Course 27411

Biological Data analysis and chemometrics

February 11th, 2008

”Missing in Action?” – please check-in

Anders C. Raffalt	s001831	Cecilie Kyrø	s043146
Julie K. Høgh	**s020322**	Louise M. Jørgensen	s050438
Lasse R. Bech	**s021876**	**Fhayaz A. Sayyad**	**s060064**
Mikkel T. Pleman	s022252	Emmanouil Papadakis	s060724
Christina H. Kærgaard	s030513	Anders S. R. Ødum	s061888
Tina F. Rasmussen	s030526	Maria M. Bermejo	s071138
Ken Sejling	**s030632**	Carlos N. Bartolomé	s071139
Anders S. Laier	**s031755**	Asli Ozen	s071208
Rene J. Larsen	s031891	Silvia Bergamaschi	s073238
Jacob B. Reves	s032280	Michal Strejcek	s073676
Lars Poulsen	s032289	Gry H. Svendsen	s961360
Kasper Jensen	**s042114**	Casper T. Holst	s991097

Groups and Presentations

Group #1 (PCA):	
Lars Poulsen	s032289
Jacob B. Reves	s032280
Group #2 (MLR):	
Anders C. Raffalt	s001831
Casper T. Holst	s991097
Group #3 (PCR):	
Maria M. Bermejo	s071138
Carlos N. Bartolomé	s071139
Group #4 (PLS-R):	
Christina H. Kærgaard	s030513
Tina F. Rasmussen	s030526
Group #5 (Ridge regress.):	
Louise M. Jørgensen	s050438
Mikkel T. Pleman	s022252
Group #6 (Corresp. Analysis):	
Cecilie Kyrø	s043146
Gry H. Svendsen	s961360

Group #7 (Cluster Analysis):	
Rene J. Larsen	s031891
Emmanouil Papadakis	s060724
Group #8 (Classification I):	
Anders S. R. Ødum	s061888
Asli Ozen	s071208
Group #9 (Classification II):	
Silvia Bergamaschi	s073238
Michal Strejcek	s073676
Group #10 (PCO):	
Fhayaz A. Sayyad	s060064
Kasper Jensen	s042114
Anders S. Laier	s031755
Ken Sejling	s030632
Julie K. Høgh	s020322
Lasse R. Bech	s021876

Lecture #2
Principal Component Analysis

Michael Adsetts Edberg Hansen
Assisting Professor, CMB/DTU
meh@bio.dtu.dk

February 11th, 2008

”Principal what?”

- Principal Component Analysis – PCA.
- “Invented” by Karl Pearson in 1901:

  Pearson, K. (1901) On lines and planes of closest fit to systems of points in space. Philosophical Magazine *(6) 2: 559-572.*

- A.k.a. or closely related to:
  - Singular Value Decomposition (SVD)
  - Karhunen-Loéve Expansion
  - Eigenvector Analysis
  - Latent Vector Analysis
  - Characteristic Vector Analysis
  - Hotelling Transformation

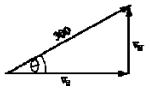
!!!! FIRST !!!!

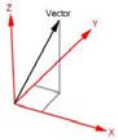
- Lets recap from linear algebra

Scalars, Vectors and Matrices

- **Scalar**: variable described by a single number (magnitude)
  - Temperature = 20 °C
  - Density = 1 g.cm^{-3}
  - Image intensity (pixel value) = 2546 a. u.
- **Vector**: variable described by magnitude and direction

$$\mathbf{v} = \begin{bmatrix} v_n \\ v_e \end{bmatrix}$$





Column vector

$$\mathbf{b} = \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}$$

Row vector

$$\mathbf{d} = \begin{bmatrix} 3 & 4 & 9 \end{bmatrix}$$

- **Matrix**: rectangular array of scalars

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 5 & 4 & 1 \\ 6 & 7 & 4 \end{bmatrix} \qquad \mathbf{C} = \begin{bmatrix} 1 & 4 \\ 2 & 7 \\ 3 & 8 \end{bmatrix} \quad \begin{matrix} 2 \\ \\ 3 \end{matrix} \qquad \mathbf{D} = \begin{bmatrix} d_{11} & d_{12} & d_{13} \\ d_{21} & d_{22} & d_{23} \\ d_{31} & d_{32} & d_{33} \end{bmatrix}$$

