Data analysis and stochastic modeling

Lecture 2 – Descriptive and exploratory statistics

Guillaume Gravier

guillaume.gravier@irisa.fr

What are we here for?

1. data from observations
   - see what the data looks like
   - describe the data: distribution, clustering, etc.
   - summarize the data

2. models for decision
   - infer more general properties
   - make a (stochastic) model of the data
   - make decisions: classification, simulation, etc.

$\Rightarrow$ Provide the elementary tools and techniques

What are we gonna talk about today?

- Representing and viewing data
  - $\rightarrow$ tables and graphics
- Describing 1D data
  - $\rightarrow$ mean, median, standard deviation, quartiles, mode, etc.
- Measuring the relation between variables
  - $\rightarrow$ correlations
- Exploring multidimensional data
  - $\rightarrow$ principal component analysis, correspondence analysis, factor analysis, etc.

In short: have a feeling for a distribution, describe a distribution, identify clusters, identify important factors.

What data and where from?

Data come from **various sources**: Physical measures, experimental results, descriptive features, etc.

1. they happen to be here
   $\rightarrow$ may not be representative
2. you design their collect
   $\rightarrow$ sample representative data
   $\rightarrow$ database design

Data come in **various flavors**

- categorical: ordered or no, coded or not
- numerical (sum has a meaning)
- scalar or not

DESCRIBING 1-DIM DATA

Representing 1D data

From a single variable X observed on n samples, we want to

- describe the variable
- summarize the information

Data are usually organized as tables.

Example. Number of suicides per year and per state observed in 14 states over 14 years [Source: Saporta, reporting Von Bortkiewicz 1898])

Nombre de suicides x_i	0	1	2	3	4	5	6	7	8	9	$\geqslant 10$
Effectif n_i	9	19	17	20	15	11	8	2	3	5	3
Total $n = 112$											

Representing 1D data (cont'd)

→ for continuous data, group into classes!

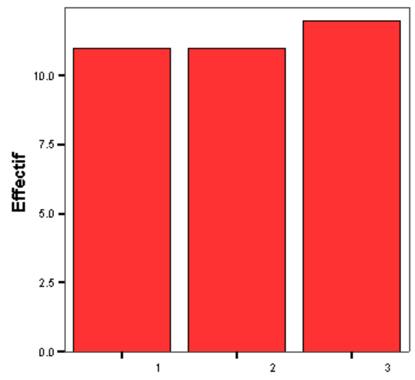
Tranche des revenus en francs	$\ln(R - 2\,500)$	% du nombre total de contribuables	% cumulés
2 500			
		0.67	
5 000	3.39		0.67
		30.18	
10 000	3.87		30.85
		27.50	
15 000	4.10		58.35
		17.09	
20 000	4.24		75.44
		14.45	
30 000	4.44		39.89
		7.01	
50 000	4.68		96.90
		1.66	
70 000	4.83		98.56
		0.81	
100 000	4.99		99.37
		0.51	
200 000	5.30		99.88
		0.10	
400 000	5.60		99.98
		0.02	
			100

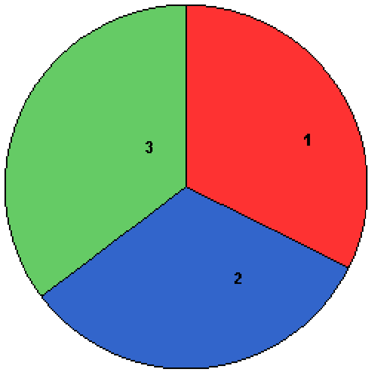
[Source: Saporta 2002, p. 117]

Viewing empirical frequencies

Categorical data

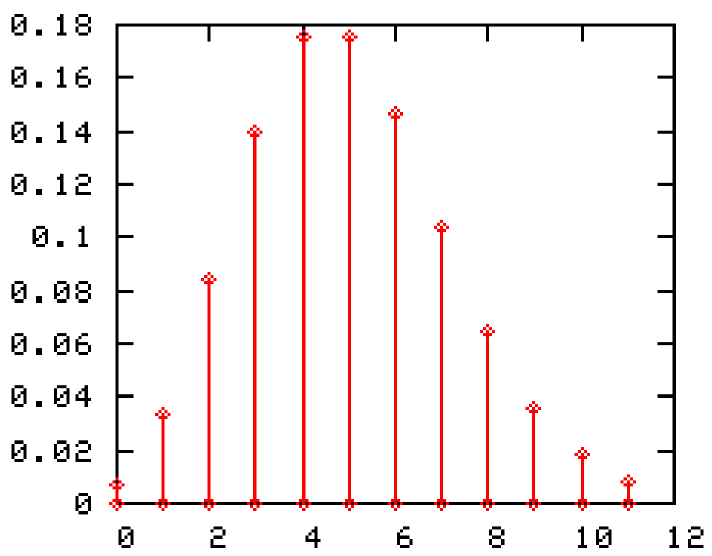
- bar graph
- pie chart





Discrete data

- empirical distribution



Empirical distributions: Histograms

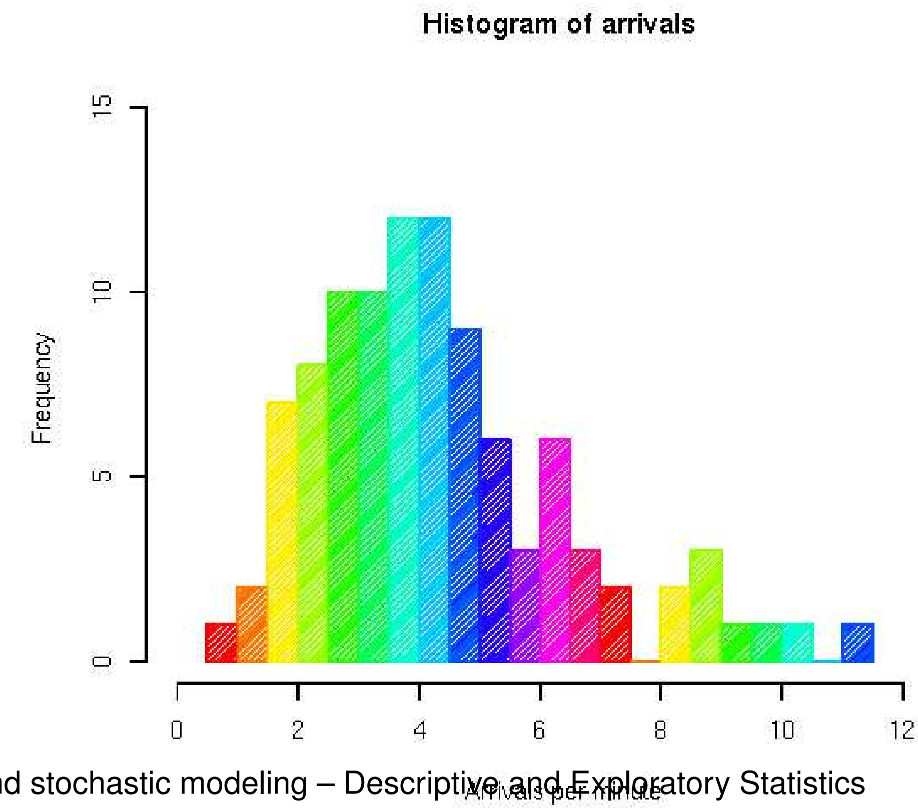
Histograms are used to **display the empirical distribution** of the variable so as to

- have an idea of the underlying distribution
- check the behavior of the data
  - → outliers, number of modes, etc.

but

- how many classes?
- equal class amplitudes or not?
  - → if not then how?

Note: The <u>area</u> of each rectangle is proportional to f_i.



Empirical distributions: Histograms (cont'd)

Smoother histograms can be obtained from wiser techniques:

- Use of a sliding window

  $\rightarrow$ count the population in all intervals $[x - \frac{\Delta}{2}, x + \frac{\Delta}{2}[$

- possibly with a kernel to weight differently samples in the interval

$$f(x) = \frac{1}{n\Delta} \sum_{i=1}^{n} K\left(\frac{x - x_i}{\Delta}\right)$$

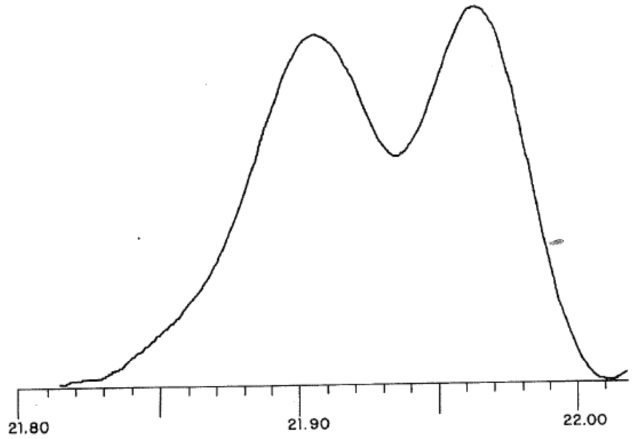
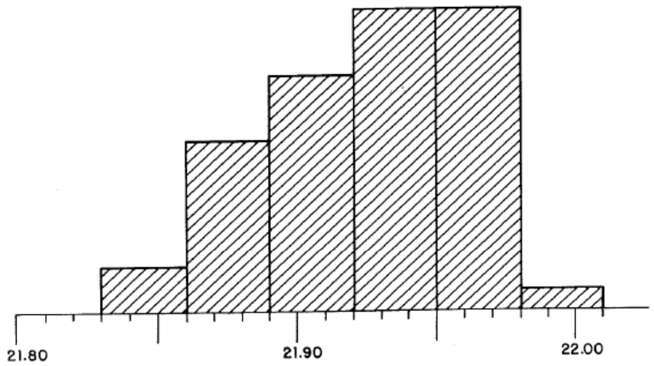


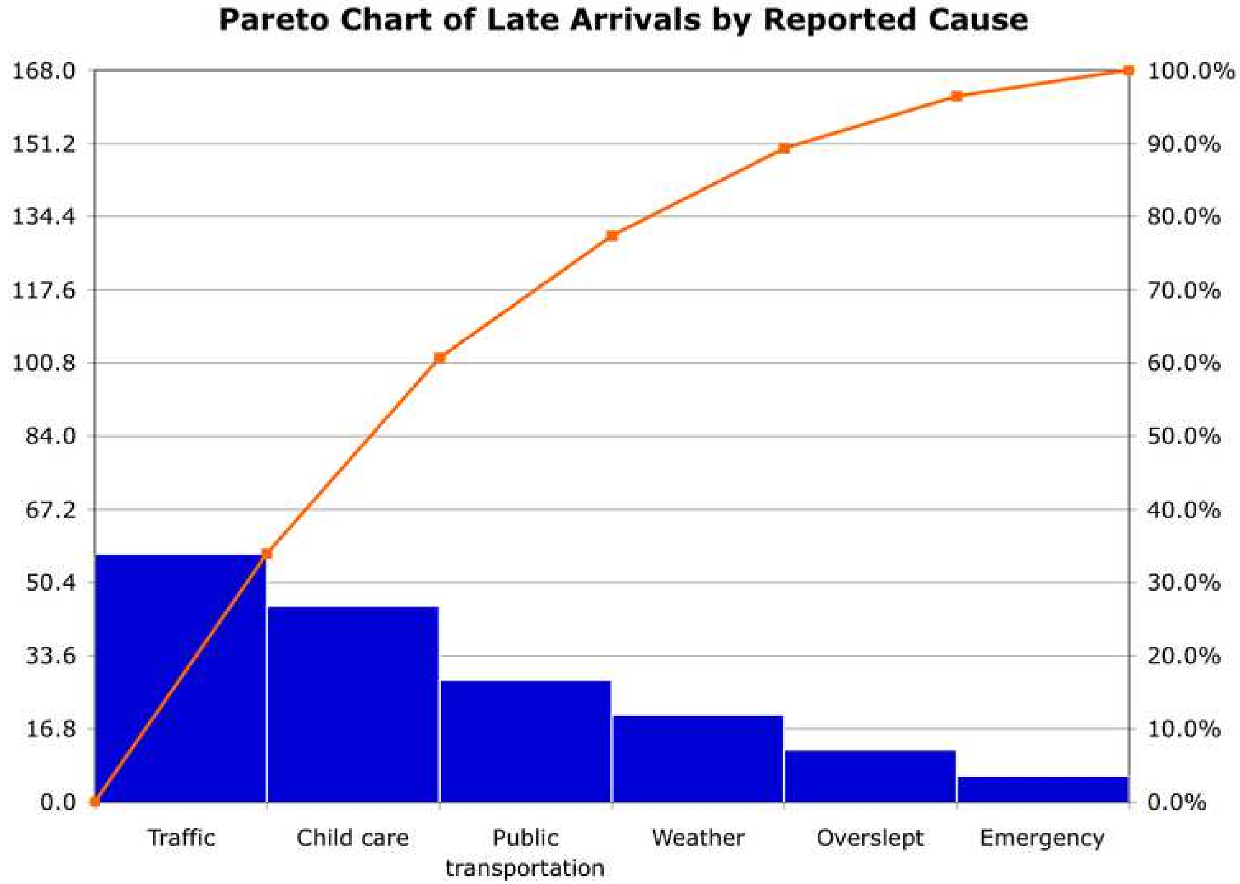
FIG. 6.3



Empirical distributions: Stem and leaf plots

[ⓒEliazar]

