Wednesday, November 26, 2014

Multivariate analysis

Principal Component Analysis

How to find patterns in high dimensional multivariate data sets?

Principal component analysis (PCA) offers a means to identify patterns in high dimensional multivariate data sets, and to represent this data in a way that highlights similarities, differences, and groups.

PCA also allows to reduce the number of dimensions without loss of information (i.e. data compression).

"PCA is generally considered to be the working horse of multivariate data analysis, since so many methods are merely a variation on the same basic theme."

— K Faber

http://www.chemometry.com/Research/PCA.html
What is the basic idea behind PCA?

Eigen vectors and eigen values

Outline of the PCA algorithm:

1. Select a normalized direction in n-dimensional space along which the variance in X is maximized. Save this vector as p1.
2. Find the next direction along which variance is maximized; however, restricting the search to all directions orthogonal to all previous selected directions. Save this vector as p2.
3. Repeat this procedure until m vectors are selected.

The resulting ordered set of p’s are the principal components.

Linear algebra offers simple analytical solutions to the above algorithm.

Thus, PCA amounts to:

1. Organize the data as an m x n matrix, where m is the number of measurement types and n is the number of samples.
2. Subtract off the mean for each measurement type.
3. Calculate the SVD or the eigenvectors of the covariance matrix.
Principal Component Analysis

Application to multivariate data analysis and dimensionality reduction

Applications to image analysis, and alignment of structures
What kind of patterns in data will not be teased apart by PCA (or LDA)?
A word about
Clustering
and
Multidimensional Scaling

(analysis of the distances)

Support Vector Machines
(SVM)