Square (3 x 3) Rectangular (3 x 2) d_{ij} : i^{th} row, j^{th} column

Vector Operations

- Transpose operator

$$\mathbf{b} = \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix} \quad \mathbf{b}^T = \begin{bmatrix} 1 & 1 & 2 \end{bmatrix} \qquad \mathbf{d} = \begin{bmatrix} 3 & 4 & 9 \end{bmatrix} \quad \mathbf{d}^T = \begin{bmatrix} 3 \\ 4 \\ 9 \end{bmatrix}$$

column → row row → column

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 5 & 4 & 1 \\ 6 & 7 & 4 \end{bmatrix} \quad \mathbf{A}^T = \begin{bmatrix} 1 & 5 & 6 \\ 2 & 4 & 7 \\ 3 & 1 & 4 \end{bmatrix}$$

- Outer product = matrix

$$\mathbf{x}\mathbf{y}^T = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \begin{bmatrix} y_1 & y_2 & y_3 \end{bmatrix} = \begin{bmatrix} x_1y_1 & x_1y_2 & x_1y_3 \\ x_2y_1 & x_2y_2 & x_2y_3 \\ x_3y_1 & x_3y_2 & x_3y_3 \end{bmatrix}$$

Vector Operations

- Inner product = scalar

$$\mathbf{x}^T\mathbf{y} = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = x_1y_1 + x_2y_2 + x_3y_3 = \sum_{i=1}^{3} x_iy_i$$

$$\mathbf{x}^T\mathbf{y} = \begin{bmatrix} x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \sum_{i=1}^{n} x_iy_i$$

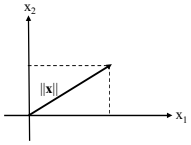
- Length of a vector

$\| \mathbf{x} \| = (x_1^2 + x_2^2)^{1/2}$

$\| \mathbf{x} \| = (x_1^2 + x_2^2 + x_3^2)^{1/2}$

Inner product of a vector with itself = (vector length)^2

$\mathbf{x}^T\mathbf{x} = x_1^2 + x_2^2 + x_3^2 = (\| \mathbf{x} \|)^2$



Matrix Operations

- Addition (matrix of same size)
  - Commutative: **A+B=B+A**
  - Associative: **(A+B)+C=A+(B+C)**

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix} + \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 3 & 3 \\ 3 & 3 \end{bmatrix}$$

Matrix Operations

- Multiplication
  - number of columns in first matrix = number of rows in second

C = **A** **B**

(m x p) = (m x n) (n x p)

c_{ij} = inner product between i^{th} row in **A** and j^{th} column in **B**

$$\mathbf{A}\,\mathbf{B} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \begin{bmatrix} 7 & 1 \\ 8 & 2 \\ 9 & 3 \end{bmatrix} = \begin{bmatrix} 1\times7+2\times8+3\times9 & 1\times1+2\times2+3\times3 \\ 4\times7+5\times8+6\times9 & 4\times1+5\times2+6\times3 \end{bmatrix} = \begin{bmatrix} 50 & 14 \\ 122 & 32 \end{bmatrix}$$

2 x 3 3 x 2 2 x 2

  - Associative: **(A B) C = A (B C)**
  - Distributive: **A (B+C) = A B + A C**
  - Not commutative: **AB ≠ BA!!!**
  - **$(\mathbf{A}\,\mathbf{B})^T = \mathbf{B}^T\mathbf{A}^T$**

Some Definitions …

- Identity Matrix

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \mathbf{I}\,\mathbf{A} = \mathbf{A}\,\mathbf{I} = \mathbf{A}$$

- Diagonal Matrix

$$\mathbf{D} = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 7 \end{bmatrix}$$

- Symmetric Matrix

$$\mathbf{B} = \begin{bmatrix} 3 & 1 & 0 \\ 1 & 5 & 2 \\ 0 & 2 & 7 \end{bmatrix} \qquad \mathbf{B} = \mathbf{B}^T \qquad b_{ij} = b_{ji}$$