Empirical distributions: Pareto diagrams

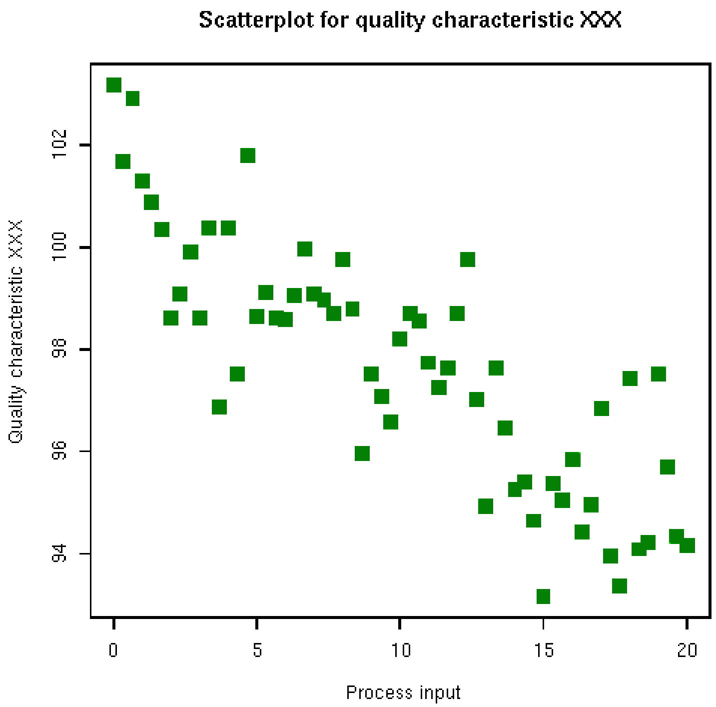


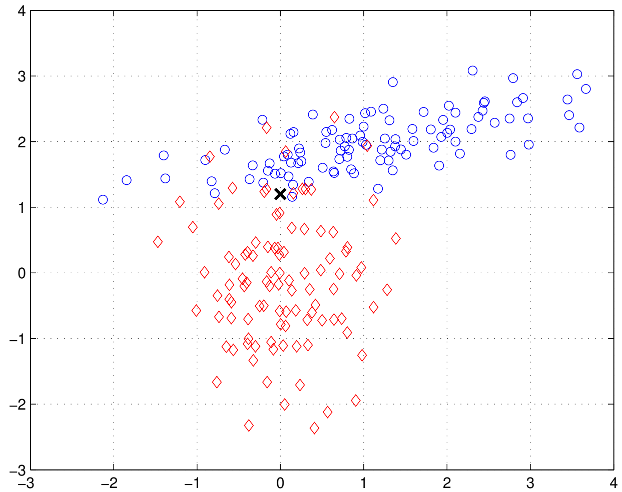
[Illustration: Metacomet (wikipedia)]

→ highlight the most important factors (e.g., main source of defects, the most frequent reasons for customer complaints. etc.)

→ aka 20 % of the causes generates 80 % of the outcome

Scatter plots for numerical variables





[Illustration: Metacomet (wikipedia)]

Numerical summaries

It is often practical to describe a distribution by a few numbers summarizing

1. central characteristics

   $\longrightarrow$ mean, median, mode, *etc*.

2. deviation around the central point

   $\longrightarrow$ extrema, standard deviation, quantiles, *etc*.

3. overall shape

   $\longrightarrow$ skewness, kurtosis, *etc*.

A single value is <u>not</u> sufficient to describe a distribution!

Yule's condition

A good statistical summary should

- be defined objectively
- be dependent on all the observations
- have a concrete and clear meaning
- be simple to compute
- be insensitive to sampling fluctuations
- be easily handled and support algebraic transformations

Georges U. Yule

1871–1951

Central characteristics of a variable

- **empirical** mean ... but sensitive to outliers!
- α**-truncated mean**: empirical mean after discarding the (α %) extremum value

$$\cancel{x_1 \leq x_2} \leq \underbrace{x_3 \leq \ldots \leq x_{20} \leq x_{21} \leq \ldots \leq x_{38}}_{\text{arithmetic mean}} \leq \cancel{x_{39} \leq x_{40}}$$

- **median**: value of the middle sample after sorting

$$x_1 \leq x_2 \leq x_3 \leq \ldots \leq \underbrace{x_{20} \leq x_{21}} \leq \ldots \leq x_{38} \leq x_{39} \leq x_{40}$$

$$\overline{X} = \frac{x_{20} + x_{21}}{2}$$

- **mode**: local extremum of the histogram

For perfectly symmetric distributions, mean = median = mode.

These statistics are not to be confused with the theoretical expectations!

Dispersion and shape of a variable

Fortunately, not all individuals are the same and, hence, mean isn't everything!

Dispersion

- minimum, maximum and range
- variance and standard deviation
- quantiles
  - bounds of the intervals dividing the data in equal parts

$$x_1 \leq \ldots \leq \underbrace{x_{10}}_{Q_1} \leq x_11 \leq \ldots \leq \underbrace{x_{20}}_{Q_2} \leq x_{21} \leq \ldots \leq \underbrace{x_{30}}_{Q_3} \leq x_{31} \leq \ldots \leq x_{40}$$

  $\rightarrow$ median (2), quartile (4), deciles (10), percentile (100)

  - interquartile range IQR $= Q_3 - Q1$

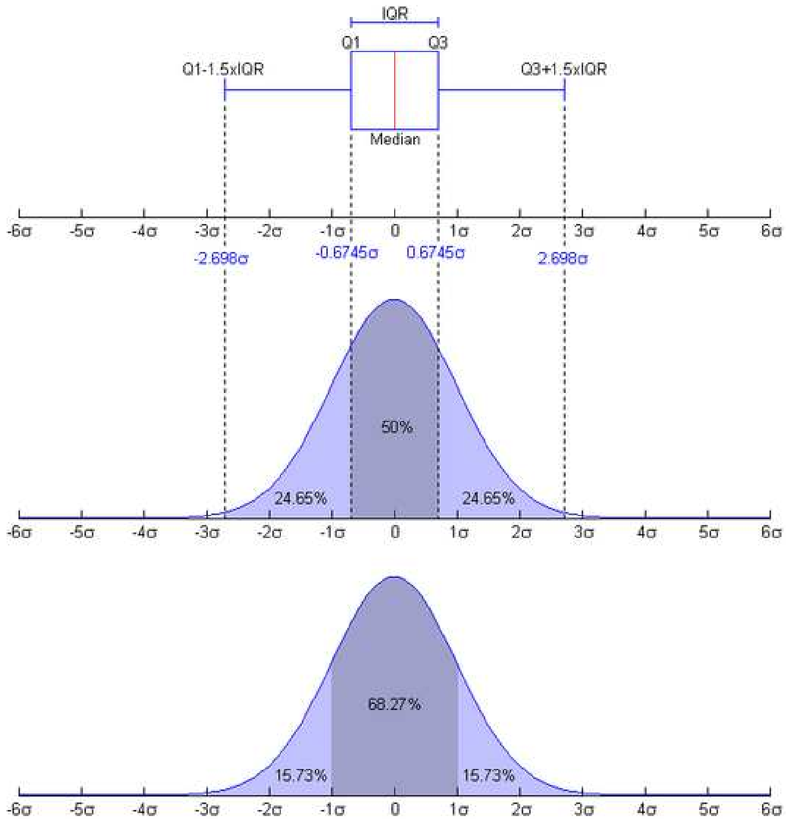
Shape

- skewness and kurtosis

Box and whisker plots

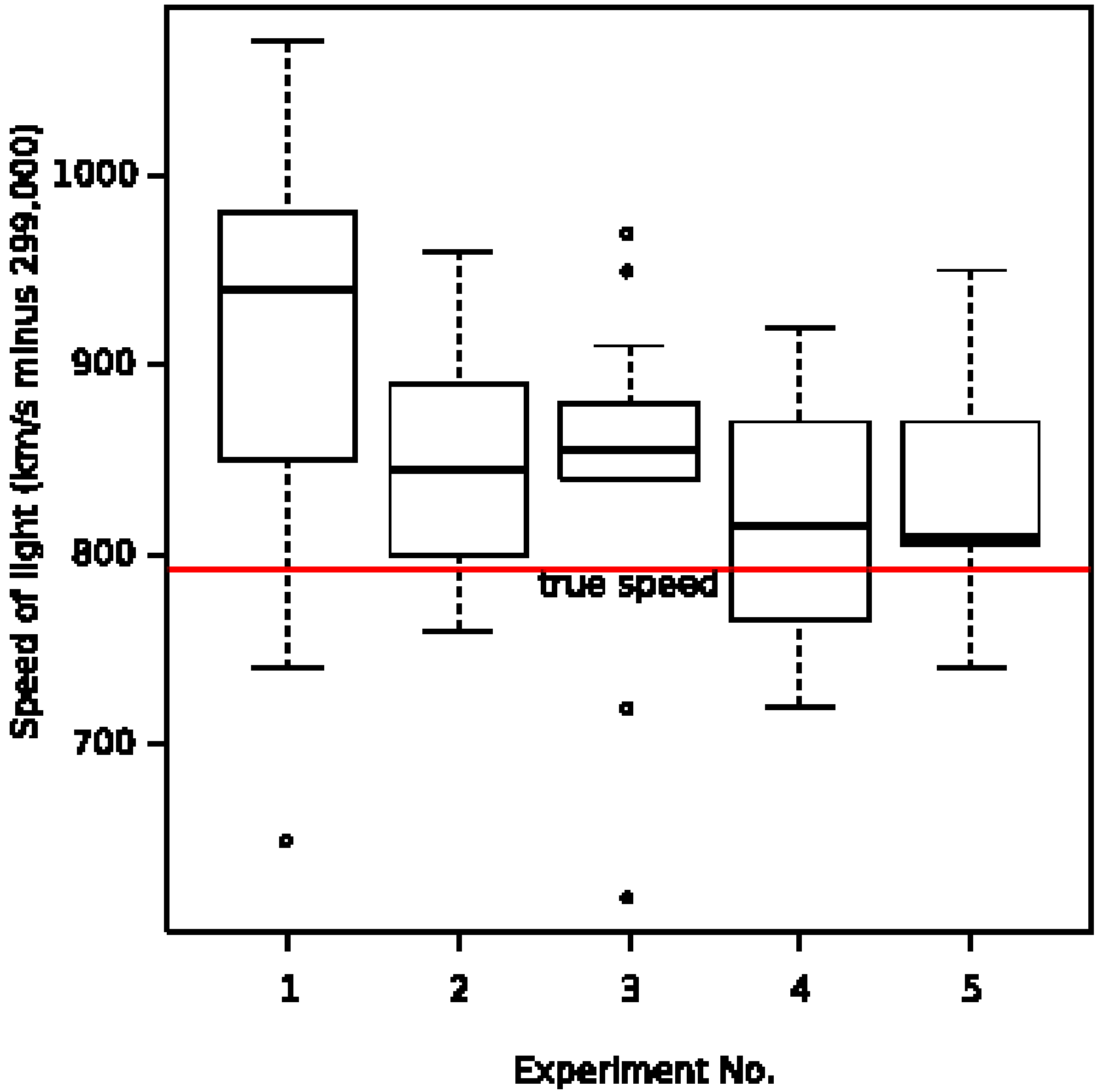
Compact representation of the mean and dispersion

