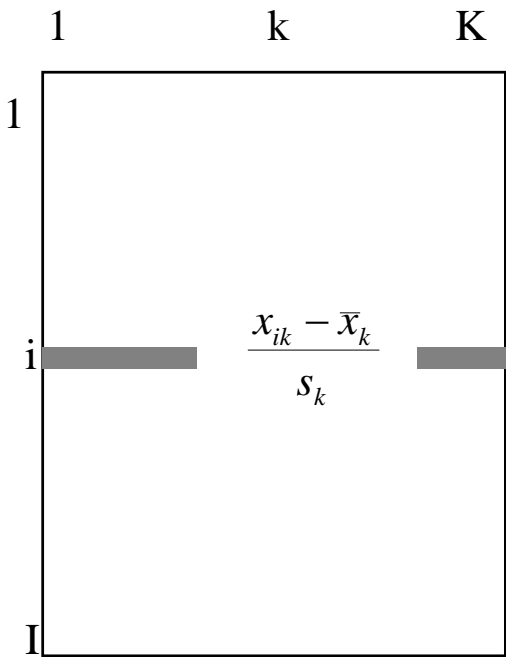


However, it is usual to analyze the relationships between variables through the linear correlation coefficient

$$r(k, h) = \frac{1}{I} \sum_{i \in I} \left( \frac{x_{ik} - \bar{x}_k}{s_k} \right) \cdot \left( \frac{x_{ih} - \bar{x}_h}{s_h} \right) = \frac{s_{k,h}}{s_k s_h}$$

In the individual space

$$d^2(i,l)=\sum_k\left(\frac{x_{ik}-x_{lk}}{s_k}\right)^2=\sum_k(z_{ik}-z_{lk})^2$$



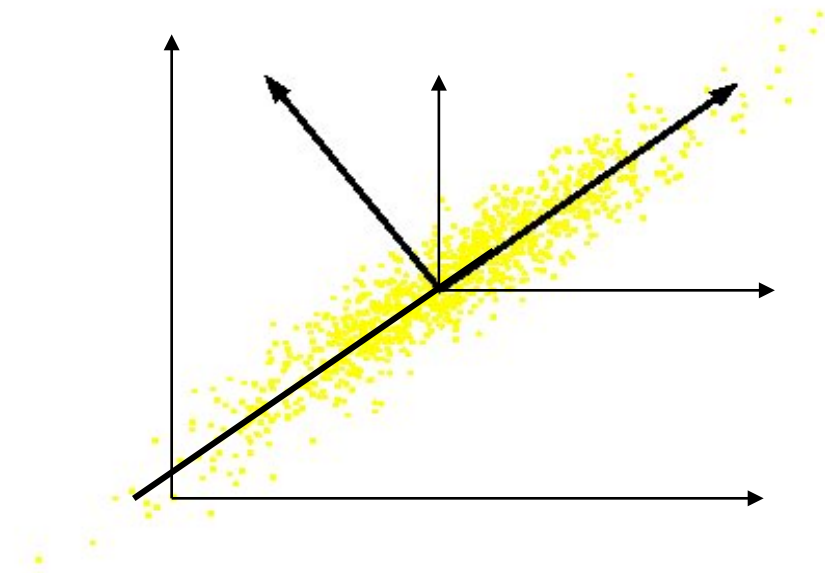
- $E((\mathbf{z} \cdot \mathbf{u})^2) = E((\mathbf{z} \cdot \mathbf{u})^T (\mathbf{z} \cdot \mathbf{u})) = E(\mathbf{u}^T \cdot \mathbf{z}^T \cdot \mathbf{z} \cdot \mathbf{u})$
- The matrix  $\mathbf{C} = \mathbf{Z}^T \cdot \mathbf{Z}$  contains the correlations (similarities) of the original axes based on how the data values project onto them
- So we are looking for  $\mathbf{w}$  that maximizes  $\mathbf{u} \mathbf{C} \mathbf{u}^T$ , subject to  $\mathbf{u}$  being unit-length
- It is maximized when  $\mathbf{w}$  is the principal eigenvector of the matrix  $\mathbf{C}$ , in which case
  - $\mathbf{u}^T \mathbf{C} \mathbf{u} = \mathbf{u}^T \lambda \mathbf{u} = \lambda$  if  $\mathbf{u}$  is unit-length, where  $\lambda$  is the principal eigenvalue of the correlation matrix  $\mathbf{C}$
  - The eigenvalue denotes the amount of variability captured along that dimension

- **Maximize**  $\mathbf{u}^T \mathbf{Z}^T \mathbf{Z} \mathbf{u}$   
s.t  $\mathbf{u}^T \mathbf{u} = 1$
- Construct Lagrangian  $\mathbf{u}^T \mathbf{Z}^T \mathbf{Z} \mathbf{u} - \lambda \mathbf{u}^T \mathbf{u}$
- Vector of partial derivatives set to zero
$$\mathbf{Z}^T \mathbf{Z} \mathbf{u} - \lambda \mathbf{u} = (\mathbf{Z}^T \mathbf{Z} - \lambda \mathbf{I}) \mathbf{u} = 0$$
- As  $\mathbf{u} \neq \mathbf{0}$  then  $\mathbf{u}$  must be an **eigenvector** of  $\mathbf{Z}^T \mathbf{Z}$  with eigenvalue  $\lambda$
- Diagonalization  $\mathbf{Z}^T \mathbf{Z} = \mathbf{U} \mathbf{\Sigma} \mathbf{U}^T$ , where  $\mathbf{\Sigma}$  diag matrix with eigenvalues

- The first root is called the principal eigenvalue which has an associated orthonormal ( $\mathbf{u}^T \mathbf{u} = 1$ ) *eigenvector*  $\mathbf{u}$
- Subsequent roots are ordered such that  $\lambda_1 > \lambda_2 > \dots > \lambda_s$  with  $\text{rank}(\mathbf{C})$  ( $S \leq \text{Min}(I, K)$ ) non-zero values.
- Eigenvectors form an orthonormal basis i.e.  $\mathbf{u}_i^T \mathbf{u}_j = \delta_{ij}$
- The eigenvalue decomposition of  $\mathbf{Z}^T \mathbf{Z} = \mathbf{U} \mathbf{\Sigma} \mathbf{U}^T$   
where  $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_s]$  s-by-s matrix whose columns are the eigenvectors of  $\mathbf{Z}^T \mathbf{Z}$  and  $\mathbf{\Sigma} = \text{diag}[\lambda_1, \lambda_2, \dots, \lambda_s]$
- Similarly the eigenvalue decomposition of  $\mathbf{Z} \mathbf{Z}^T = \mathbf{V} \mathbf{\Sigma} \mathbf{V}^T$
- The SVD is closely related to the above  $\mathbf{Z}^T = \mathbf{U} \mathbf{\Sigma}^{1/2} \mathbf{V}^T$
- The left eigenvectors  $\mathbf{U}$ , right eigenvectors  $\mathbf{V}$ ,  
singular values = square root of eigenvalues.