Matrix Inverse

- $\mathbf{A}^{-1}\,\mathbf{A} = \mathbf{A}\,\mathbf{A}^{-1} = \mathbf{I}$

$$\mathbf{D}\,\mathbf{D}^{-1} = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 7 \end{bmatrix} \begin{bmatrix} \frac{1}{3} & 0 & 0 \\ 0 & \frac{1}{5} & 0 \\ 0 & 0 & \frac{1}{7} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \mathbf{I}$$

Properties

$\mathbf{A}^{-1}$ only exists if $\mathbf{A}$ is **square** (n x n)

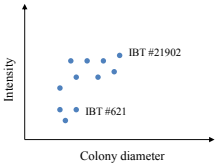
If $\mathbf{A}^{-1}$ exists then $\mathbf{A}$ is **non-singular** (invertible)

$(\mathbf{A}\,\mathbf{B})^{-1} = \mathbf{B}^{-1}\,\mathbf{A}^{-1}$; $\quad \mathbf{B}^{-1}\,\mathbf{A}^{-1}\,\mathbf{A}\,\mathbf{B} = \mathbf{B}^{-1}\,\mathbf{B} = \mathbf{I}$

$(\mathbf{A}^{T})^{-1} = (\mathbf{A}^{-1})^{T}$; $\quad (\mathbf{A}^{-1})^{T}\,\mathbf{A}^{T} = (\mathbf{A}\,\mathbf{A}^{-1})^{T} = \mathbf{I}$

Data as observations

- Samples as observations in a Multi/hyper-dimensional Space:
  - Objects are a collection of features.
  - Features are dimensions.
  - Objects are points in a multidimensional space.

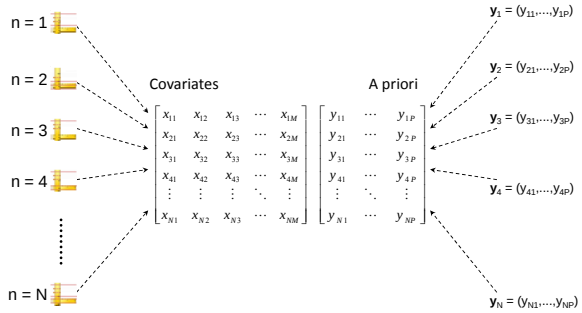


- Mathematical notation
  - N is the number of observations
  - M is the number of variables/features

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_N \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1M} \\ x_{21} & x_{22} & \cdots & x_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NM} \end{bmatrix}$$

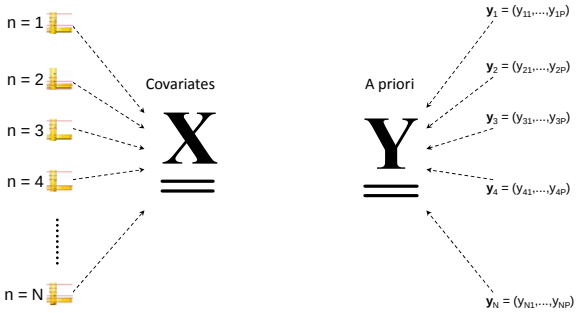
Data as observations

- For each sample



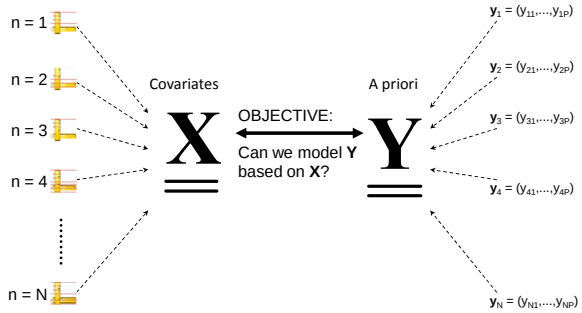
Data as observations

- For each sample



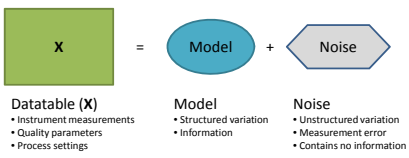
Data as observations

- For each sample



Principal Component Analysis (PCA)

- Projection method
- Exploratory data analysis
- Extract information and remove noise
- Reduce dimensionality / Compression
- Classification and clustering