- 1st and 3rd quartiles
- higher value $\leq Q_3 + 1.5(Q_3 - Q_1)$
- smaller value $\geq Q_1 - 1.5(Q_3 - Q_1)$
- outliers



[John W. Tukey (1915–2000)]

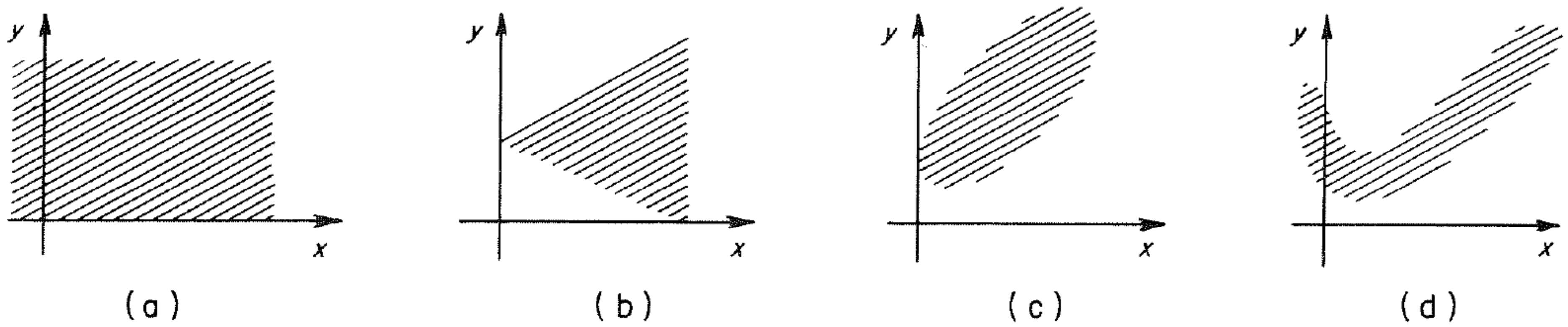
Box and whisker plots (cont'd)



MEASURING RELATIONS

About correlation

There exists various types of correlations between two variables X and Y.



(a) no correlation

(b) no correlation in mean (but correlation in dispersion)

(c) positive linear correlation

(d) non linear correlation

Correlation coefficients

Pearson's linear correlation coefficient

$$r_{XY} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \widehat{\mu}_x)(y_i - \widehat{\mu}_y)}{\widehat{\sigma}_x^2 \widehat{\sigma}_y^2}$$

$\longrightarrow$ measures the strength and direction of the relationship

$\longrightarrow$ only for linear dependencies

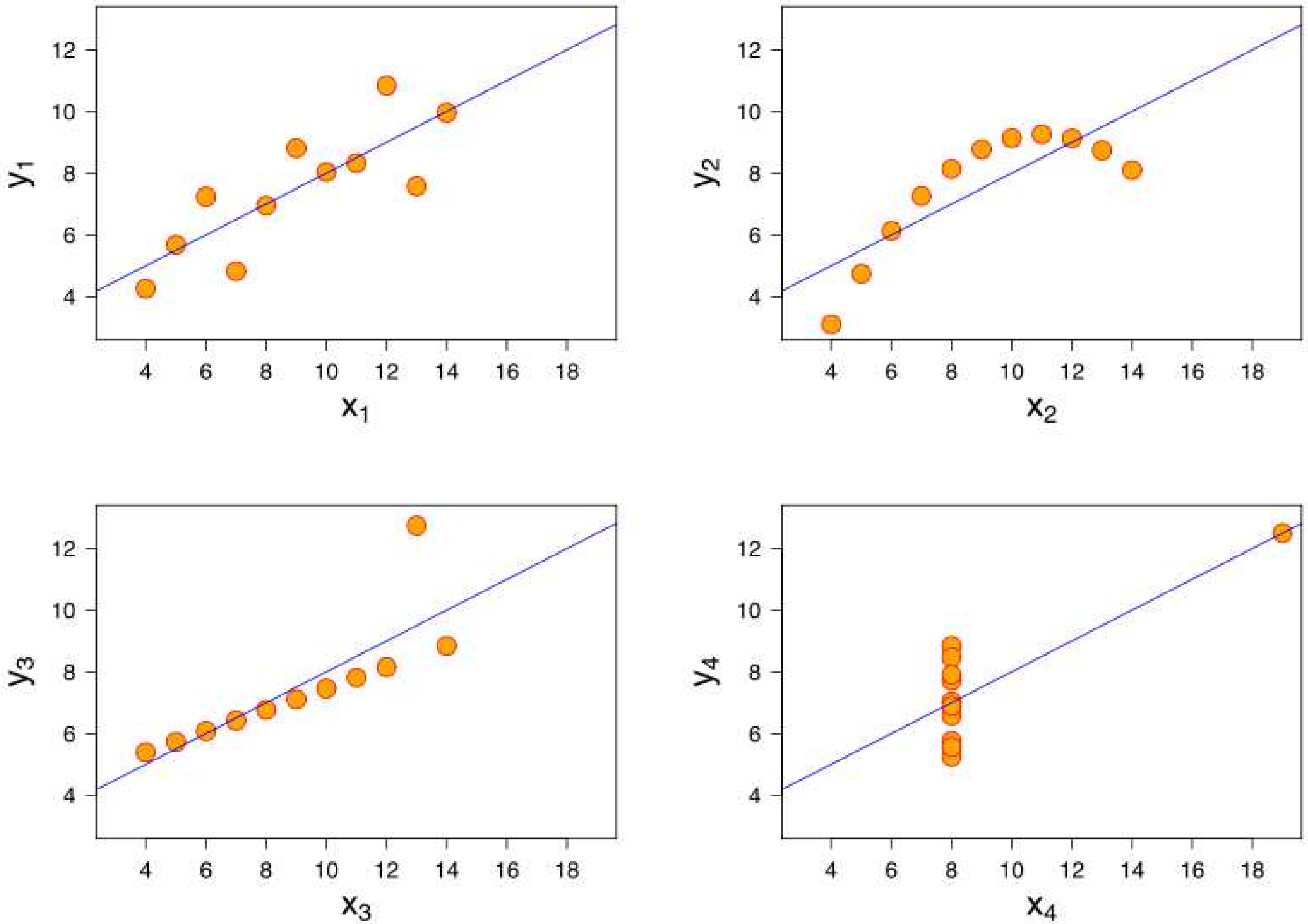
Spearman's rank correlation coefficient

$$\rho_{XY} = 1 - \frac{6\sum_{i=1}^{n}(r(x_i) - r(y_i))^2}{n(n^2-1)}$$

$\Rightarrow$ non linear monotonous dependencies

$\Rightarrow$ less sensitive to outliers

Pearson's linear correlation coefficient



Kendall's τ rank correlation

- Measure if two random variables X and Y vary in the same direction
- Idea: look at the sign of the product $(X_1 - X_2)(Y_1 - Y_2)$
- For all pairs (x_i, y_i) and (x_j, y_j)
  - count 1 if same order (i.e., $x_i < x_j$ and $y_i < y_j$)
  - count -1 otherwise

$$\tau_{XY} = \frac{2S}{n(n-1)}$$

x_i	1	2	3	4	5	6	7	8	9	10
y_i	3	1	4	2	6	5	9	8	10	7

10 wines ranked by two experts

$$\rho = 0.84 \qquad \tau = 0.64$$

Correlation ratio

For mixed situations where X is categorical and Y numerical

$$\eta^2_{Y|X} = \frac{\frac{1}{n}\sum_i n_i(\overline{\mu}_{Y|X=i} - \overline{\mu}_Y)^2}{\sum_y (y - \overline{\mu}_Y)^2} = \frac{\sigma^2_{\overline{\mu}_{Y|X=i}}}{\sigma^2_Y}$$

$\eta = 0 \Rightarrow$ no dispersion of the mean across categories

$\eta = 1 \Rightarrow$ no dispersion within the respective categories

Measure of association for categorical variables

Contingency table for two categorical variables X and Y

	Right-handed	Left-handed	Total
Male	43	9	52
Female	44	4	48
Total	87	13	100

Deviation from independency

- empirical independence if all line and column profiles are identical

$$\Rightarrow n_{ij} = \frac{n_{i\cdot}\, n_{\cdot j}}{n}$$

- χ^2 independence test statistics

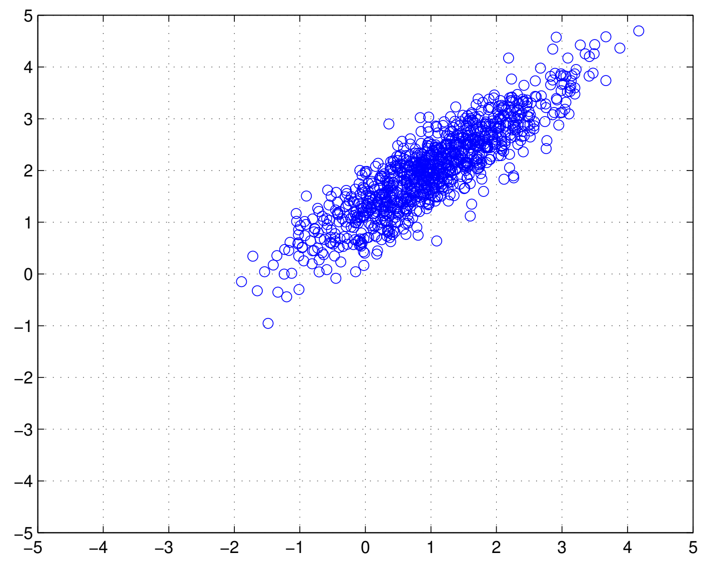
$$\chi^2 = \sum_i \sum_j \frac{\left(n_{ij} - \frac{n_{i\cdot}\, n_{\cdot j}}{n}\right)^2}{\frac{n_{i\cdot}\, n_{\cdot j}}{n}}$$

MULTIDIMENSIONAL DATA ANALYSIS

Projection *vs.* Clustering

We observe variables $X \in \mathbb{R}^p$. **What to do if p is large?**

- either display variables in $\mathbb{R}^q$, with $q \ll p$
  - PCA
  - LDA & the likes
  - correspondence analysis
  - factor analysis
- clustering
  - k-means & the likes
  - bottom-up clustering
  - spectral clustering



The general idea of factor analysis

Explain observed variables in terms of a smaller number of unobserved, or latent, variables.

$$x_i \quad = \quad \mu + l_{i1} f_1 + \ldots + l_{ik} f_k + \epsilon_i \qquad\qquad i = 1, \ldots, n$$

