Principal component analysis

Course 27411
Biological dataanalysis and chemometrics

Thanks to Lars Nørgaard, CAMO, Michael Edberg and Harald Martens for some of the figures

Jens C. Frisvad
Department of Systems Biology
Building 221
Technical University of Denmark
2800 Kgs. Lyngby – Denmark
E-mail: jcf@bio.dtu.dk

Principal component analysis
OUTLINE

• Some matrix algebra
• What is PCA?
• Goals of PCA
• Projection
• Q and R analysis
• Matrix notation
• Double example (continous variables and a series of single variables)
• Connection to PCR and PLSR
• Validation (short)
• How to do it!
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

- **Matrix**: rectangular array of scalars

<table>
<thead>
<tr>
<th>A</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1 2 3]</td>
<td>[1 4]</td>
<td>d₁₁</td>
</tr>
<tr>
<td>5 4 1</td>
<td>2 7 3</td>
<td>d₂₁</td>
</tr>
<tr>
<td>6 7 4</td>
<td>3 8 3</td>
<td>d₃₁</td>
</tr>
</tbody>
</table>

Square (3 x 3)    Rectangular (3 x 2)

- Column vector: \( \mathbf{b} = \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix} \)
- Row vector: \( \mathbf{d} = \begin{bmatrix} 3 & 4 & 9 \end{bmatrix} \)

Vector Operations

- **Transpose operator**

  \[ \mathbf{b}^T = \begin{bmatrix} 1 & 1 & 2 \end{bmatrix} \]

  \[ \mathbf{d}^T = \begin{bmatrix} 3 \\ 4 \\ 9 \end{bmatrix} \]

- **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
  \[ \mathbf{x}^T \mathbf{y} = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = x_1y_1 + x_2y_2 + x_3y_3 = \sum_{i=1}^{3} x_i y_i \]

- Length of a vector
  \[ \| \mathbf{x} \| = (x_1^2 + x_2^2 + x_3^2)^{1/2} \]

Matrix Operations

- Addition (matrix of same size)
  - Commutative: \( \mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A} \)
  - Associative: \( (\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C}) \)

\[ \mathbf{A} + \mathbf{B} = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix} + \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 3 & 3 \\ 3 & 3 \end{bmatrix} \]
Matrix Operations

• Multiplication
  - number of columns in first matrix = number of rows in second
  - Associative: \((A B) C = A (B C)\)
  - Distributive: \(A (B+C) = A B + A C\)
  - Not commutative: \(AB \neq BA!!!\)
  - \((A B)^T = B^T A^T\)

\[
\begin{bmatrix}
1 & 2 & 3 \\
4 & 5 & 6 \\
2 & 3 & 2
\end{bmatrix}
\begin{bmatrix}
7 & 1 \\
8 & 2 \\
9 & 3
\end{bmatrix}
= \begin{bmatrix}
1 \times 7 + 2 \times 8 + 3 \times 9 & 1 \times 1 + 2 \times 2 + 3 \times 3 \\
4 \times 7 + 5 \times 8 + 6 \times 9 & 4 \times 1 + 5 \times 2 + 6 \times 3
\end{bmatrix}
= \begin{bmatrix}
50 & 14 \\
122 & 32
\end{bmatrix}
\]

Some Definitions ...

• Identity Matrix
  \[
  I = \begin{bmatrix}
  1 & 0 & 0 \\
  0 & 1 & 0 \\
  0 & 0 & 1
  \end{bmatrix}
  \]
  \[IA = AI = A\]

• Diagonal Matrix
  \[
  D = \begin{bmatrix}
  3 & 0 & 0 \\
  0 & 5 & 0 \\
  0 & 0 & 7
  \end{bmatrix}
  \]

• Symmetric Matrix
  \[
  B = \begin{bmatrix}
  1 & 5 & 2 \\
  5 & 2 & 7 \\
  2 & 7 & 1
  \end{bmatrix}
  \]
  \[B = B^T \quad b_{ij} = b_{ji}\]
Matrix Inverse