- So, the new axes are the eigenvectors of the matrix of correlations of the original variables, which captures the similarities of the original variables based on how data samples project to them



- Geometrically: centering followed by rotation
  - Linear transformation

- The first PC retains the greatest amount of variation in the sample
- The  $k^{\text{th}}$  PC retains the  $k^{\text{th}}$  greatest fraction of the variation in the sample
- The  $k^{\text{th}}$  largest eigenvalue of the correlation matrix  $C$  is the variance in the sample along the  $k^{\text{th}}$  PC
- **The least-squares view:** PCs are a series of linear least squares fits to a sample, each orthogonal to all previous ones

Coordinates of the individuals on  $u_s$ :  $\mathbf{F}_s = \mathbf{Z}\mathbf{u}_s$

$\mathbf{F}_s$  is:

- The vector of the coordinates of the individuals on axis  $s$  (*=one new variable*)
- a lineal combination of the original variables (coefficients = elements of  $u_s$ )
- centered with variance  $\lambda_s$

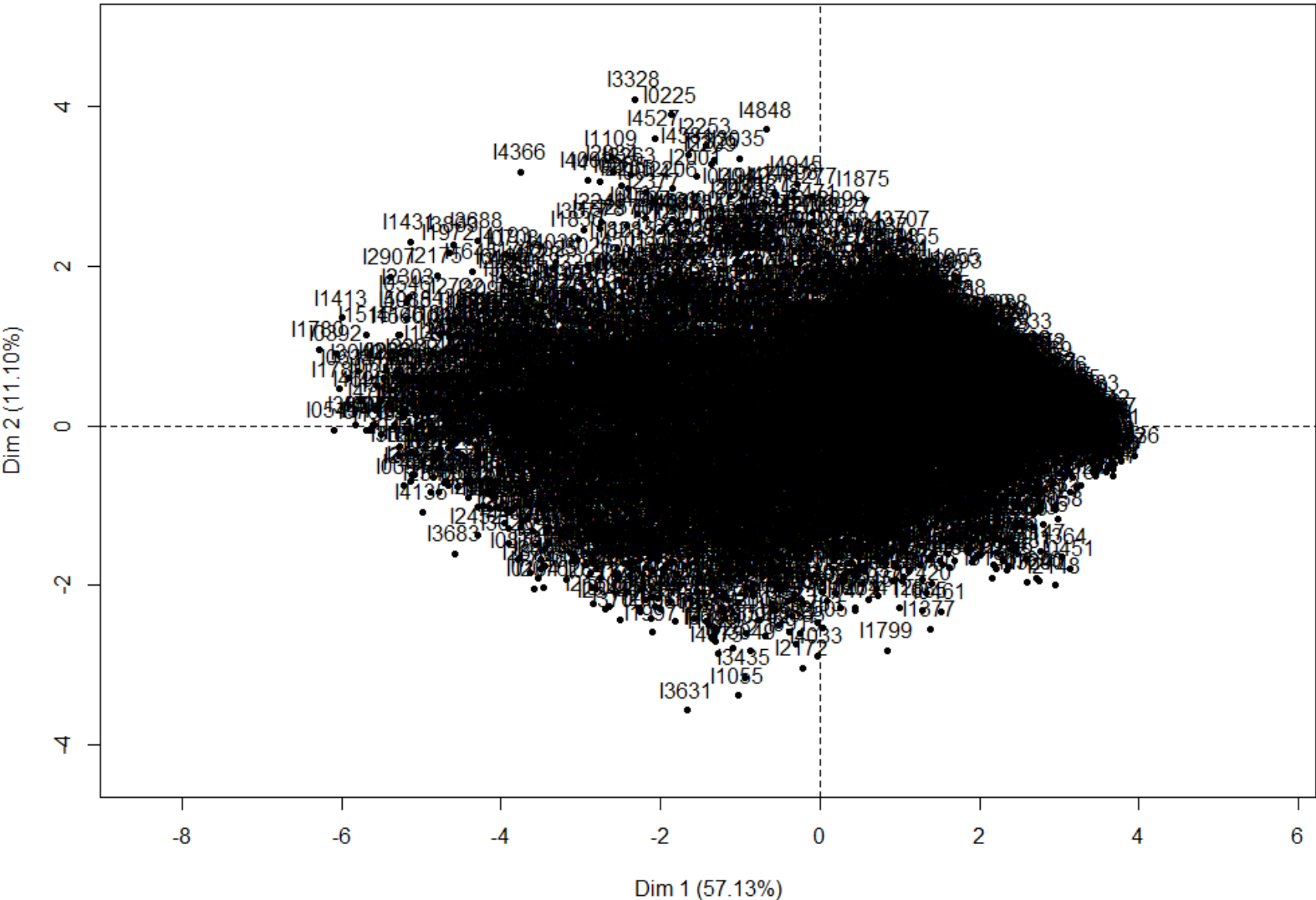
$\mathbf{F}_s$  is called the  $s_{th}$  principal component ( $1 \times 1$ )

The principal axes  $\mathbf{u}_s$  are orthogonal.

Thus a series of synthetic variables is defined, called **principal components**.

They are uncorrelated and constitute the best summary of the initial variables.

Individuals factor map (PCA)

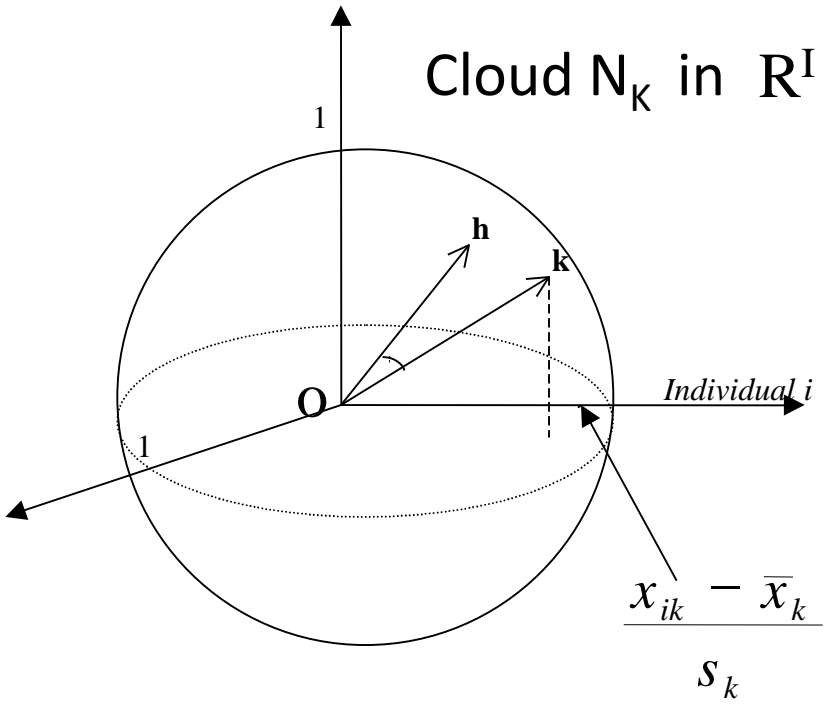
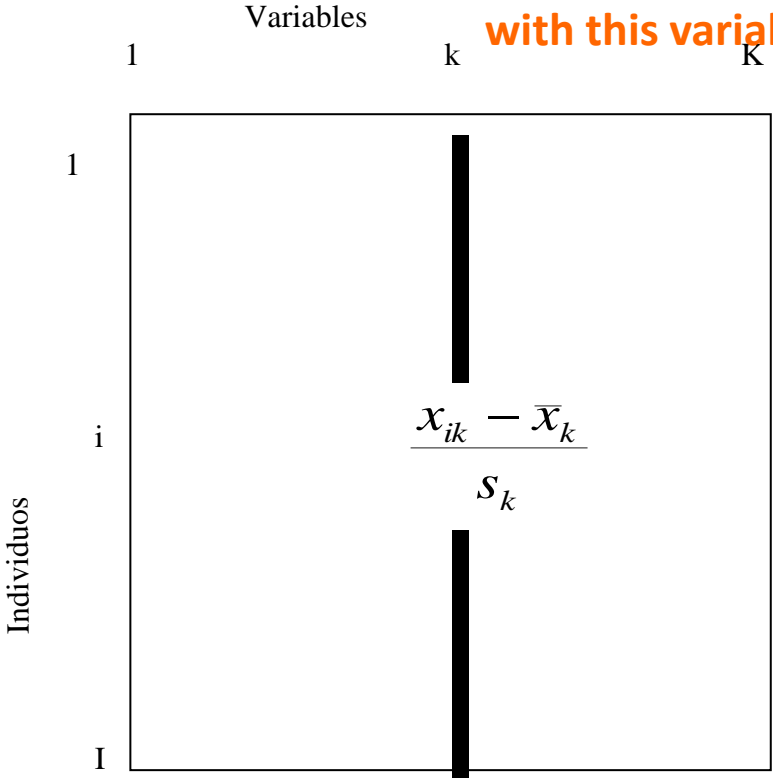