Principal Component Analysis = linear projection

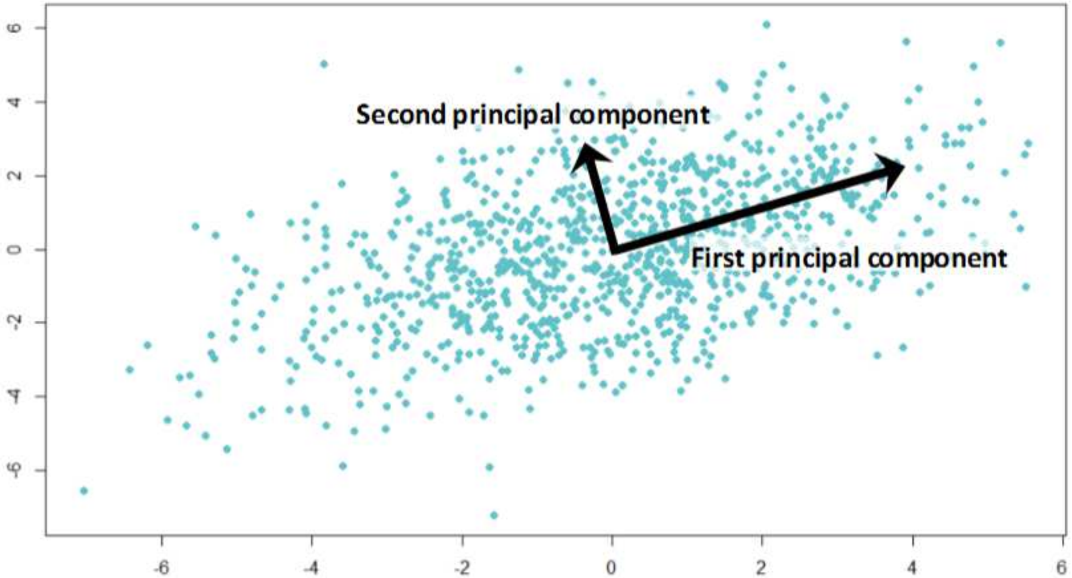
In PCA, we rescrict ourselves to linear transformations, i.e.

1. the new reference $\mathbf{u}$ is a linear combination of $\mathbf{v}$
2. $\mathbf{x}_u$ is a linear combination of $\mathbf{x}_v$, possibly with dimensionality reduction

$$\mathbf{y}_{q\times 1} = \mathbf{U}_{q\times p} \quad \mathbf{x}_{p\times 1}$$

$$\begin{pmatrix} y_1 \\ \vdots \\ y_q \end{pmatrix} = \begin{pmatrix} u_{11} & \cdots & u_{1p} \\ \vdots & \ddots & \vdots \\ u_{q1} & \cdots & u_{qp} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix}$$

What’s a good linear projection



- keep distances unchanged
- maximize variance
- maximize inertia
- least square error

The algorithmics of PCA (1)

Consider n *observations* with p *variables* each

$$\mathbf{X} = \begin{pmatrix} x_{11} & \dots & x_{1i} & \dots & x_{1p} \\ \vdots & \ddots & \vdots & & \vdots \\ x_{j1} & \dots & x_{ij} & \dots & x_{jp} \\ \vdots & & \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{ni} & \dots & x_{np} \end{pmatrix}$$

- for sake of simplification, we will assume that
  1. the data exhibit a null empirical mean (centered)
  2. all observations are equally important with a wieght $1/n$
- $\mathbf{V} \propto \mathbf{X}'\mathbf{X}$ and $\mathbf{R} \propto \tilde{\mathbf{X}}'\tilde{\mathbf{X}}$

The algorithmics of PCA (2)

- **Good projection = keep unaltered** (as much as possible) **the distances between individuals**
- Maximize the inertia of the projected data

$$\text{Trace}(\mathbf{Y}'\mathbf{Y}) = \text{Trace}(\mathbf{VP})$$

- PCA derives from the two following theorems:
  - Theorem 1: *If F_k is the subspace of dimension k with maximal inertia, the subspace of dimension $k+1$ with maximal inertia is the direct sum of F_k and of the 1-dimensional subspace orthogonal to F_k with maximal inertia.* $\Rightarrow$ **the solutions are intricated**
  - Theorem 2: *The subspace F_k is the subspace generated by the k eigen vectors of V associated with the k highest eigen values of V.*

The algorithmics of PCA (3)

1. Compute the covariance matrix $\mathbf{V}_{p\times p} = \frac{1}{n}\mathbf{X}'\mathbf{X}$ (or the correlation matrix $\mathbf{R}$)
2. Compute eigen system

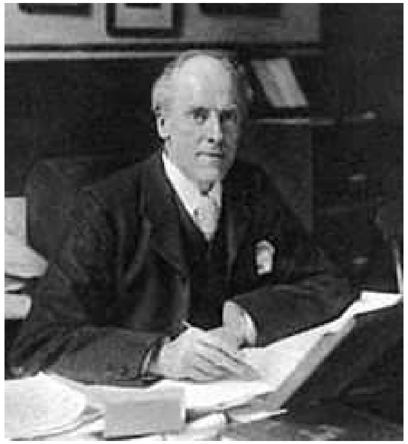
$$\mathbf{V}_{p\times p} = \mathbf{U}_{p\times p}\mathbf{\Lambda}_{p\times p}\mathbf{U}^{-1}_{p\times p} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}'$$

   - NIPALS algorithm for very high dimension data
3. Sort eigen values and retain the highest ones, sorting $\mathbf{U}_{p\times p}$ accordingly to yield $\overline{\mathbf{U}}_{q\times p}$
4. Reconstruct or project $\mathbf{X}$

$$\mathbf{Y}_{q\times n} = \overline{\mathbf{U}}_{q\times p}\mathbf{X}_{p\times n}$$

The guys behind PCA

Karl Pearson. On lines and planes of closest fit to systems of points in space. Philosophical Magazine, Series 6, 2(11):559–572, 1901.



Karl Pearson

1857–1936

Harold Hotelling. Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology, 24(6–7):417–441,498–520, 1933.

Harold Hotteling

1895–1973

Observation and variable spaces

	Var 1	Var 2
A	4	2
B	7	6
C	10	4

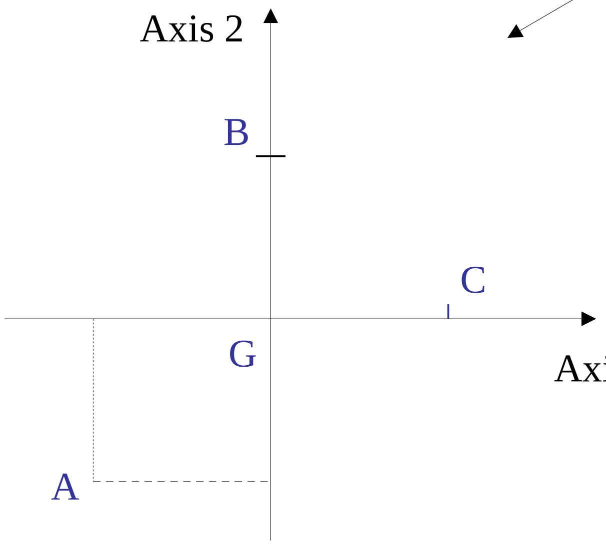
$\overline{x_1} = 7$ $\quad \overline{x_2} = 4$

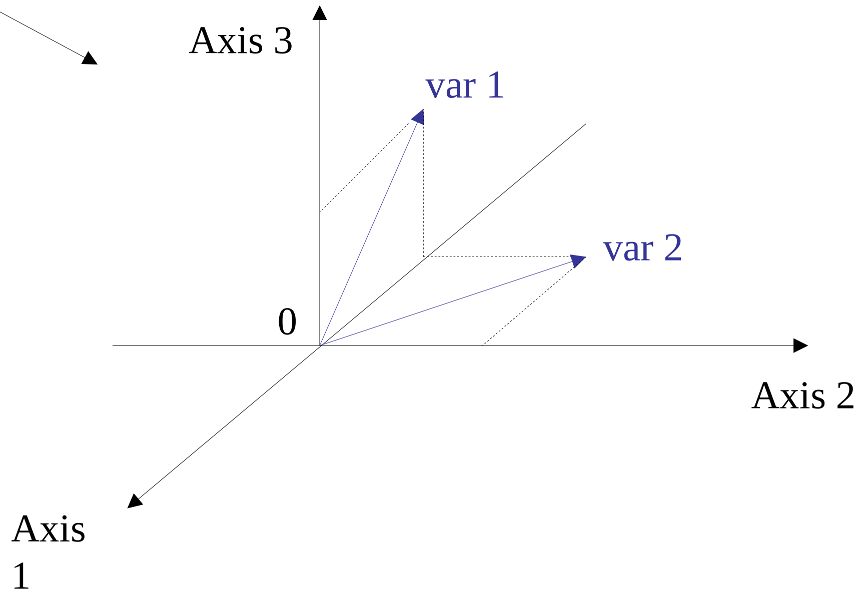
$s_j^2 = 6$ $\quad s_j^2 = 8/3$

$$\begin{pmatrix} -\sqrt{\frac{3}{2}} & -\sqrt{\frac{3}{2}} \\ 0 & \sqrt{\frac{3}{2}} \\ \sqrt{\frac{3}{2}} & 0 \end{pmatrix} = \begin{pmatrix} -1{,}22 & -1{,}22 \\ 0 & 1{,}22 \\ 1{,}22 & 0 \end{pmatrix}$$

Observation and variable spaces (cont'd)

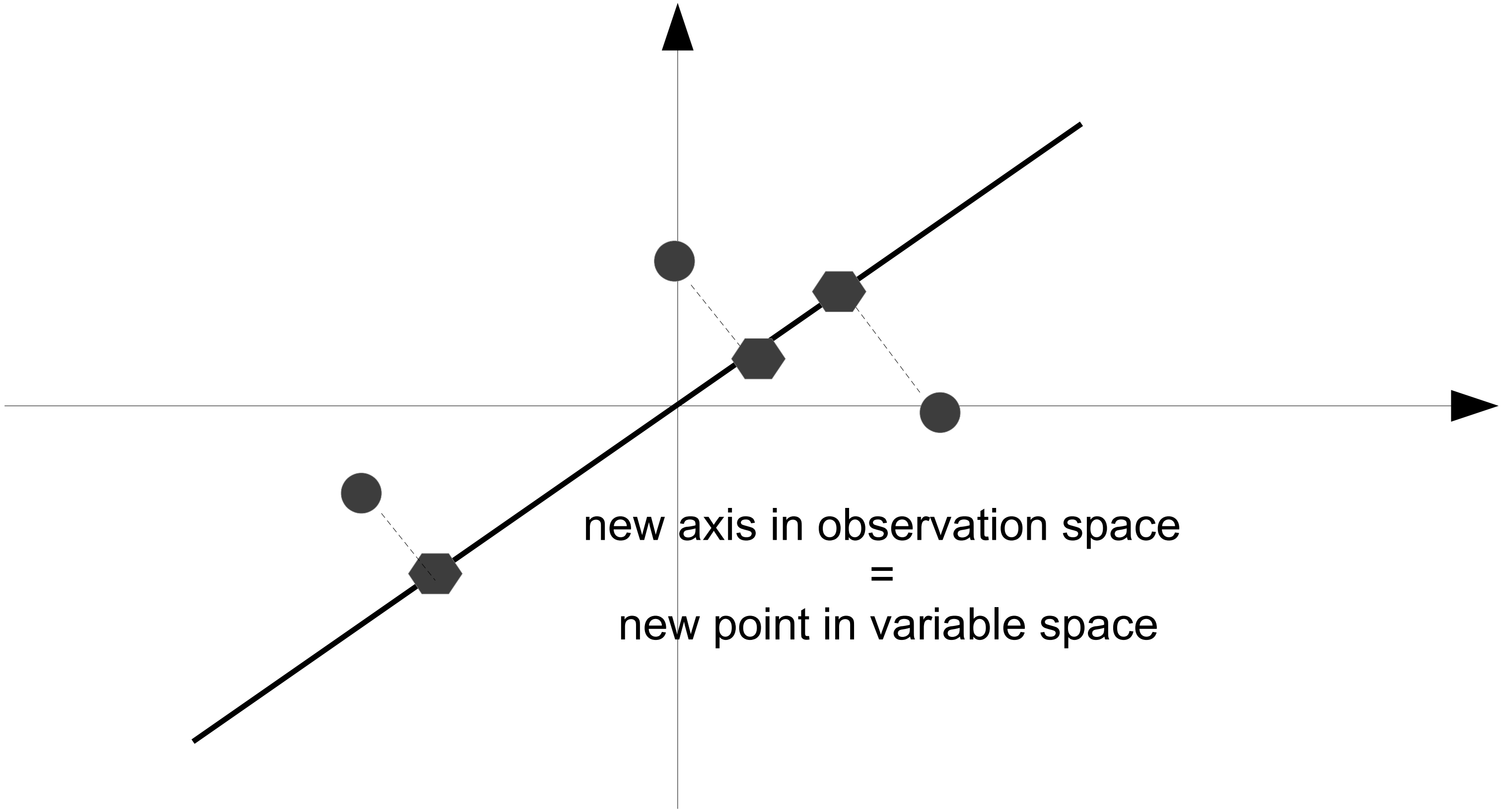
$$\begin{pmatrix} -\sqrt{\frac{3}{2}} & -\sqrt{\frac{3}{2}} \\ 0 & \sqrt{\frac{3}{2}} \\ \sqrt{\frac{3}{2}} & 0 \end{pmatrix} = \begin{pmatrix} -1{,}22 & -1{,}22 \\ 0 & 1{,}22 \\ 1{,}22 & 0 \end{pmatrix}$$