• $A^{-1} A = A A^{-1} = I$

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

Properties

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

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

$(A B)^{-1} = B^{-1} A^{-1}$;  \hspace{1cm} $B^{-1} A^{-1} A B = B^{-1} B = I$

$(A^T)^{-1} = (A^{-1})^T$;  \hspace{1cm} $(A^{-1})^T A^T = (A A^{-1})^T = I$

Data as observations

• Samples as observations in a Multi/hyper-dimensional Space:
  - Objects are characterized by a profile 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

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

- Developed by Karl Pearson in 1901

- Also called:
  - Singular value decomposition
  - Karhunen-Loéve expansion
  - Eigenvector analysis
  - Latent vector analysis
  - Characteristic vector analysis
  - Hotelling transformation

Principal Component Analysis (PCA)

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

\[
X = \text{Model} + \text{Noise}
\]

**Datatable (X)**
- Instrument measurements
- Quality parameters
- Process settings

**Model**
- Structured variation
- Information

**Noise**
- Unstructured variation
- Measurement error
- Contains no information
Goals of PCA

• Simplification
• Data reduction
• Modeling
• Outlier detection
• Variable selection
• Classification (unsupervised)
• Unmixing (curve resolution)

PCA, which distance measure is used?

• PCA uses Euclidean distance!
PCA in a nutshell

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

Projection of data

- Linear transformation

![Feature observations in 3D space](image1.png) ![Feature observations in reduced 2D space](image2.png)
Projection of data

• By proper linear transformation

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

• Fisher approach:
  – 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

• Subtract the mean:


Principal Component Analysis

• Take this as our new coordinate system:
Principal Component Analysis

• Calculate the direction in which the variance is maximal:

\[ x_2 \]
\[ x_1 \]

\[ p_1 \]

Principal Component Analysis

• And repeat this for the next orthonormal axis (direction with second most variance):

\[ x_2 \]
\[ x_1 \]

\[ p_1 \]
\[ p_2 \]
Principal Component Analysis

• Leaving us with a rotated grid:

![Diagram of a rotated grid](image)

Principal Component Analysis

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

![Diagram of a normal position](image)
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

- Assuming variation equals species diversity...:

\[ x_2 \]
\[ x_1 \]

Principal Component Analysis

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

\[ p_2 \]
\[ p_1 \]
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

Data as observations
There are often two matrices to compare

For each sample

- \( n = 1 \)  
- \( n = 2 \)  
- \( n = 3 \)  
- \( n = 4 \)  
- \( n = N \)  

\[ \begin{bmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1M} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2M} \\ x_{31} & x_{32} & x_{33} & \cdots & x_{3M} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & x_{N3} & \cdots & x_{NM} \end{bmatrix} \]

\[ \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} y_{11} & \cdots & y_{1P} \\ y_{21} & \cdots & y_{2P} \\ y_{31} & \cdots & y_{3P} \\ \vdots & \ddots & \vdots \\ y_{N1} & \cdots & y_{NP} \end{bmatrix} \]
Data as observations

- For each sample
  \( n = 1 \)
  \( n = 2 \)
  \( n = 3 \)
  \( n = 4 \)
  \( \cdots \)
  \( n = N \)

Covariates \( X \)

A priori \( Y \)

\( y_1 = (y_{11}, \ldots, y_{1p}) \)
\( y_2 = (y_{21}, \ldots, y_{2p}) \)
\( y_3 = (y_{31}, \ldots, y_{3p}) \)
\( y_4 = (y_{41}, \ldots, y_{4p}) \)
\( \cdots \)
\( y_N = (y_{N1}, \ldots, y_{NP}) \)

Data as observations

Extensions of PCA: PCR, PLS etc.

- For each sample
  \( n = 1 \)
  \( n = 2 \)
  \( n = 3 \)
  \( n = 4 \)
  \( \cdots \)
  \( n = N \)

Covariates \( X \)

OBJECTIVE: Can we model \( Y \) based on \( X \)?