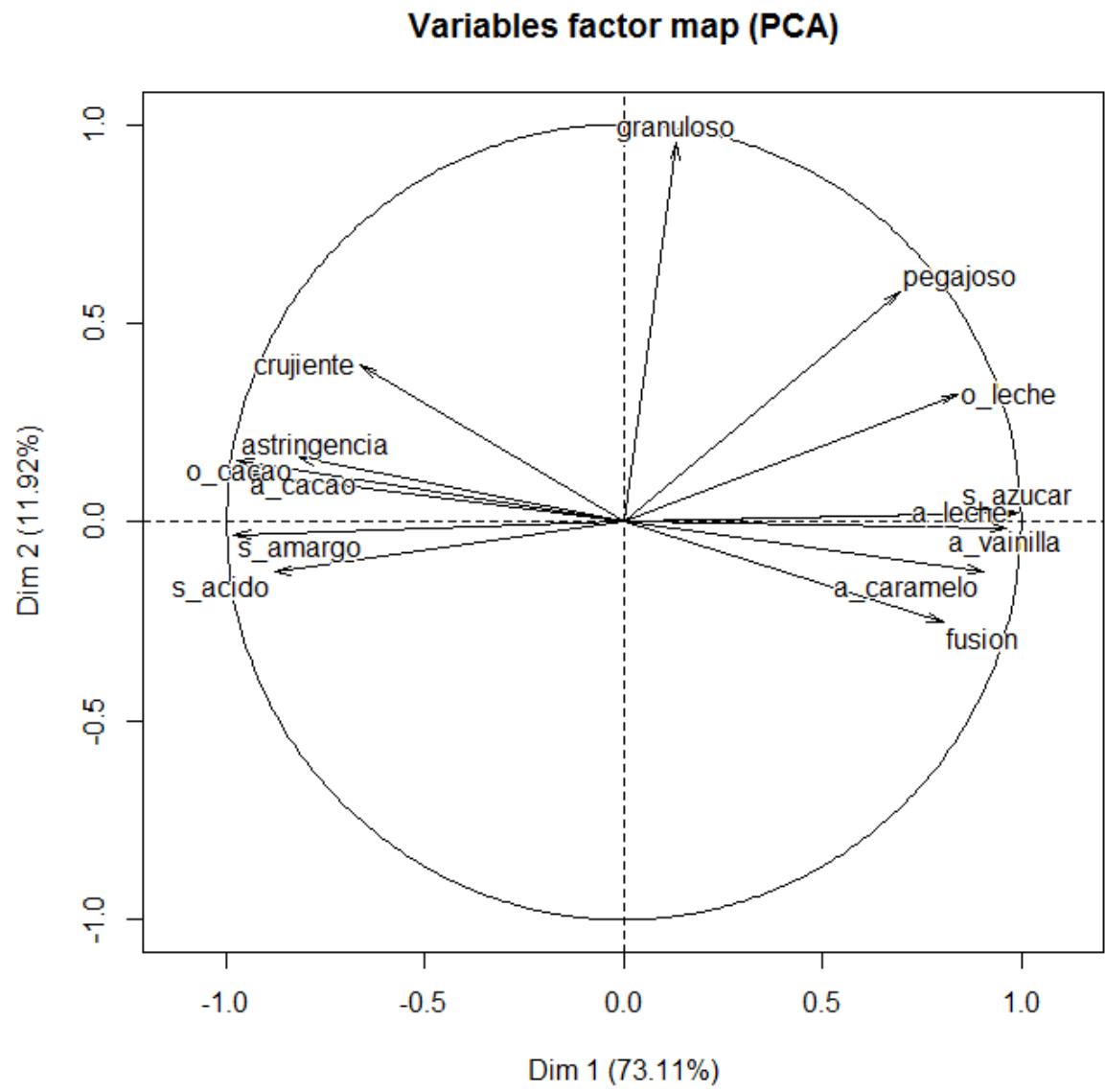
In  $R'$ , because of the variables normalization, all the variables lie on the hypersphere with radius 1 . To center the variable means:

In  $R'$ , the cosine of the angle between two vectors is equal to the correlation between the two variables.

As the variables are centered and standardized, **the projection on  $v_1$  of any variable is equal to the coefficient of correlation with this variable.**

