Statistics and Quantitative Biology

Thursday, October 8, 2009

Multivariate Data,
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 $m$-dimensional space along which the variance in $X$ is maximized. Save this vector as $p_1$.
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 $p_i$.
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 data as an $m \times n$ 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
R: biplot(prcomp(data))
Some applications
Strongylocentrotus purpuratus

uncaged speract response
200 frames/s
Adan's visual method

Automated method based on eigen vectors of cov matrix of green dots
Which translation and rotation?
Mouthon et al. (1995) Invariance and restriction toward a limited set of self-antigens characterize neonatal IgM antibody repertoires and prevail in autoreactive repertoires of healthy adults. Proc Natl Acad Sci USA
Prakash, Fezel et al. (2006) Clusters of cytokines determine malaria severity in Plasmodium falciparum-infected patients from endemic areas of Central India. J Infect Diseases
What kind of patterns in data will not be teased apart by PCA (or LDA)?
FIG. 6 Example of when PCA fails (red lines). (a) Tracking a person on a ferris wheel (black dots). All dynamics can be described by the phase of the wheel $\theta$, a non-linear combination of the naive basis. (b) In this example data set, non-Gaussian distributed data and non-orthogonal axes causes PCA to fail. The axes with the largest variance do not correspond to the appropriate answer.