A priori \( Y \)

\( y_1 = (y_{11}, \ldots, y_{1p}) \)
\( y_2 = (y_{21}, \ldots, y_{2p}) \)
\( y_3 = (y_{31}, \ldots, y_{3p}) \)
\( y_4 = (y_{41}, \ldots, y_{4p}) \)
\( \cdots \)
\( y_N = (y_{N1}, \ldots, y_{NP}) \)
PCA

• PCA is used for analysis of one datamatrix, in a hope to distinguish latent signals from noise in data

\[
x = X + T_x \times P' + E
\]

Meas. data  Av. of all X  Scores  Loadings  Errors

Nr. samples (n)  Nr. variables (p)
PCA is depending on feature size and scale

- Variables = features = characteristics = attributes = characters
- When in anyone variable the largest value is 8 times greater than the lowest value, use log transformation $x_{\log 1} = \log (x+1)$
- Often the data are centered: $x_{ik \text{ (new)}} = x_{ik} - x_i$

### Mean, variance and standard deviation

Mean (avg)

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Variance

$$Var(x) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

Standard deviation (std)

$$S_x = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$
Autoscaling = standardization

• To avoid the problem of different scales the data are often centered **and** all values are divided by the standard deviation for each variable

\[ X_{\text{aut}} = \frac{X - X_{\text{avg}}}{X_{\text{std}}} \]

Q versus R analysis

• Sometimes the object could as well be the variables, in that case the matrix X can be transposed (to X')
• Objects are rarely transformed regarding the variables
• Normalization is rarely done (all variables have sum 1 or 100%): one risks closure of the data
  – Fx used in GC: % area (but do not use this as raw data, rather standardize your method)
Objects and variables

- In contingency tables, objects and variables may have equal status. Often correspondence analysis is used on such data.
- In some cases only the objects or only the variables are interesting.
  - Example: Fluorescence or other spectrophotometric data are occasionally used just for characterization of the objects. Those soft curves are often corrected f.ex. by Multiple Scatter Correction before PCA, but they are not always used in the final evaluation of the data.

Spectral measurements on sugar samples from three different factories (B, E, F): 34 samples and 1023 variables (Lars Nørgaard example)
Raw data

Centering

Lars Nørgaard, Foss
Two dimensional score plot, PC1 vs. PC2

Two-dimensional score plot, PC1 vs. PC3

Lars Nørgaard, Foss
Chemical/physical measurements on sugar samples from 3 different factories (B, E, F); 34 samples, 10 variables

Lars Nørgaard, Foss
Score plot

Principal component analysis

Lars Nørgaard, Foss
Loading plot

Biplot

Lars Nørgaard, Foss
**NIPALS algorithm**

**Non-linear Iterative PArtial Least Squares**

1. Center or autoscale X
2. Choose a random column in X as a start guess on t
3. Solve $X = tp' + E$ in regard to p; "$p = X/t$" (normalize $p$ to length one: $p = p/|p|$)
4. Solve $X = tp' + E$ in regard to t: "$t = X/p''$
5. Repeat step 3 & 4 until convergence
6. Orthogonalize $X$: $X_{\text{new}} = X - tp$
7. Start from 2. again with $X_{\text{new}}$ to extract another factor (scores and loadings)

---

**Principal Component Analysis**

- The $i$'th principal component of $X$ is the projection
  \[ Y_i = p_i^T X \]

- The vector $Y$
  \[ Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_i \end{bmatrix} = P^T X \]
  is called the vector of principal components.
  \[ \text{cov}(Y) = \text{cov}(P^T X) = P^T \Sigma P = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_i \end{pmatrix} \]
Principal Component Analysis

- New variables $Y_i$ that are linear combination of the original variables ($x_i$):
  
  $$Y_i = a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{ip}x_p ; i = 1..p$$

- The new variables $Y_i$ are derived in decreasing order of importance.

- They are called “principal components”.

Principal Component Analysis

- The direction of $p_1$ is given by the eigenvector $\lambda_1$ corresponding to the largest eigenvalue of matrix $C$.
- The second vector that is orthogonal (uncorrelated) to the first is the one that has the second highest variance which comes to be the eigenvector corresponding to the second eigenvalue.
- And so on ...
Principal Component Analysis