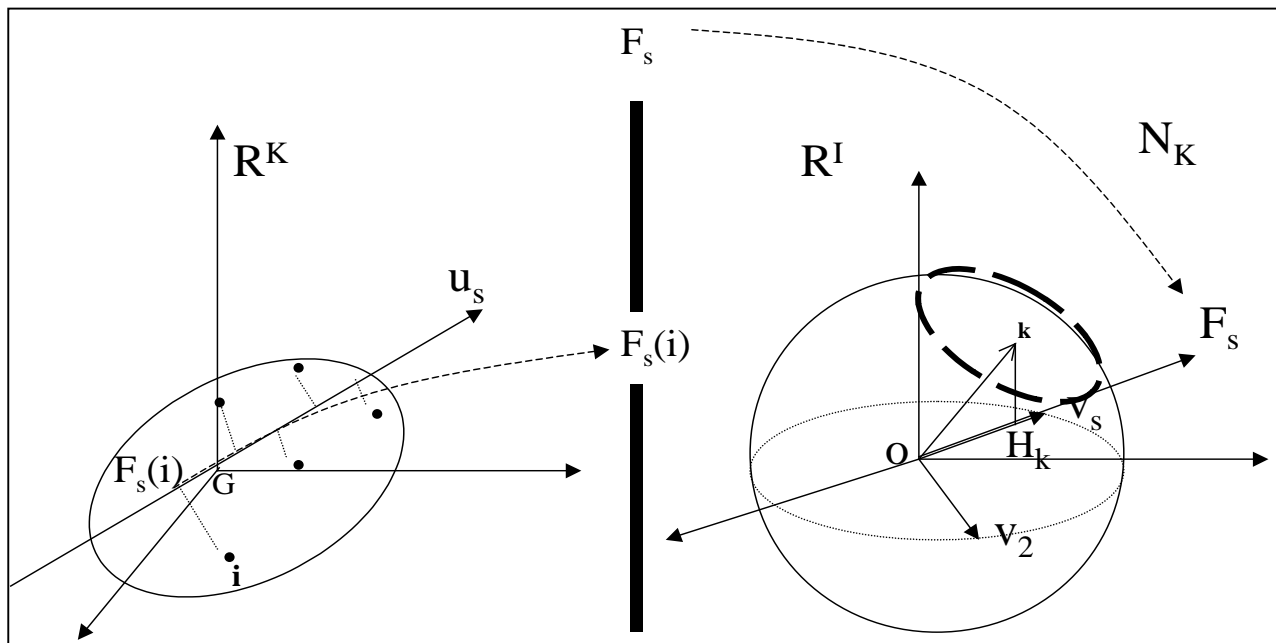
## Cloud of variables





## Principal components

Factor  $s$  or principal component  $s$  on individuals: projection of all points of the cloud of individuals on the axis  $s$  noted  $F_s$

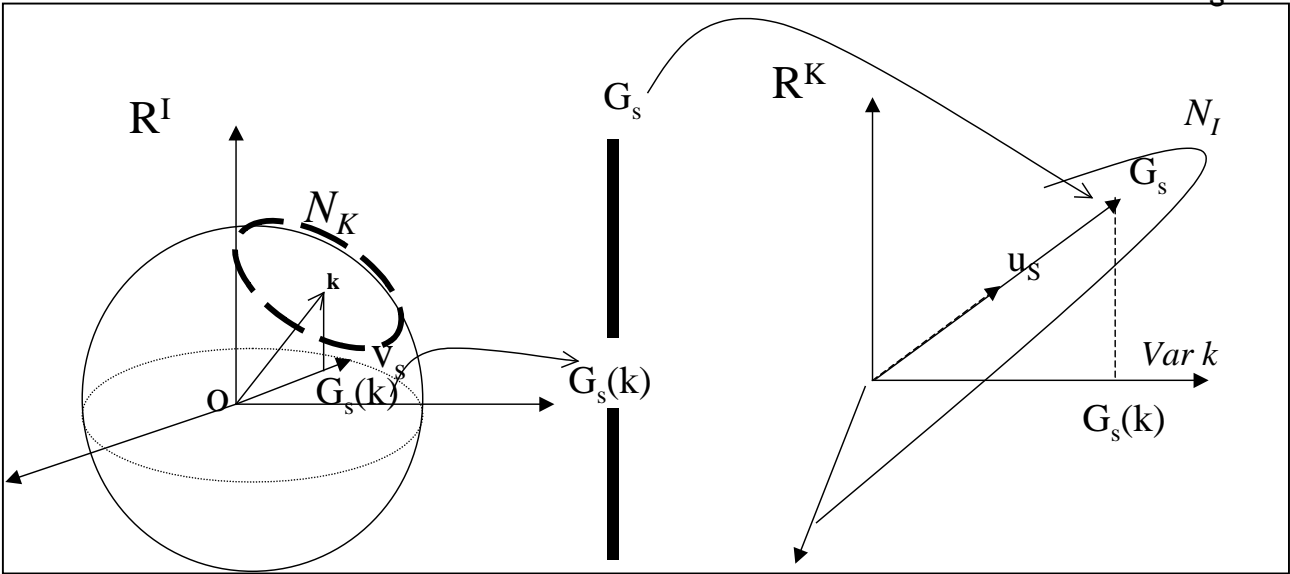


$$v_s = \frac{1}{\sqrt{\lambda_s}} F_s$$

Relationship between the factor  $F_s$  and factorial axis  $v_s$

Factor  $s$  on the variables  $G_s$ : projection of the  $K$  variables on the factorial axis

The set of values is the factor  $s$  on the variables noted  $G_s$



Relationship between axes  $u_s$  and factor  $G_s$

$$u_s = \frac{1}{\sqrt{\lambda_s}} G_s$$



## Transition relationships

They are deduced from the relationships between axes and factors:

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_k \frac{x_{ik} - \bar{x}_k}{s_k} G_s(k)$$

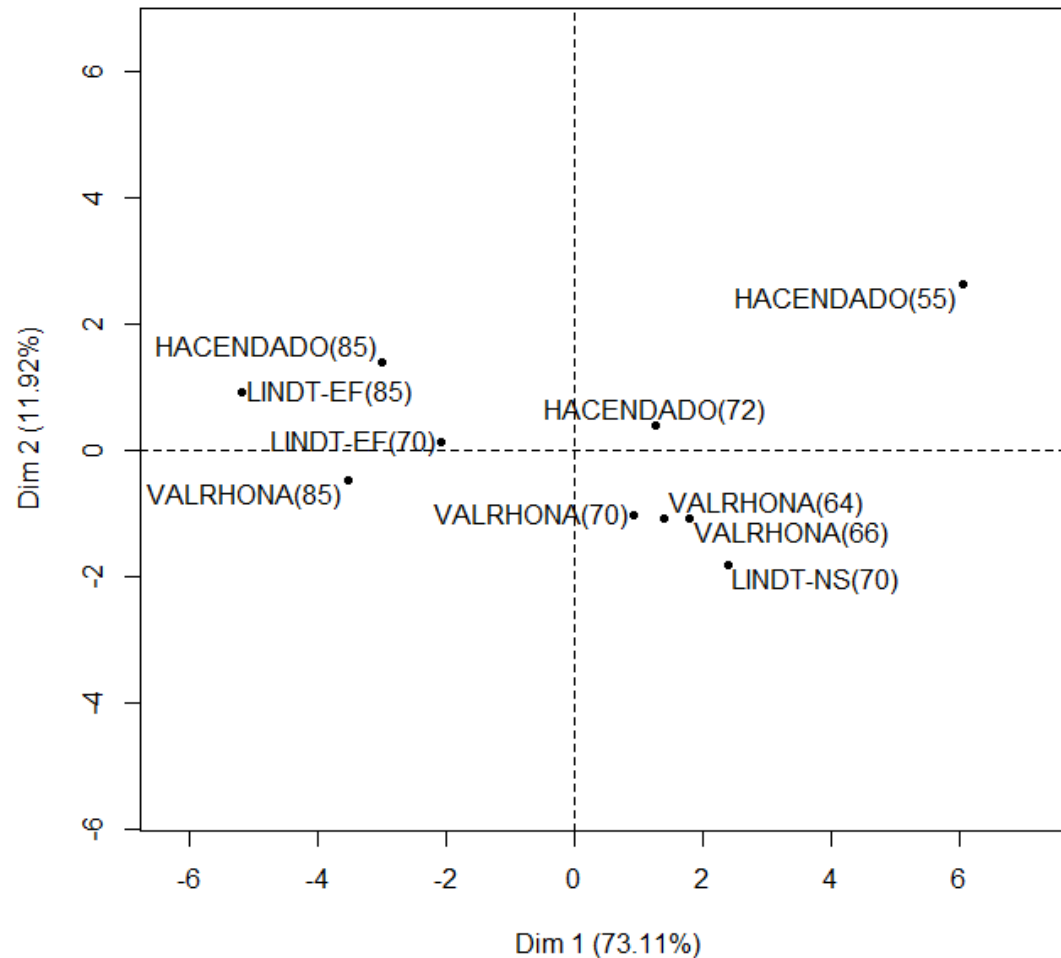
$$G_s(k) = \frac{1}{I} \frac{1}{\sqrt{\lambda_s}} \sum_i \frac{x_{ik} - \bar{x}_k}{s_k} F_s(i)$$