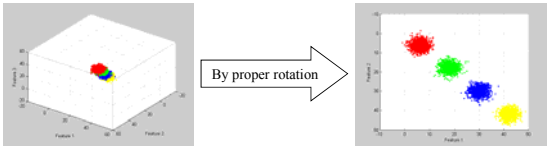


PCA in a nutshell

- First...
  ... The illustrational/intuitive approach:

Projection of data

- Linear transformation

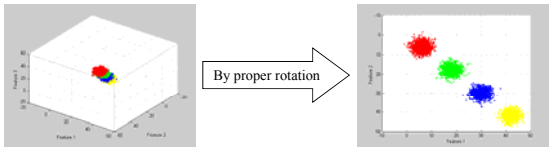


Feature observations in 3D space

Feature observations in reduced 2D space

Projection of data

- By proper linear transformation



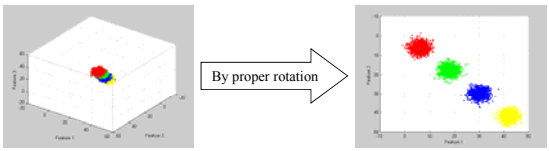
Feature observations in 3D space

Feature observations in reduced 2D space

- The PCA approach:
  - Rotation according to maximum variance in data.

Projection of data

- By proper linear transformation



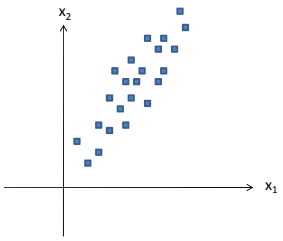
Feature observations in 3D space

Feature observations in reduced 2D space

- The PCA approach:
  - Rotation according to maximum variance in data.
- Fisher approach (**later lecture**):
  - Rotation according to maximum discrimination between groups.

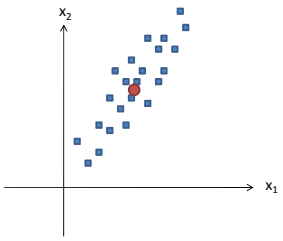
Principal Component Analysis

- For a given dataset:



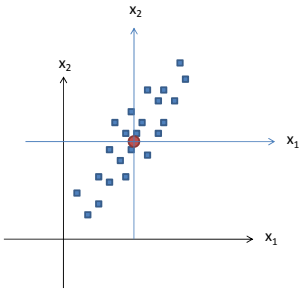
Principal Component Analysis

- Calculate the centroid (=”mean in all directions”):



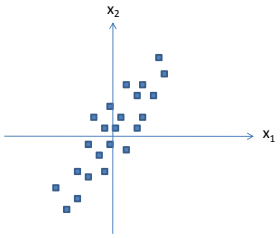
Principal Component Analysis

- Shift the grid to the centroid:



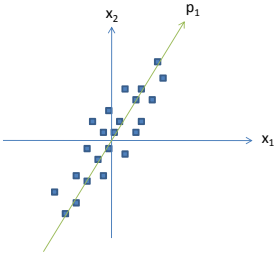
Principal Component Analysis

- Take this as our new coordinate system:



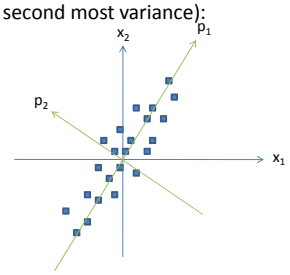
Principal Component Analysis

- Calculate the direction in which the variance is maximal:



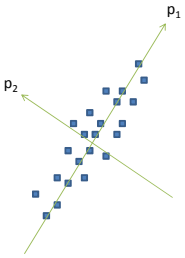
Principal Component Analysis

- And repeat this for each next perpendicular axis (direction with second most variance):



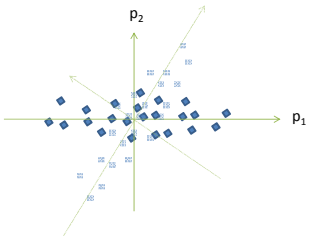
Principal Component Analysis

- Leaving us with a rotated grid:



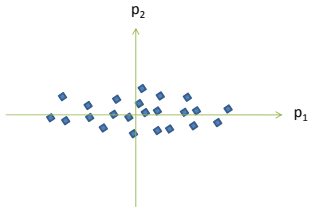
Principal Component Analysis

- Which we can rotate to a “normal” position:



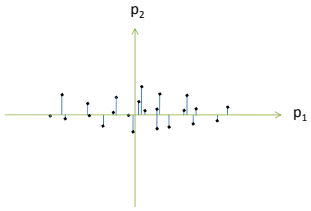
Principal Component Analysis

- Showing us maximal variance:



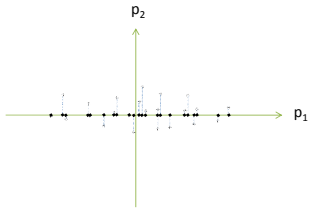
Principal Component Analysis

- We can also use this to reduce the complexity of the data set:



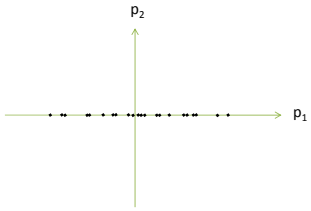
Principal Component Analysis

- By eliminating a number of axis by projection of the points:



Principal Component Analysis

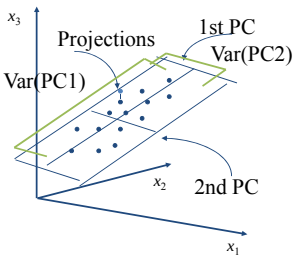
- In this example moving from two…:



Principal Component Analysis

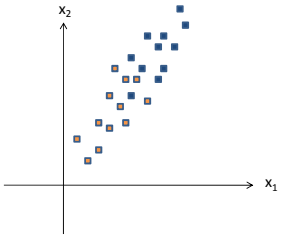
- … to one dimensional data points:

In 3D



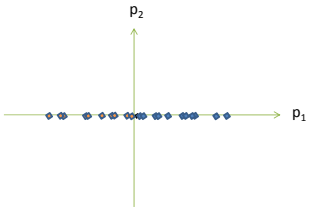
Principal Component Analysis

- Assumuming variation equals species diversity...:



Principal Component Analysis

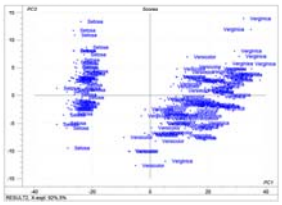
- ... the first PCA depicts this information...:



Scores and Loadings

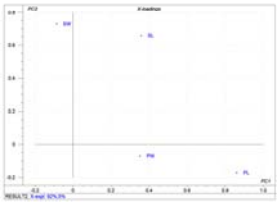
Scores

- Map of samples
- Displays distribution of samples in the new space defined by the PC's



Loadings

- Map of variables
- Shows how the original variables are related to the PC's



Principal Component Analysis

- We center the data.
  - But what about scaling?
- Especially when the covariates (columns) represent different scales (i.e. comparing appels and bananas) its important to divide by the standard deviation!

PCA in a nutshell

- Next...
  ... lets describe the illustrations in mathematical terms:

Some (boring) definitions – 1 of 2

- PCA is mathematically defined as an **orthogonal linear transformation** that transforms the data to a new coordinate system **such that the greatest variance by any projection of the data comes to lie on the first coordinate** (called the first principal component), the second greatest variance on the second coordinate, and so on.

Some (boring) definitions – 2 of 2

- PCA can be used for **dimensionality reduction** in a data set **by retaining those characteristics of the data set that contribute most to its variance**, by keeping lower-order principal components and ignoring higher-order ones.
- Such low-order components often contain the "most important" aspects of the data.
  - I.e. **assuming that information is equal variation**!
- However, depending on the data this may (obviously) not always be the case.