Projection = creating new variables

$$v_1' \begin{pmatrix} y_1(1) & y_2(1) & y_3(1) \end{pmatrix} = \begin{pmatrix} u_1 & u_2 \end{pmatrix} \begin{pmatrix} x_1(1) & x_2(1) & x_3(1) \\ x_1(2) & x_2(2) & x_3(2) \end{pmatrix}$$

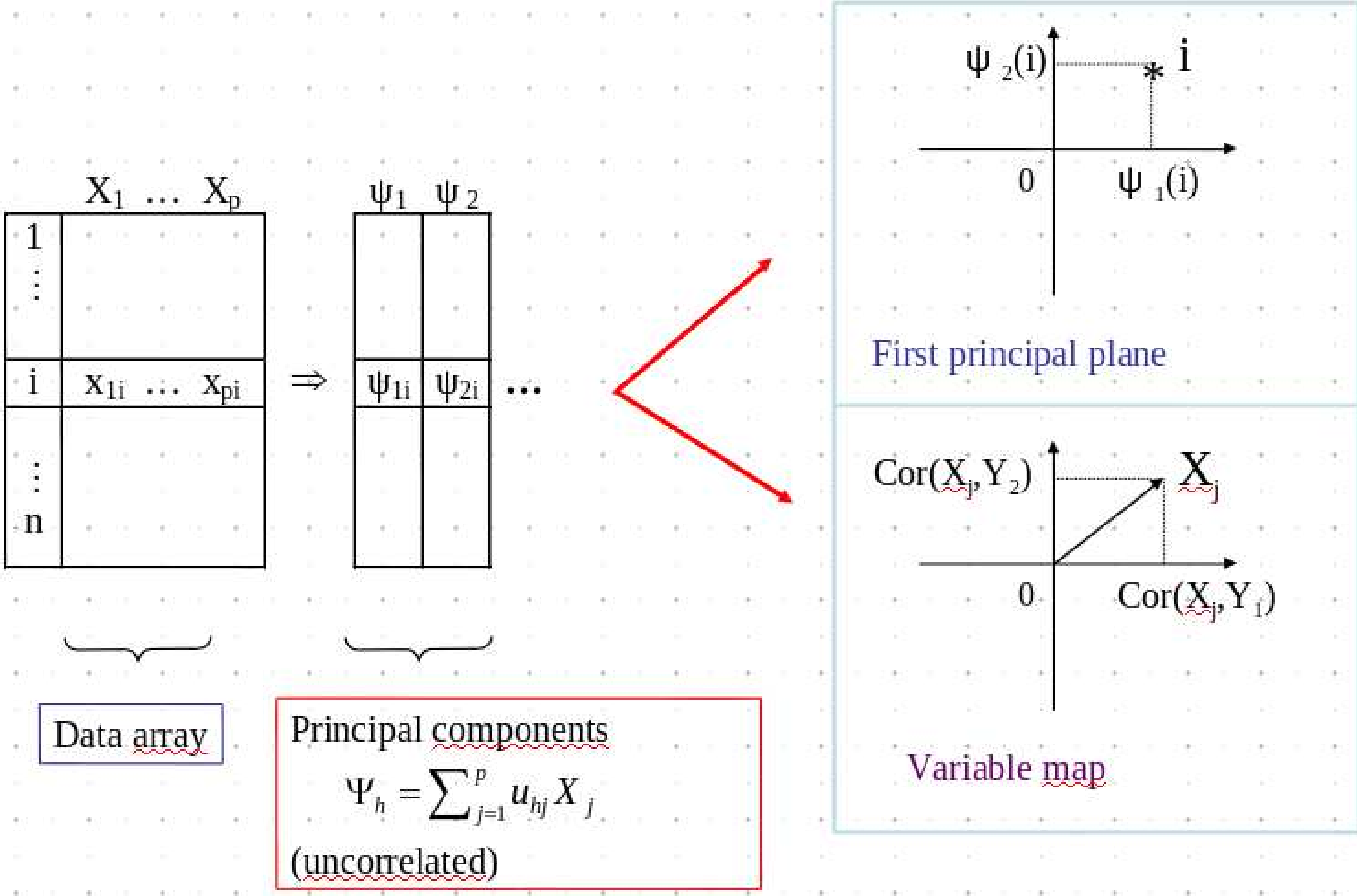


The vocabulary of PCA

- $\mathbf{u}_i$ = ith **principal axis** or factor = linear combination of descriptive variables
  - Note that there actually is a difference between axis and factor if a metric $\mathbf{M} \neq \mathbf{I}$ is used.
- $\mathbf{c}_i = \mathbf{X}\mathbf{u}_i$ is the ith **principal components** (homogeneous to a variable)
  - $V[\mathbf{c}_i] = \lambda_i$
  - principal components are the eigen vectors of the (n, n) matrix $\mathbf{X}\mathbf{X}'$ (⇒ relation to the variable space)

In summary: PCA replaces the correlated variables $\mathbf{x}_1 \ldots \mathbf{x}_p$ with new variables, the principal components $\mathbf{c}_1 \ldots \mathbf{c}_q$, uncorrelated linear combination of the variables $\mathbf{x}_i$ with maximum variance.

Result interpretation and quality



Result interpretation and quality

- Interpretation
  - correlation between components and variables

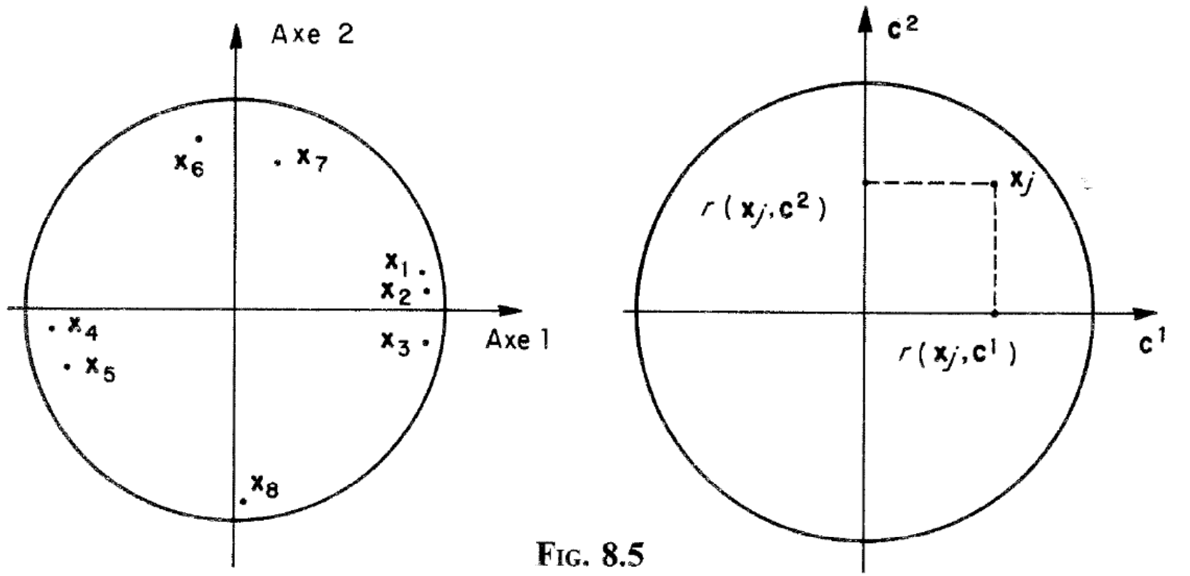


FIG. 8.5

  - contribution of each sample to an axis
- Measure of quality
  - Global measurement: Fraction of the total inertia retained = $\frac{\lambda_1+\ldots+\lambda_q}{I_g}$
  - Local measurement: angle between the principal plan and a sample
  $\rightarrow$ small angle $\Rightarrow$ good representation

Example

- Data table

		Pain ordinaire PAO	Autre pain PAA	Vin ordinaire VIO	Autre vin VIA	Pommes de terre POT	Légumes secs LEC	Raisin de table RAI	Plats préparés PLP
Exploitants agricoles	AGRI	167	1	163	23	41	8	6	6
Salariés agricoles	SAAG	162	2	141	12	40	12	4	15
Professions indépendantes	PRIN	119	6	69	56	39	5	13	41
Cadres supérieurs	CSUP	87	11	63	111	27	3	18	39
Cadres moyens	CMOY	103	5	68	77	32	4	11	30
Employés	EMPL	111	4	72	66	34	6	10	28
Ouvriers	OUVR	130	3	76	52	43	7	7	16
Inactifs	INAC	138	7	117	74	53	8	12	20

(*Source* : A. Villeneuve, « La consommation alimentaire des Français », *Collections de l'INSEE*, M 34.)