In practice, they are computed :

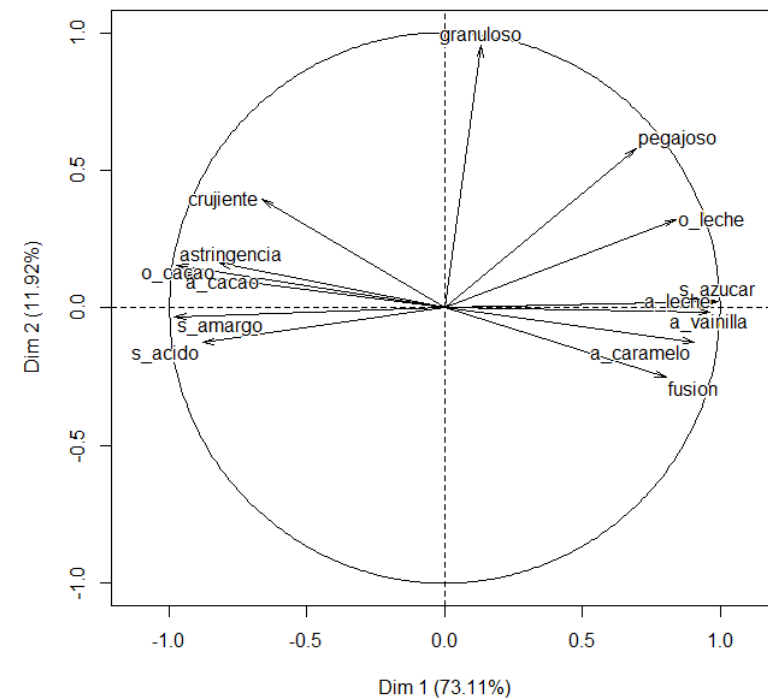
$$G_s(k) = \sqrt{\lambda_s} u_{sk}$$

$$G_s(k) = \text{corr}(k, F_s)$$

Individuals factor map (PCA)



Variables factor map (PCA)



## Principal component analysis

### *Helps to interpretation*

Quality representation of the projection of the cloud on an axis, a plane, etc.

$$\frac{\sum_{s=1}^q \lambda_s}{\sum_{s=1}^K \lambda_s}$$

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	4.57	57.13	57.13
comp 2	0.89	11.10	68.23

Quality of representation of an element on an axis:

$$QLT_s(i) = \frac{(OH_i^s)^2}{(Oi)^2} = \cos^2 \theta$$

Contribution of a row-element to the inertia of an axis  $\frac{1}{I} \cdot \frac{F_s^2(i)}{\lambda_s}$

Contribution of a column-element to the inertia of an axis  $\frac{G_s^2(k)}{\lambda_s}$