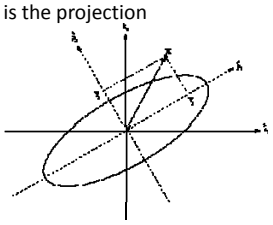
PCA in a nutshell

- Finally... Lets wrap it up in math:

Principal Component Analysis

- The i'th principal component of $\underline{\mathbf{X}}$ is the projection

$$\underline{\mathbf{Y}}_i = \underline{\mathbf{p}}_i^T \underline{\mathbf{X}}$$



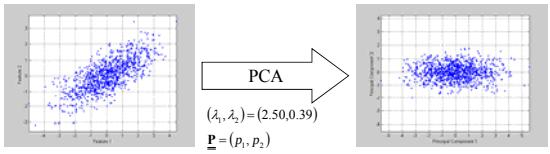
- The vector $\underline{\mathbf{Y}}$

$$\underline{\mathbf{Y}} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_k \end{pmatrix} = \underline{\underline{\mathbf{P}}}^T \underline{\mathbf{X}}$$

is called the vector of principal components.

$$\mathrm{cov}(\underline{\mathbf{Y}}) = \mathrm{cov}\left(\underline{\underline{\mathbf{P}}}^T \underline{\mathbf{X}}\right) = \underline{\underline{\mathbf{P}}}^T \underline{\underline{\boldsymbol{\Sigma}}} \underline{\underline{\mathbf{P}}} = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_k \end{pmatrix}$$

Principal Component Analysis

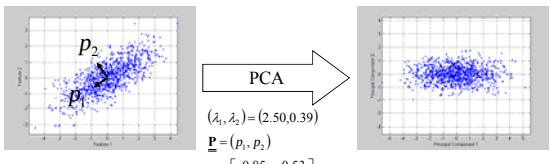


$\boldsymbol{\mu} = (0.0,0.0)$
$\boldsymbol{\Sigma} = \begin{bmatrix} 1.87 & 0.96 \\ 0.96 & 0.99 \end{bmatrix}$

$(\lambda_1, \lambda_2) = (2.50,0.39)$
$\underline{\mathbf{P}} = (p_1, p_2)$
$= \begin{bmatrix} -0.85 & 0.53 \\ -0.53 & -0.85 \end{bmatrix}$

$\boldsymbol{\mu} = (0.0,0.0)$
$\boldsymbol{\Sigma} = \begin{bmatrix} 2.50 & 0.00 \\ 0.00 & 0.39 \end{bmatrix}$

Principal Component Analysis

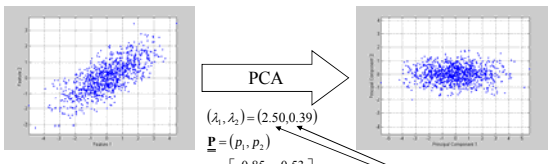


$\boldsymbol{\mu} = (0.0,0.0)$
$\boldsymbol{\Sigma} = \begin{bmatrix} 1.87 & 0.96 \\ 0.96 & 0.99 \end{bmatrix}$

$(\lambda_1, \lambda_2) = (2.50,0.39)$
$\underline{\mathbf{P}} = (p_1, p_2)$
$= \begin{bmatrix} -0.85 & 0.53 \\ -0.53 & -0.85 \end{bmatrix}$

$\boldsymbol{\mu} = (0.0,0.0)$
$\boldsymbol{\Sigma} = \begin{bmatrix} 2.50 & 0.00 \\ 0.00 & 0.39 \end{bmatrix}$

Principal Component Analysis



$\boldsymbol{\mu} = (0.0,0.0)$
$\boldsymbol{\Sigma} = \begin{bmatrix} 1.87 & 0.96 \\ 0.96 & 0.99 \end{bmatrix}$

$(\lambda_1, \lambda_2) = (2.50,0.39)$
$\underline{\mathbf{P}} = (p_1, p_2)$
$= \begin{bmatrix} -0.85 & 0.53 \\ -0.53 & -0.85 \end{bmatrix}$

$\boldsymbol{\mu} = (0.0,0.0)$
$\boldsymbol{\Sigma} = \begin{bmatrix} 2.50 & 0.00 \\ 0.00 & 0.39 \end{bmatrix}$

Principal Component Analysis

- From the PCA we may extract the set of linear combinations that explains the most variation

$$\frac{\lambda_1 + \cdots + \lambda_m}{\lambda_1 + \cdots + \lambda_m + \cdots + \lambda_k} \geq q$$

- And hereby condense and reduce the dimensionality of the featurespace.
- From the example before we see, that the linear combinations explain

$$\underline{\underline{\mathbf{P}}} = (p_1, p_2) = (86.3\%, 13.7\%)$$

of the total variance.