• Let $C = \text{cov}(X)$

• The eigenvalues $\lambda_i$ of $C$ are found by solving the equation

$$\det(C - \lambda I) = 0$$

• Eigenvectors are columns (rows) of the matrix $P$

Example

• Let us take two variables with covariance $c > 0$

$$C = \text{cov}(X) = \begin{pmatrix} 1 & c \\ c & 1 \end{pmatrix}$$

• Solving the eigenvalue problem

$$\det(C - \lambda I) = \det\begin{pmatrix} 1 - \lambda & c \\ c & 1 - \lambda \end{pmatrix} = 0$$

$$\Leftrightarrow (1 - \lambda)^2 - c^2 = 0$$

• Solving this we find: $\lambda_1 = c + 1$

$$\lambda_2 = c - 1 (< \lambda_1)$$
Principal Component Analysis

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

PCA

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

\[ \mu = (0.0, 0.0, 0.0) \]
\[ \Sigma = \begin{bmatrix} 2.50 & 0.00 \\ 0.00 & 0.39 \end{bmatrix} \]
Principal Component Analysis

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

\[
\frac{\lambda_1 + \cdots + \lambda_m}{\lambda_1 + \cdots + \lambda_m + \cdots + \lambda_q} \geq q
\]

- And hereby condense and reduce the dimensionality of the feature space.

- From the example before we see, that the linear combinations explain

\[
\mathbf{P} = (p_1, p_2) = (86.3\%, 13.7\%)
\]

of the total variance.
Output

• Loadings
  – The weights
• Scores
• Plots
  – Loadings plot
  – Scores plot
  – Biplot

Overview of PCA

• Select objects (Sneath rule of thumb: use 24 objects per supposed class)
• Select variables (in general use many)
• Examine the raw data (plot and calculate averages and standard deviations)
• Consider transformations of raw data
• Perform the PCA
• Use cross-validation to estimate the number of components A
• Other tests for the number of components can also be used (broken stick etc.)
• In general it is good to have more than 75 % of the variance in the data explained
• Plot scores
• Plot loadings (maybe also biplots)
• Look for outliers (representatives probably of a new class (or more often an error) (leverage or high residual outliers)
• Be conservative: Do not take too many (object) outliers out before a new PCA
Validation

• Cross-validation (will be explained in detail in lecture on PCR)
• Jack-knifing
• Bootstrapping
• PTP analysis & Monte Carlo simulations

Crossvalidation (number of PC’s ?)

- Predicted error on Y
- Under fit
- Over fit
- Optimum
- Number of PC’s
Crossvalidation (methods)

• “Leave one out” or “full validation” leaves one sample out each time

• Segmented (category) validation
  • On replicates (on independent replicates), leaves one replicate group (one segment) out each time
  • Mixtures, leaves one mixture group (one segment) out each time

• Remember that repeated experiments are dangerous to use in cross-validation (i.e. use real biological/chemical replicates!!)
Bootstrap analysis

1. Cladogram calculated from original data matrix

<table>
<thead>
<tr>
<th>Characters</th>
<th>1 2 3 4 5 6 7 8 9 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Taxon A</td>
<td>0 0 0 0 1 0 0 1 0</td>
</tr>
<tr>
<td>Taxon B</td>
<td>1 1 0 0 0 0 0 0 0</td>
</tr>
<tr>
<td>Taxon C</td>
<td>1 1 1 1 1 1 1 2 0</td>
</tr>
<tr>
<td>Taxon D</td>
<td>1 1 1 0 1 1 2 0 1</td>
</tr>
<tr>
<td>Taxon E</td>
<td>1 1 0 0 1 1 1 1 0</td>
</tr>
</tbody>
</table>

2. Data matrix resampled at random with replacement to construct new data matrix and cladogram

<table>
<thead>
<tr>
<th>Characters</th>
<th>1 2 3 4 5 6 7 8 9 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Taxon A</td>
<td>1 0 0 0 1 0 0 0 0</td>
</tr>
<tr>
<td>Taxon B</td>
<td>0 0 0 0 0 0 0 0 1 0</td>
</tr>
<tr>
<td>Taxon C</td>
<td>0 1 1 0 1 1 1 1 1</td>
</tr>
<tr>
<td>Taxon D</td>
<td>2 0 1 1 1 1 1 1 1</td>
</tr>
<tr>
<td>Taxon E</td>
<td>1 1 0 1 1 0 0 1 1</td>
</tr>
</tbody>
</table>

3. Repeat until 1,000 randomly constructed data matrices have been analysed. Bootstrap value = % of times branch appears in 1,000 cladograms