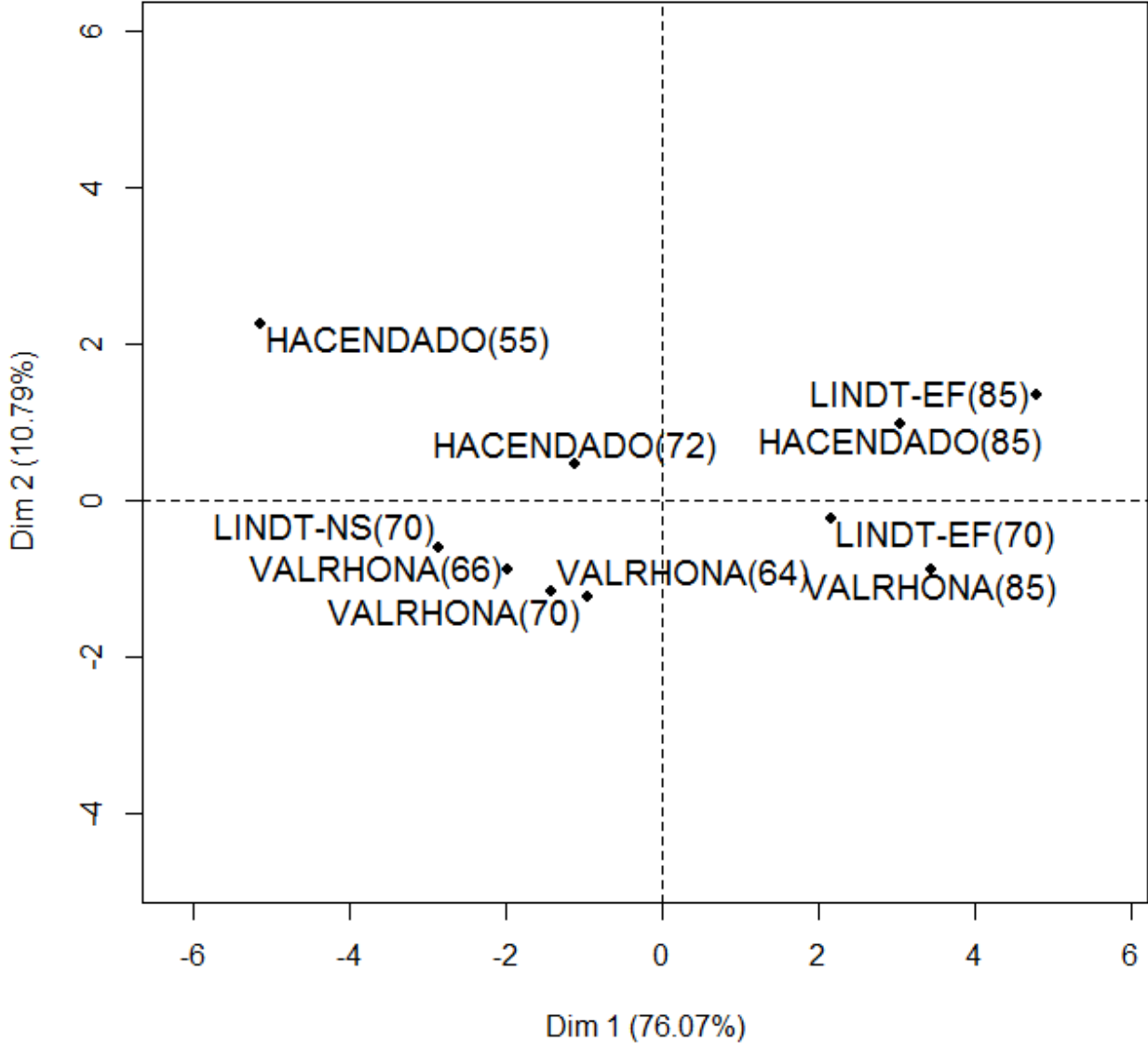
## Individuals

	Dist	Dim.1	ctr	cos2	Dim.2	ctr	cos2
VALRHONA(85)	3.974	-3.536	12.212	0.792	-0.461	1.276	0.013
VALRHONA(70)	2.028	0.919	0.824	0.205	-1.030	6.354	0.258
VALRHONA(66)	2.530	1.792	3.137	0.502	-1.087	7.082	0.185
VALRHONA(64)	2.161	1.401	1.918	0.420	-1.086	7.064	0.252
LINDT-NS(70)	3.548	2.390	5.581	0.454	-1.803	19.478	0.258
LINDT-EF(85)	5.477	-5.176	26.178	0.893	0.923	5.104	0.028
LINDT-EF(70)	2.878	-2.083	4.239	0.524	0.141	0.119	0.002
HACENDADO(85)	3.472	-2.998	8.783	0.746	1.383	11.458	0.159
HACENDADO(72)	1.603	1.256	1.542	0.614	0.401	0.965	0.063
HACENDADO(55)	6.598	6.035	35.586	0.837	2.619	41.100	0.158

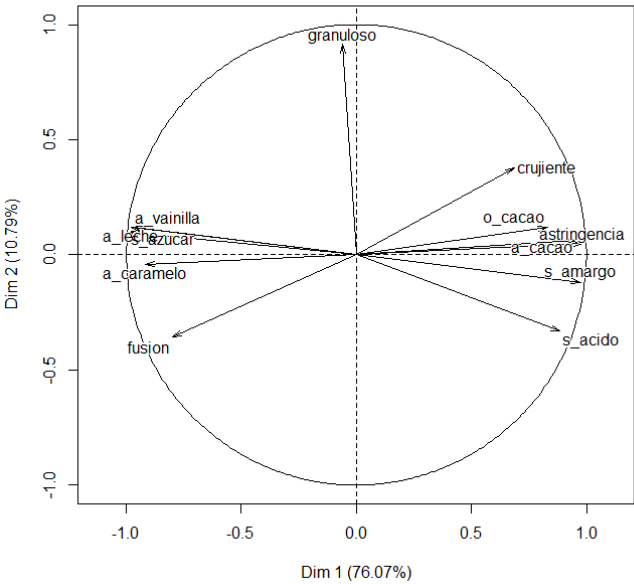
## Variables (the 10 first)

	Dim.1	ctr	cos2	Dim.2	ctr	cos2
o_cacao	-0.818	6.536	0.669	0.162	1.568	0.026
o_leche	0.841	6.904	0.707	0.323	6.236	0.104
s_azucar	0.993	9.624	0.985	0.022	0.029	0.000
s_acido	-0.878	7.534	0.771	-0.127	0.966	0.016
s_amargo	-0.980	9.378	0.960	-0.033	0.065	0.001
a_cacao	-0.949	8.807	0.901	0.126	0.949	0.016
a_leche	0.983	9.431	0.965	0.025	0.037	0.001
a_caramelo	0.907	8.041	0.823	-0.124	0.926	0.015
a_vainilla	0.962	9.036	0.925	-0.014	0.011	0.000
astringencia	-0.971	9.221	0.944	0.155	1.440	0.024

Individuals factor map (PCA)



Variables factor map (PCA)



## Supplementary individuals and variables

- In order to project a supplementary or illustrative individual: its values for each of the variables, but centered (centered or standardized) are computed. Then the transition relationships are used to place it on every axis.

$$F_s(i^+) = \frac{1}{\sqrt{\lambda_s}} \sum_k \frac{x_{i^+k} - \bar{x}_k}{s_k} G_s(k)$$

- A supplementary quantitative variable is placed on the axes through its correlations with the principal components.

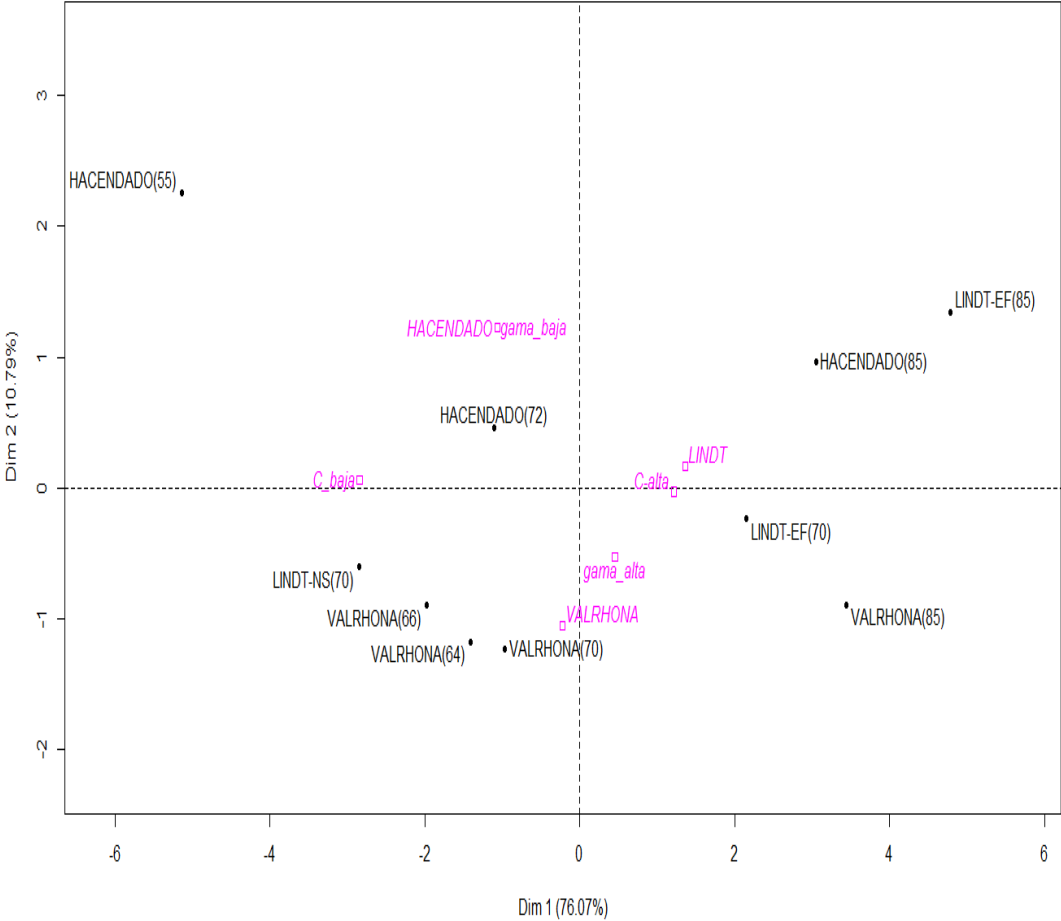
$$G_s(k^+) = \text{corr}(k^+, F_s)$$

- A category of a **supplementary categorical variable** (with  $m$  categories) is placed at the centroid of the individuals that belong to this category.