- Correlation matrix

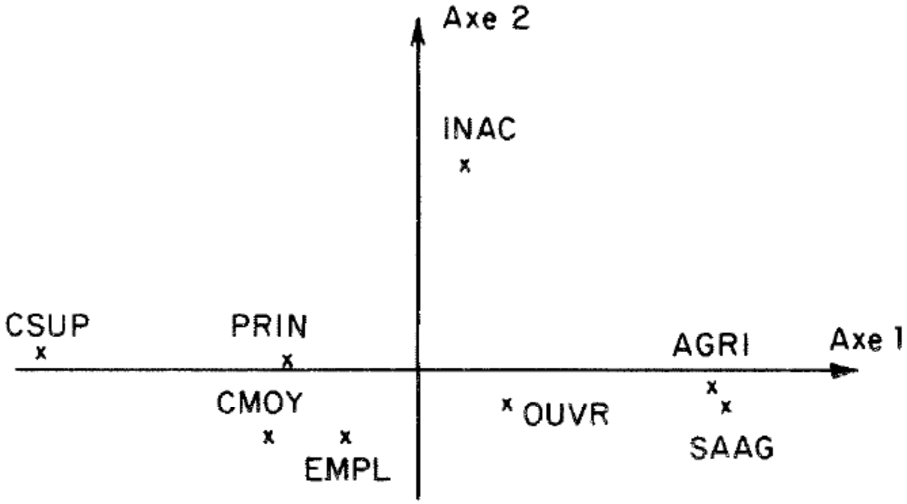
	PAO	PAA	VIO	VIA	POT	LEC	RAI	PLP
PAO	100							
PAA	– 75	100						
VIO	83	– 57	100					
VIA	– 89	90	– 73	100				
POT	66	– 30	52	– 40	100			
LEC	90	– 66	80	– 84	61	100		
RAI	– 82	96	– 65	91	– 42	– 82	100	
PLP	– 85	78	– 82	72	– 55	– 73	85	100

[Source: Saporta 2002, pp. 180–183]

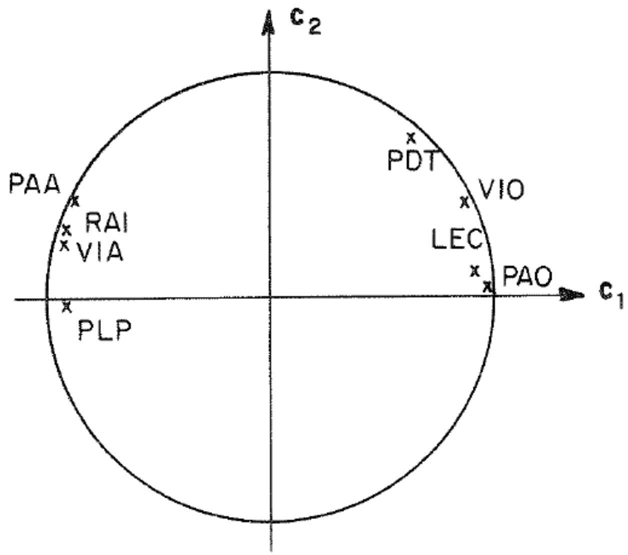
Example (cont'd)

λ	6.21	0.89	0.42	0.32	0.14	0.01	0.005
Inertia (in %)	77.57	11.21	5.26	3.99	1.74	0.11	0.06
Cumulated	77.57	88.78	94.04	98.0	99.8	99.9	100

Projection of the observations



Principal components



Eigenfaces

- Face images are represented as a vector of pixels
- PCA is used to find the principal components/faces

  $\Rightarrow$ consider the parameter space (size M) rather than the image space (of size N^2)
- Each face is represented by a linear combination of the eigenfaces

[M. Turk and A. Pentland. Face recognition using eigenfaces, in CVPR 91]

Latent semantic analysis/indexing

- each observation/document is a bag of words in $\mathbb{R}^d$
- d is the number of index terms
- x_{ji} is proportional to the frequency of term j in document i

$$
\mathbf{X}_{d\times n} \approx \mathbf{U}_{d\times r} \mathbf{Z}_{r\times n}
$$

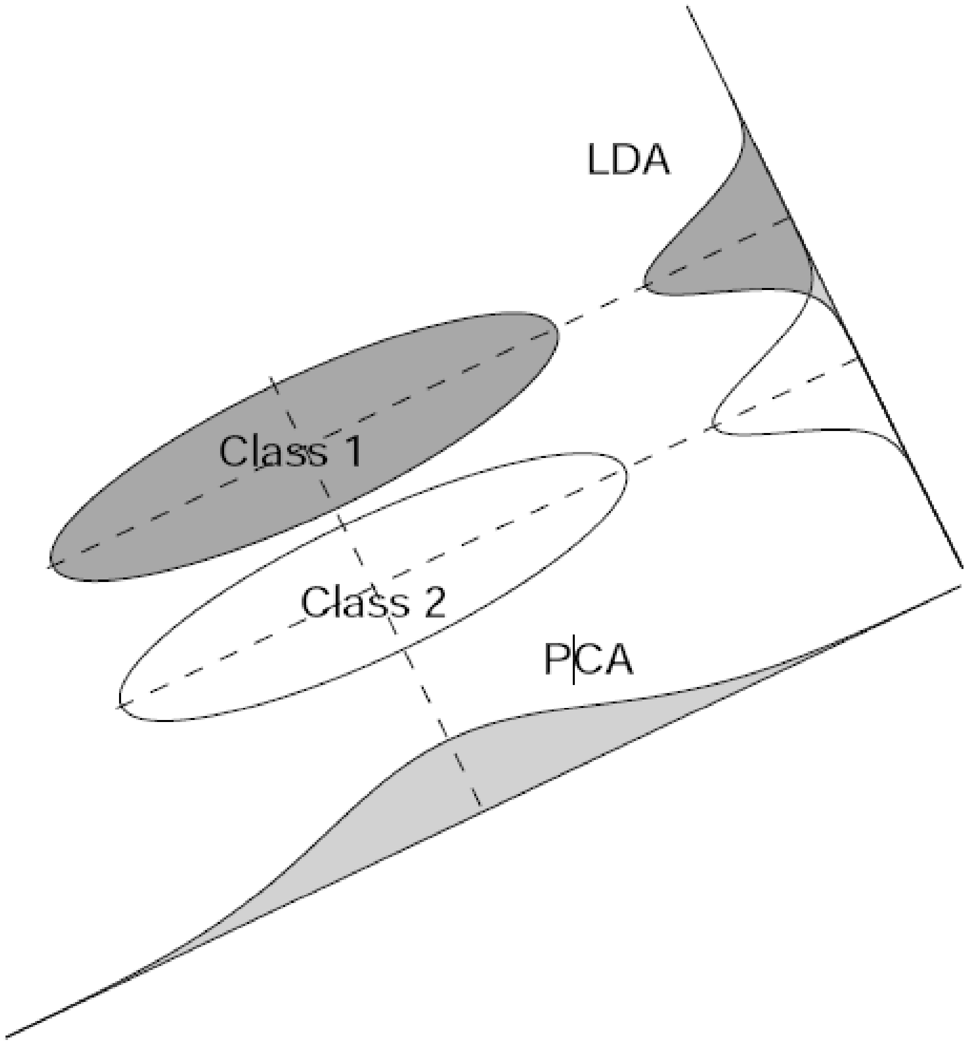
$$
\begin{pmatrix}
\text{stocks:} & 2 & \cdots & 0 \\
\text{chairman:} & 4 & \cdots & 1 \\
\text{the:} & 8 & \cdots & 7 \\
\cdots & \vdots & \cdots & \vdots \\
\text{wins:} & 0 & \cdots & 2 \\
\text{game:} & 1 & \cdots & 3
\end{pmatrix}
\approx
\begin{pmatrix}
0.4 & \cdots & -0.001 \\
0.8 & \cdots & 0.03 \\
0.01 & \cdots & 0.04 \\
\vdots & \cdots & \vdots \\
0.002 & \cdots & 2.3 \\
0.003 & \cdots & 1.9
\end{pmatrix}
\begin{pmatrix}
| & & | \\
\mathbf{z}_1 & \cdots & \mathbf{z}_n \\
| & & |
\end{pmatrix}
$$

$\Rightarrow$ documents are better represented in the *concept* subspace obtained by PCA on the term / document matrix.

[S. Deerwater *et al.* Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6):391–407, 1990.]

PCA with data from several classes

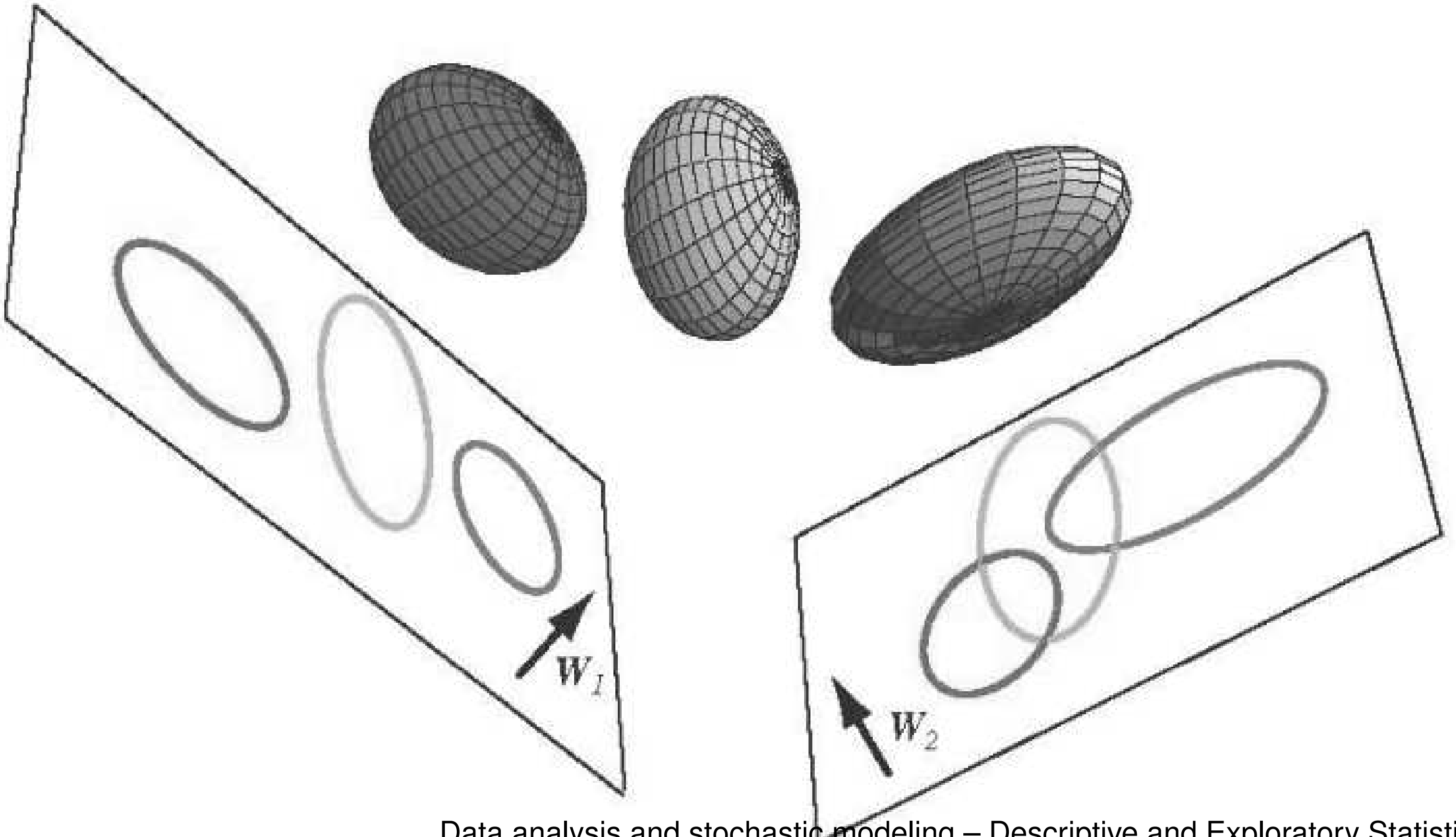
PCA disregard the information on the class of each sample!



