Mathematics for Systems Biology and Bioinformatics

Lecture Prof. Dr. Thomas Filk

Tutorials Dr. Tim Maiwald, Christian Tönsing

Exercise sheet no. 9

Submission until 9.1.2013 10:00 am in the tutorials

Homework 14: Principal Component Analysis (PCA) (10 Points)

Since pattern in data can be hard to find in data of high dimension, PCA is a powerful tool for analyzing data where the luxury of graphical representation is not available.

Perform a PCA on this 2D data set:

x	2.5	0.5	2.2	1.9	3.1	2.3	2.0	1.0	1.5	1.1
y	2.4	0.7	2.9	2.2	3.0	2.7	1.6	1.1	1.6	0.9

a) Visualize the data in a coordinate system. You will notice, that all the data lies within the upper right quadrant. Center the data around (0,0) by subtracting the mean of the variables in *each* dimension. ($X_i = x_i - \bar{x}$). Draw the centered data in *another* coordinate system.

b) Calculate the covariance matrix of the centered data set.

$$C = \begin{pmatrix} Cov(X,X) & Cov(X,Y) \\ Cov(X,Y) & Cov(Y,Y) \end{pmatrix}$$

$$Cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)}$$

c) Calculate the two eigenvalues and eigenvectors of the covariance matrix C. Note, that you will need the eigenvectors of length 1 for all further calculations.
Draw the eigenvectors in the coordinate system with the centered data set.

d) If you look at the eigenvectors and eigenvalues, you will notice that the eigenvalues are quite different values. In fact, it turns out that the eigenvector with the *highest* eigenvalue is the *principal component* of the data set.
Explain shortly which property of the data set is described via the principal component.

e) We now want to transform the initial centered data set in direction of largest eigenvector using the following equation:

$$D_{trans} = E^T \cdot D_{initial} = \begin{pmatrix} e_1^{(1)} & e_2^{(1)} \\ e_1^{(2)} & e_2^{(2)} \end{pmatrix}^T \cdot \begin{pmatrix} x_1^c & x_2^c & \cdots & x_n^c \\ y_1^c & y_2^c & \cdots & y_n^c \end{pmatrix}.$$

where $e_\beta^{(\alpha)}$ denotes the β-th component of the α-th largest eigenvector. The data matrix $D_{initial}$ contains the centered initial data set (denoted by $\cdot^c$).

Draw the transformed data in a new coordinate system.

tim.maiwald@fdm.uni-freiburg.de christian.toensing@fdm.uni-freiburg.de
http://www.fdmold.uni-freiburg.de/∼ctoens/teaching/index.html