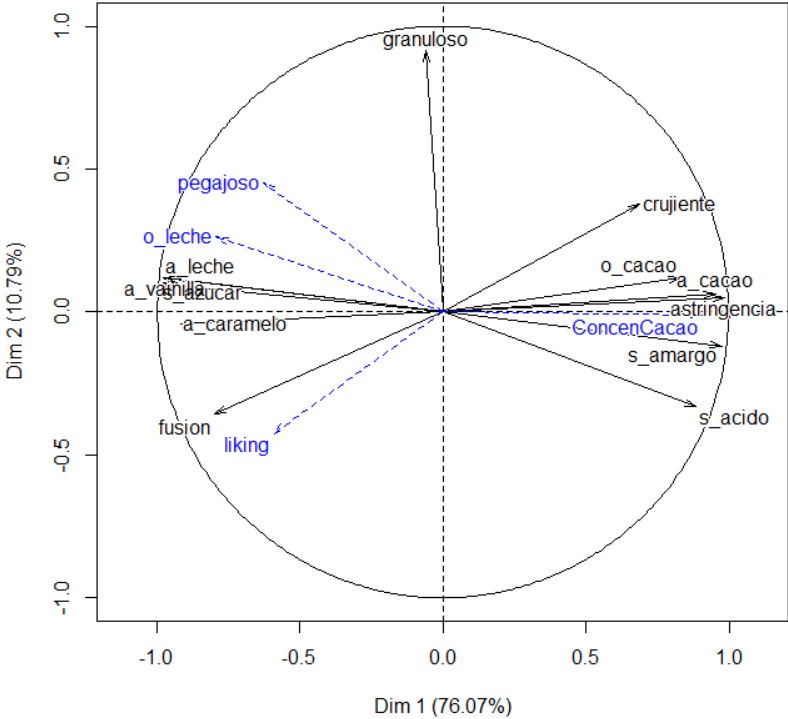
- Separate representation of both clouds
- The two clouds are not in the same space; They do not have the same referential
- Similarities between individuals are interpreted as corresponding to similar behavior as far as the active variables are concerned
- Proximities between variables are interpreted as correlations



Individuals factor map (PCA)



Variables factor map (PCA)



The total inertia of both clouds is the same

$$Inertia = \frac{1}{I} \sum_k \sum_i \left( \frac{x_{ik} - \bar{x}}{s_k} \right)^2 = \text{sum of variances}$$

Inertia total= nr of variables when the variables are standardized

In general, the inertia is equal to

- the sum of variances of the variables
- the sum of the trace of the variance-covariance matrix

Thus, this analysis performs a decomposition of the total inertia equivalent in both spaces. The inertia projected onto the same rank axis are equal.

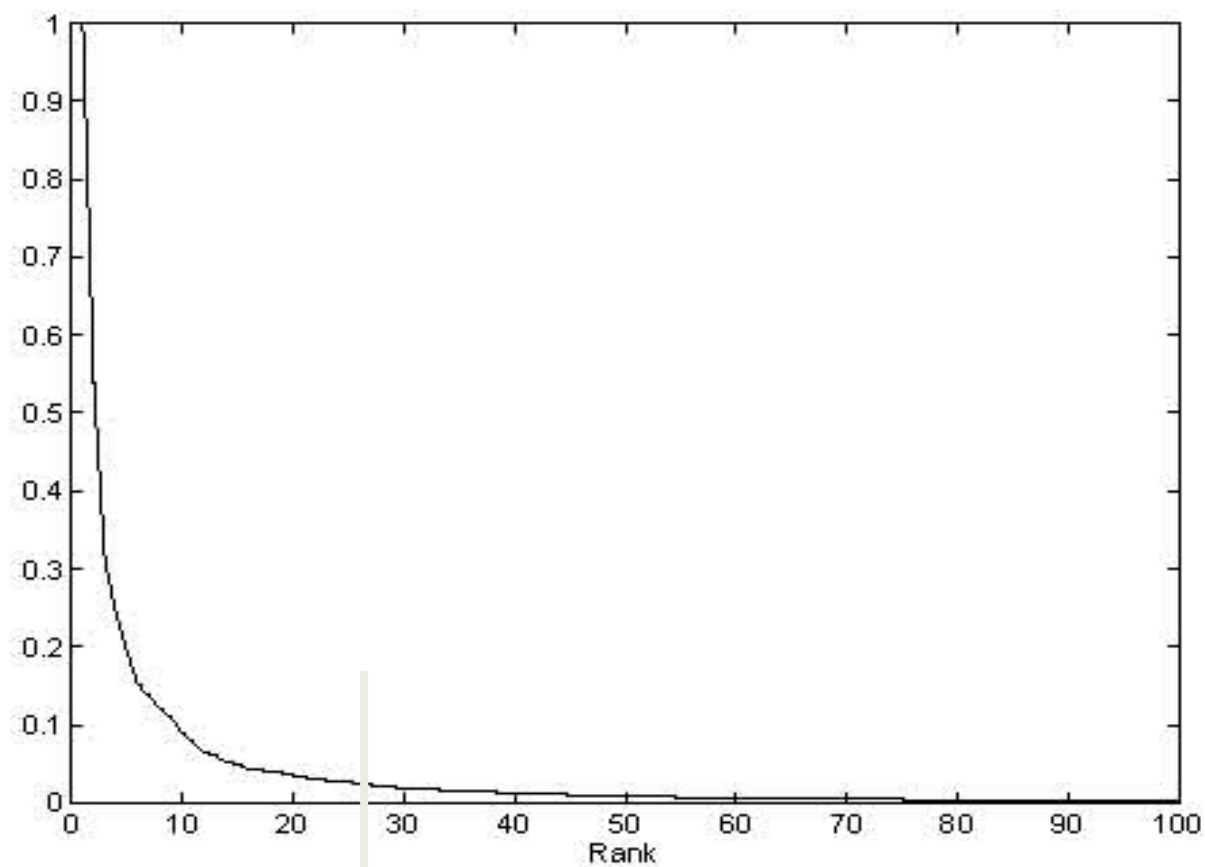
- For  $K$  original dimensions, correlation matrix is  $K \times K$ , and has up to  $K$  eigenvectors. So  $K$  PCs.
- Where does dimensionality reduction come from?