Linear discriminant analysis

Find a linear projection of the data $\mathbf{X}$ into a subspace of smaller dimension which

1. maximizes the dispertion across classes
2. minimizes the dispertion within classes



Fisher's linear discriminant

- Observations $\mathbf{x}_i$ from two classes with means μ_0 and μ_1 and covariance matrices $\mathbf{\Sigma}_0$ and $\mathbf{\Sigma}_1$
- Projection along the line $\mathbf{w}$ will result in a separation defined as

$$S = \frac{\sigma_{\text{across}}}{\sigma_{\text{within}}} = \frac{(\mathbf{w}(\mu_1 - \mu_0))^2}{\mathbf{w}'(\Sigma_0 + \Sigma_1)\mathbf{w}} \quad .$$

- Maximum separation occurs when

$$\mathbf{w} = (\Sigma_0 + \Sigma_1)^{-1}(\mu_1 - \mu_0)$$

Two-class LDA is equivalent to Fisher's linear discriminant with the assumptions that the posterior distribution $p(\mathbf{x}_i|\text{classe})$ is Gaussian and that they are homoscedastic ($\Sigma_0 = \Sigma_1 = \Sigma$)

http://www.youtube.com/watch?v=fkGpzbXnO0c

Multiclass linear discriminant analysis

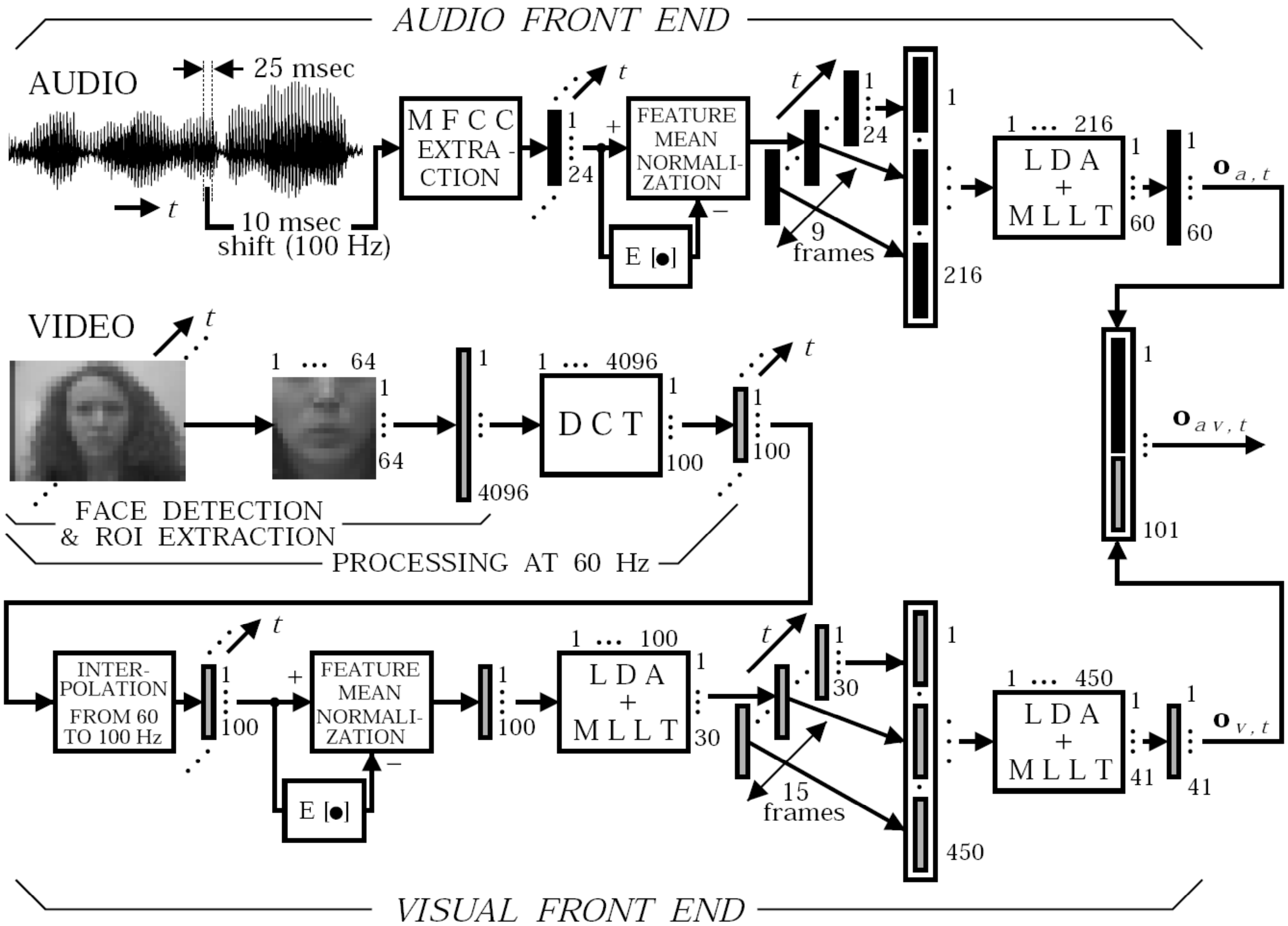
- Assume K classes of n_i samples each with respective mean μ_i
- Whithin-class scatter matrix: $\mathbf{S}_w = \sum_{i=1}^{K}\sum_{j=1}^{n_i}(x_{ij}-\mu_i)(x_{ij}-\mu_i)'$
- Across-class scatter matrix $\mathbf{S}_b = \sum_{i=1}^{K}(\mu_i-\mu)(\mu_i-\mu)'$
- search for the projection $\mathbf{y} = \mathbf{U}\mathbf{x}$ which maximizes

$$\max_{\mathbf{U}} \frac{|\mathbf{U}'\mathbf{S}_b\mathbf{U}|}{|\mathbf{U}'\mathbf{S}_w\mathbf{U}|}$$

- solution is given by the generalized eigen system

$$\mathbf{S}_b\mathbf{u}_k = \lambda_k\mathbf{S}_w\mathbf{u}_k$$

LDA front-end for audiovisual ASR



[Potamianos *et al.*. Recent advances in the automatic recognition of audio-visual speech. IEEE Proc., 2003.]

Beyond linear projections

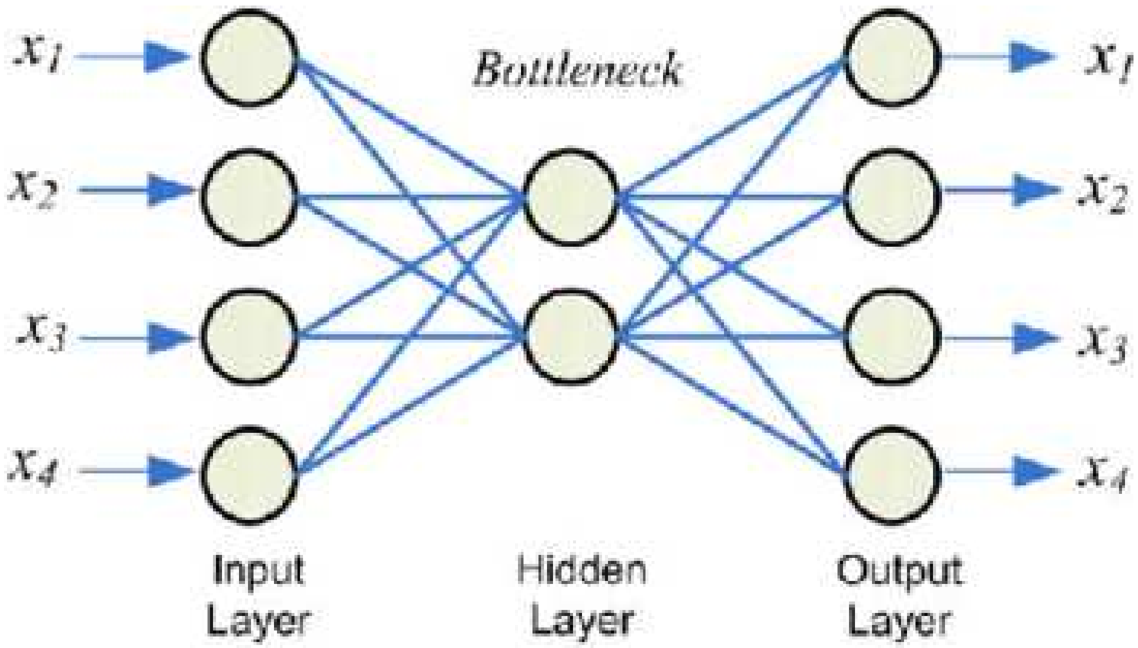
- Use a linear projection $\mathbf{y} = \mathbf{U}\mathbf{x}$ via the eigen system to
  - PCA: maximize the variance of the projected data
  - LDA: maximize discrimination between classes

- More complex forms of U can be used
  - NMF: non-negative matrix factorization
  - ICA: independent component analysis

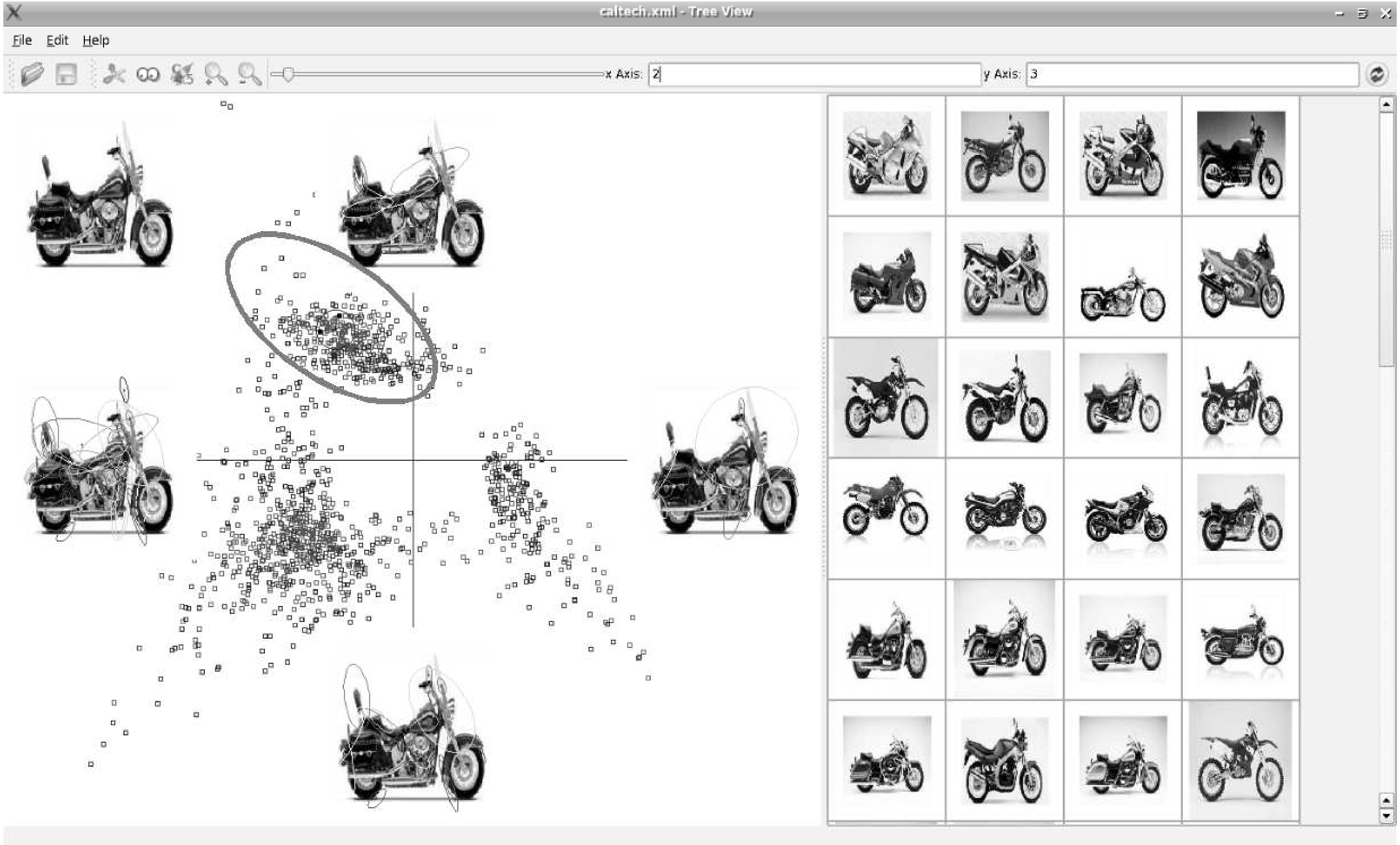
- Non linear transformations are also possible
  - use of kernels ($\rightarrow$ the kernel trick)
  - artifical neural network (Multi Layer Perceptron)