PTP analysis

1. Cladogram calculated from original data matrix and length recorded

<table>
<thead>
<tr>
<th>Characters</th>
<th>1 2 3 4 5 6 7 8 9 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Taxon A</td>
<td>0 0 0 0 0 0 1 0 1 0</td>
</tr>
<tr>
<td>Taxon B</td>
<td>1 1 0 0 0 0 0 0 0 0</td>
</tr>
<tr>
<td>Taxon C</td>
<td>1 1 1 1 1 1 1 2 0 1</td>
</tr>
<tr>
<td>Taxon D</td>
<td>1 1 1 1 0 1 1 2 0 1</td>
</tr>
<tr>
<td>Taxon E</td>
<td>1 1 0 0 1 0 1 1 1 0</td>
</tr>
</tbody>
</table>

2. Data within columns is randomized to construct new data matrix of same size. A new cladogram is constructed and its length recorded.

<table>
<thead>
<tr>
<th>Characters</th>
<th>1 2 3 4 5 6 7 8 9 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Taxon A</td>
<td>1 1 0 1 0 0 0 0 0 1</td>
</tr>
<tr>
<td>Taxon B</td>
<td>1 1 1 0 2 1 1 1 1 0</td>
</tr>
<tr>
<td>Taxon C</td>
<td>1 0 0 0 1 1 1 2 0 0</td>
</tr>
<tr>
<td>Taxon D</td>
<td>1 1 1 0 1 1 0 0 1 0</td>
</tr>
<tr>
<td>Taxon E</td>
<td>1 1 0 1 0 1 0 1 0 1</td>
</tr>
</tbody>
</table>

3. Repeat until 10,000 randomly constructed data matrices have been analysed. The cladogram has significant structure if the original tree length is less than the 95% tail of random tree lengths
Can we always use PCA?
(test not used in UNSCRAMBLER, but in NTSYS)

• Bartlett’s (1950) sphericity test (an approximate chi-squared test)

\[ \text{chi}^2 \left[ \frac{(p^2 - p)}{2} \right] = - \left[ (n-1) - \frac{(2p + 5)}{6} \right] \ln R \]

• In R is the natural log of the determinant of the correlation matrix
• The p formula is the number of degrees of freedom associated with the chi-square test statistic
• p = the number of variables
• n = the number of observations

Bartlett’s (1951a,b) test whether all of the roots except the first are equal
(in NTSYS under Eigenvectors)

\[ \text{chi}^2 = (n - \frac{2p + 5}{6}) - \frac{2}{3} (\sum_{i=2}^{p} \ln \hat{\lambda}_i + (p-1) \ln \rho) \]

where \( \rho = 1 - \frac{1}{p} \sum_{i=2}^{p} \hat{\lambda}_i \)

d.f.: \( \frac{(p-2)(p-1)}{2} - \frac{(p-1)(p-2)(1-\rho^2)}{p} \)
UNSCRAMBLER EXAMPLE

• Datafile: Fischerout, but add a first variable telling the grouping (1, 2, 3)
• Task: PCA, select objects (all), select variables (2-5: Iris petal & sepal width and length)
• Use autoscaling & cross-validation
Colouring the groups

- You may colour each class (if you know them!)
  - Fx use PCA on each of the classes, and create a group for each class, when you have these classes, you use the for sample grouping (right click on score plot) (you can use these classes for SIMCA later!)

- An alternative: use K-means clustering or fuzzy clustering to get the variable showing the groups (classes)
Pros and Cons of PCA

• Positives
  – Can deal with large datasets (both in objects and variables)
  – There are no special assumptions on the data and PCA can be applied on all data-sets

• Negatives
  – Non-linear structure is hard to model with PCA
  – The meaning of the original variables may be difficult to assess directly on latent variables (but use the loading plot) or Varimax, factor analysis etc.

• Still:
  – Non-linear PCA methods exist (kernel methods)
  – Sparseness of supervised projections can be introduced to emphasize important features

Exercises in PCA

• The classical Iris data set (150 Iris plants characterized by 4 variables): Fisherout (Excell file)

• The wine data set (178 wines characterized by 13 variables): VIN2 (Excell file): for examn!