- How many principal components to keep?
  - To choose  $K$ , the following criterion can be used:

$$\frac{\sum_{i=1}^K \lambda_i}{\sum_{i=1}^N \lambda_i} > \textit{Threshold} \quad (\text{e.g., } 0.9 \text{ or } 0.95)$$

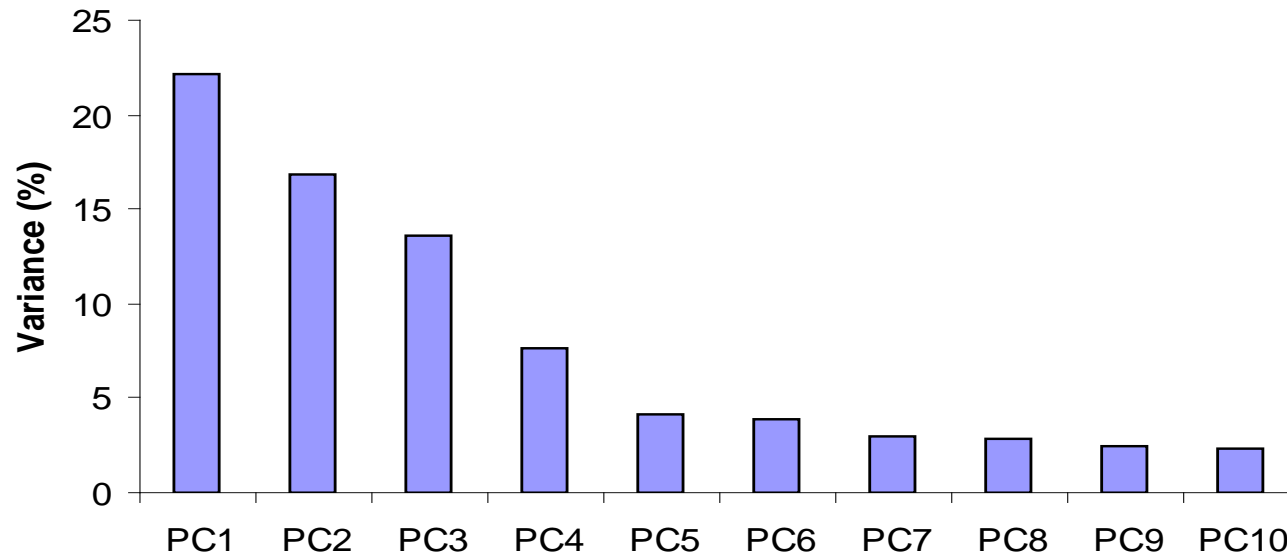
- Unfortunately for some data sets to meet this requirement we need  $K$  almost equal to  $I$ . That is, no effective data reduction is possible.

- Eigenvalue spectrum



Scree plot

Can *ignore* the components of lesser significance.



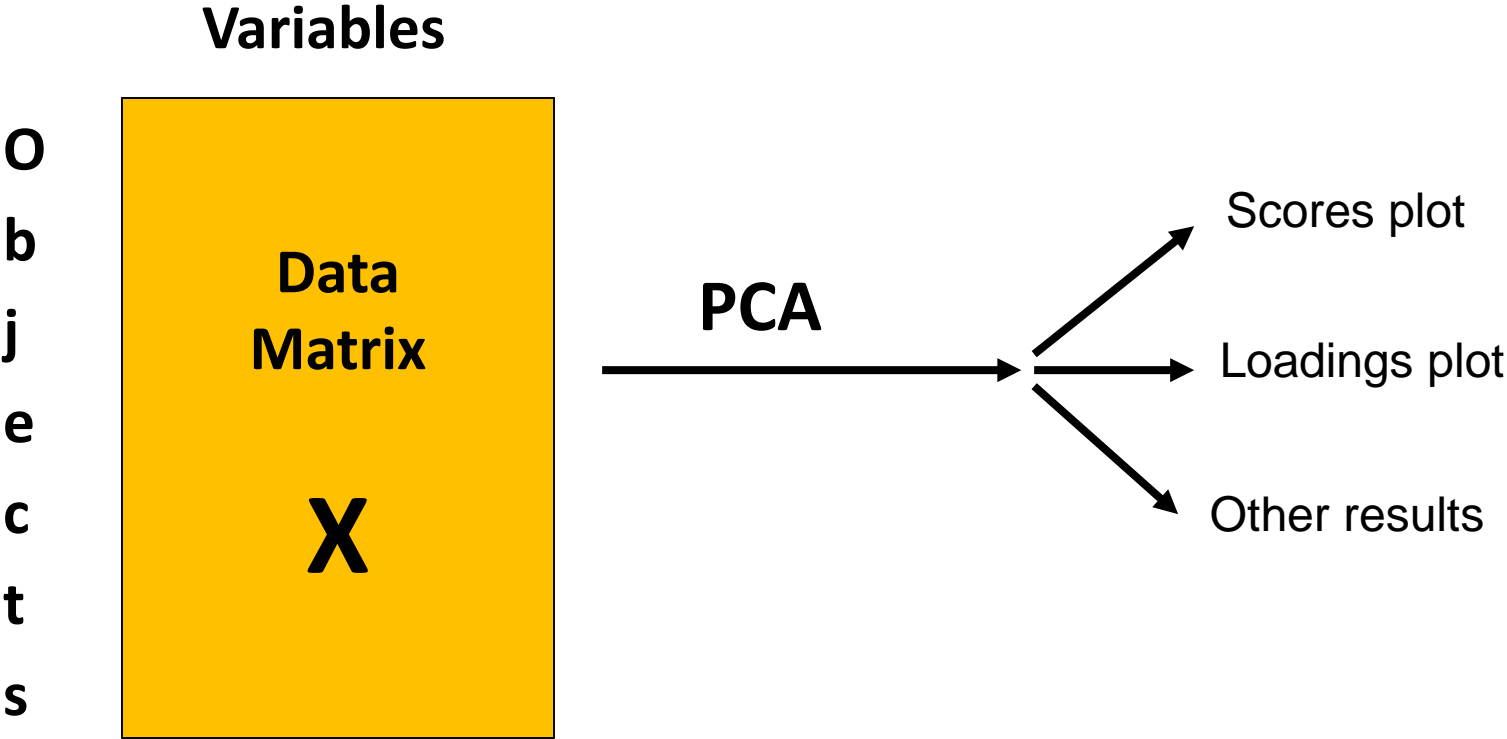
You do *lose some information*, but if the eigenvalues are small, you don't lose much

- *K* dimensions in original data
- calculate *K* eigenvectors and eigenvalues
- choose only the first *p* eigenvectors, based on their eigenvalues
- final data set has only *p* dimensions

- Probably the most widely-used and well-known of the “standard” multivariate methods
- Invented by Pearson (1901) and Hotelling (1933)
- First applied in ecology by Goodall (1954) under the name “factor analysis” (“principal factor analysis” is a synonym of PCA).

- PCA takes a data matrix of  $I$  objects by  $K$  variables, which may be correlated, and summarizes it by uncorrelated axes (principal components or principal axes) that are linear combinations of the original  $K$  variables
- The first  $k$  components display as much as possible of the variation among objects.





- **PCA Model :  $X=TP^T + E$**

The matrix  $X$  is modelled as components (systematic effects) plus residuals,  $E$  (noise)

- **Scores plot**

- For interpreting relations among samples

- **Loadings plot**

- For interpreting relations among variables

- **Explained variance plot**

- **Addition of supplementary variables is very useful for interpretation**

- Usually 2-dimensions are shown, firstly PC1 and PC2.
- Higher dimension PCs,  $(PC_k, PC_k')$  relationships have to be checked.
- For spectroscopy and other continuous measurements, 1-dimensional plots are used.

- Variables which are close have high correlation
- Samples which are close are similar
- Variables on opposite side of origin have negative correlation
- Objects on the right are dominated by variables to the right and so on....

- PCA
  - finds orthonormal basis for data
  - Sorts dimensions in order of “importance”
  - Discard low significance dimensions
- Uses:
  - Get compact description
  - Ignore noise
  - Improve classification (hopefully)
- Not magic:
  - Doesn’t know class labels
  - Can only capture linear variations
- One of many tricks to reduce dimensionality!