- Self-organizing maps

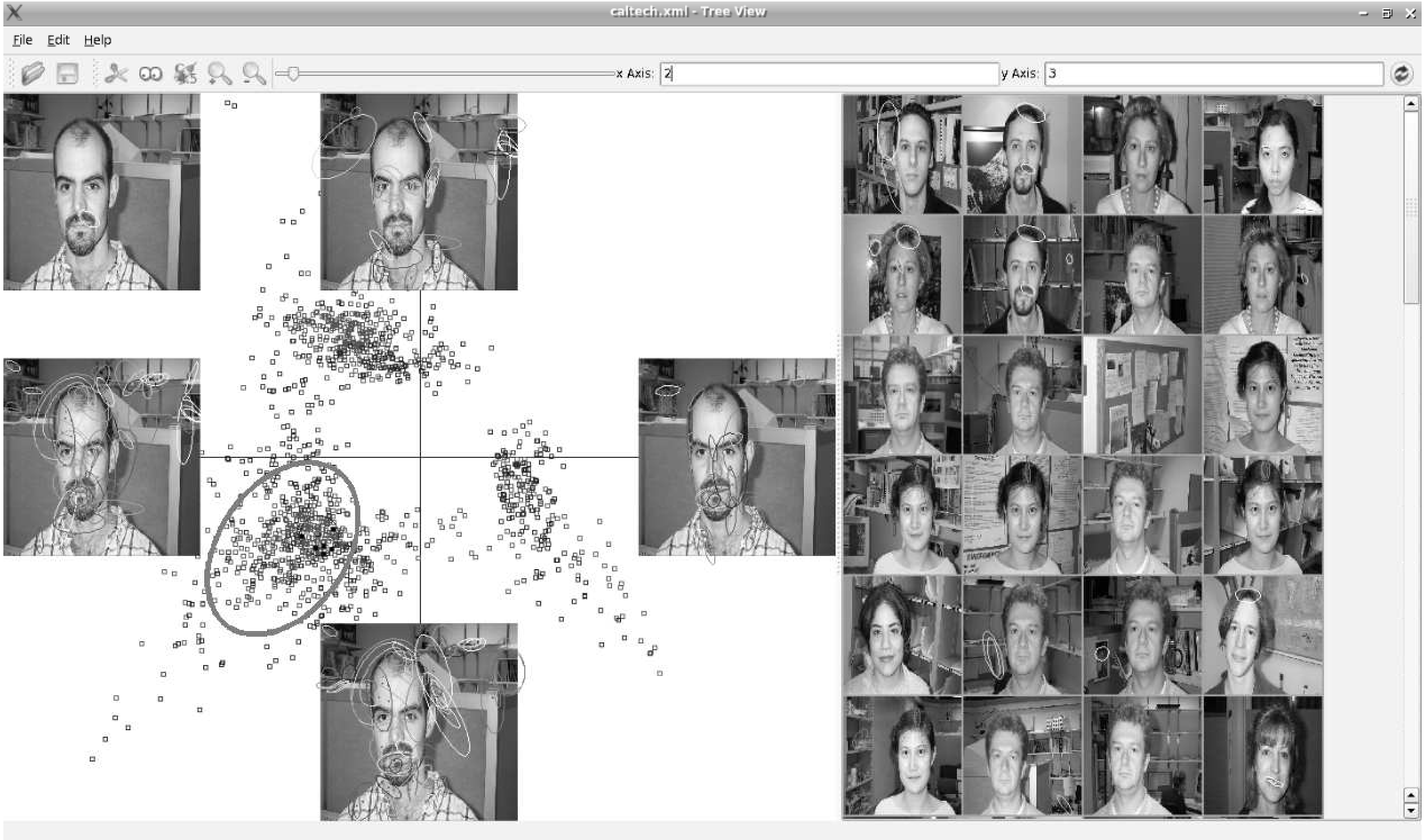
Beyond linear projections



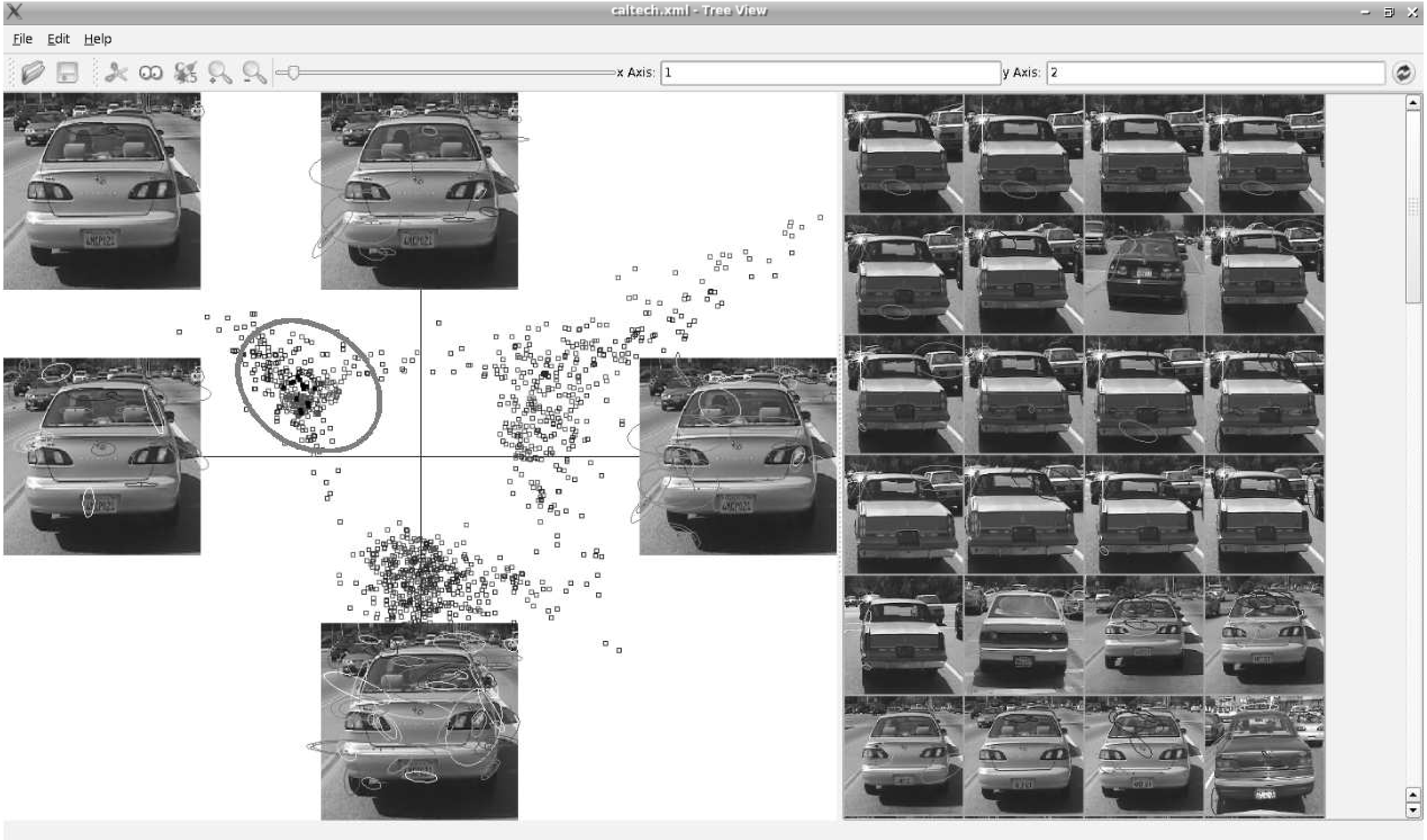
Factor analysis for images



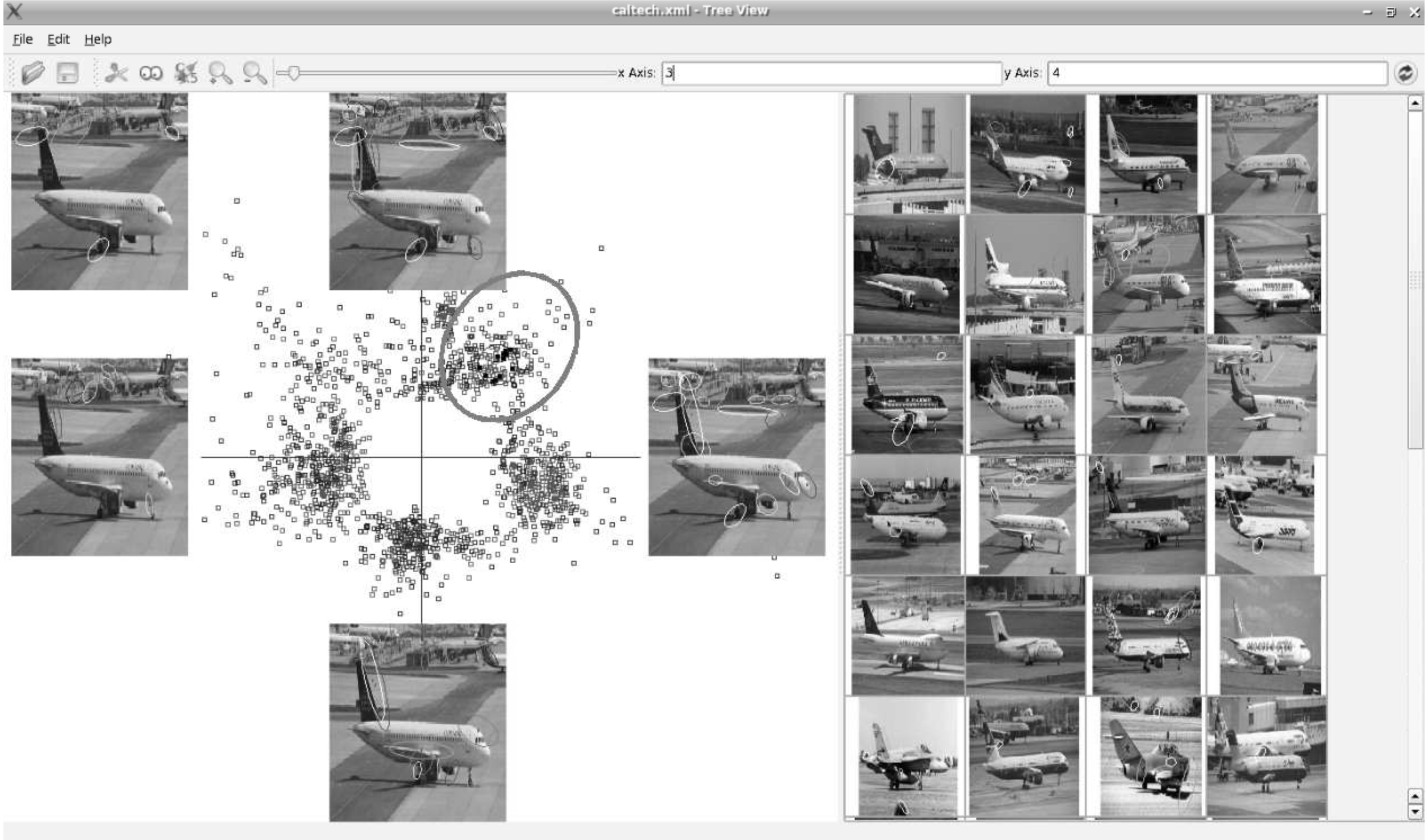
Factor analysis for images (cont'd)



Factor analysis for images (cont'd)



Factor analysis for images (cont'd)



Additional readings

- R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification, 2nd edition, Wiley-Interscience. (See in particular Chapter 3)
- C. M. Bishop. Pattern recognition and machine learning, Springer, 2006. (See in particular Chapter 12)
- Pattern recognition course of George Bebis (http://www.cse.unr.edu/ bebis/CS679/)