Output

- Loadings
  - The weights
- Scores
- Plots
  - Loadings plot
  - Scores plot
  - Biplot

Pros and Cons

- Positives
  - Can deal with large data sets.
  - There weren't done any assumptions on the data. This method is general and may be applied to any data set.
- Negatives
  - Nonlinear structure is invisible to PCA
  - The meaning of features is lost when linear combinations are formed
- Still!!!
  - Nonlinear PCA's exist (so-called kernel methods)
  - Sparseness or supervised projections can be introduced to emphasize important features

Course 27411 – Exercises 11/2 2008

PCA Exercise 1, Fisher *Iris* data

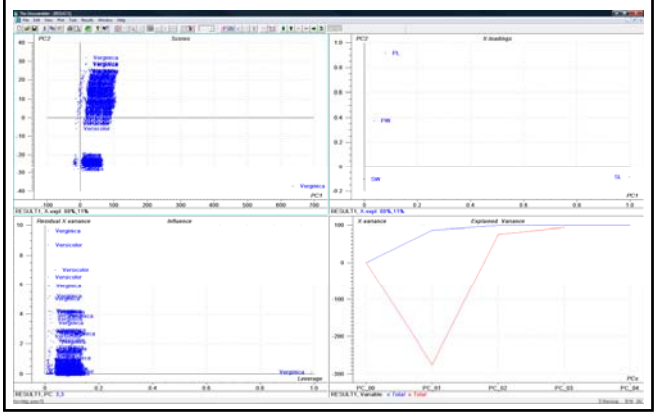
The Fisher Iris data-set is classical[1,2]. There are 150 objects, 50 *Iris setosa*, 50 *Iris versicolor* and 50 *Iris virginica*. The flowers of these 150 plants have been measured by a ruler. The variables are sepal length, sepal width, petal length and petal width (4 variables).

The original hypothesis was that *I. versicolor* was a hybrid of the two other species i.e. *I. setosa* x *virginica*. *I. setosa* is a diploid species; *I. virginica* is a tetrapoid and *I. versicolor* is a hexaploid.

1. Start The Unscrambler and open the help function. Lookup PCA and read about the different options and plots provided by the software.
   In general: Use the help function before asking! ☺ To get help on the different plots mark the plot and press F1.
2. Import the excel file fisherout.xls to UNSCRAMBLER and save the UNSCRAMBLER file
3. Usually you should examine the raw data first, but in this exercise try to perform a PCA, with "leverage correction" and with centering. Look at the four standard plots.
   a. How many principal components would you need and what does the PC1 describe?
   b. How many % of the variation is described by the first two PCs?
   c. Do you see anything "wrong" in the score-plot?
   d. If so what is wrong and what should be done?
   e. Try to correct it!
4. Does it help to standardize the data matrix and perform a new PCA (try it!).
5. How many PCs are needed to explain 70, 75% or 90% of the variance?
6. What does PC1, PC2 and PC3 show?
7. Are any variables more discriminative than others? Are any variables dispensable?
8. Can you see the presupposed classes? Any class overlap?

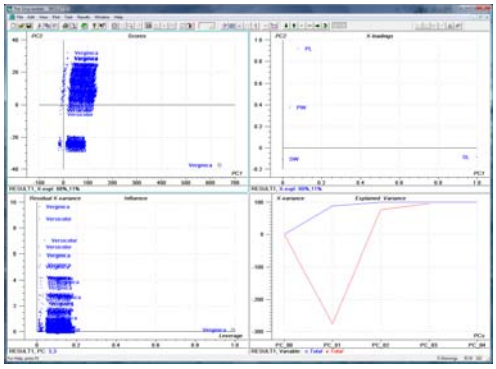
[1] Fisher, R.A. (1936). "The Use of Multiple Measurements in Taxonomic Problems". *Annals of Eugenics* 7: 179–188.
[2] Edgar Anderson (1935). "The irises of the Gaspe Peninsula". *Bulletin of the American Iris Society* 59: 2–5.

Exercise



Exercise

Exercise



Exercise