

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).

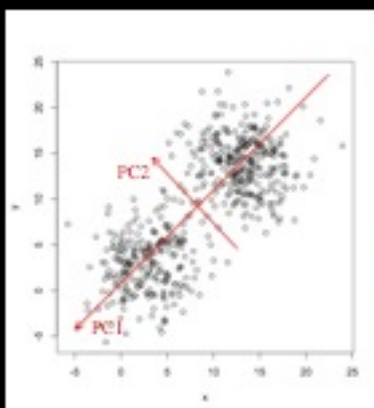
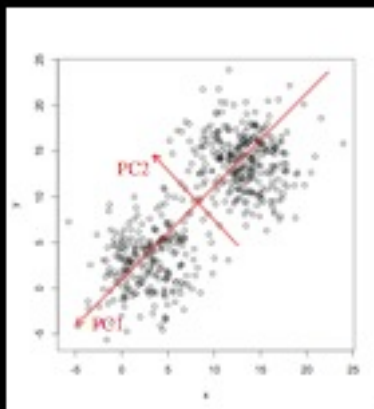
"PCA is generally considered to be the working horse of multivariate data analysis, since so many methods are merely a variation on the same basic theme."

– 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"

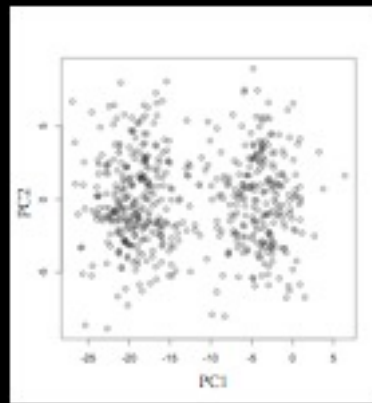
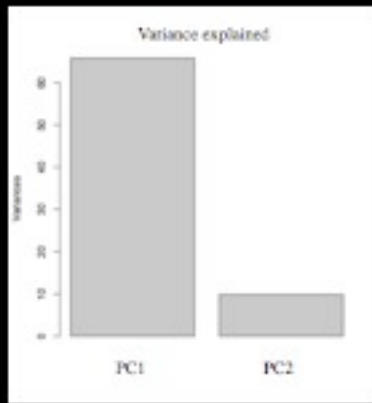
1. Select a normalized direction in  $m$ -dimensional space along which the variance in  $X$  is maximized. Save this vector as  $p_1$ .
2. Find the next direction along which variance is maximized, however, restricting the search to all directions orthogonal to all previous selected directions. Save this vector as  $p_2$ .
3. Repeat this procedure until  $m$  vectors are selected.

The resulting ordered set of  $p_i$ s are the principal components.

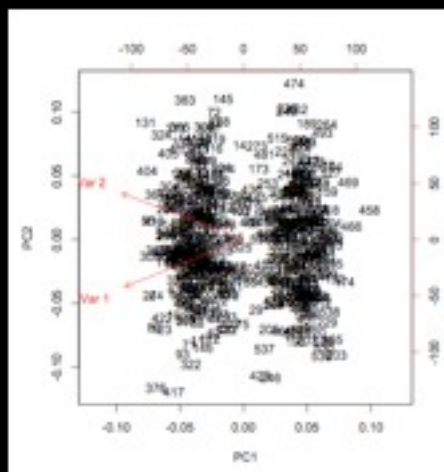
Linear algebra offers simple analytical solutions to the above algorithm.

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))**



## Principal Component Analysis

**Application** to multivariate data analysis  
and dimensionality reduction

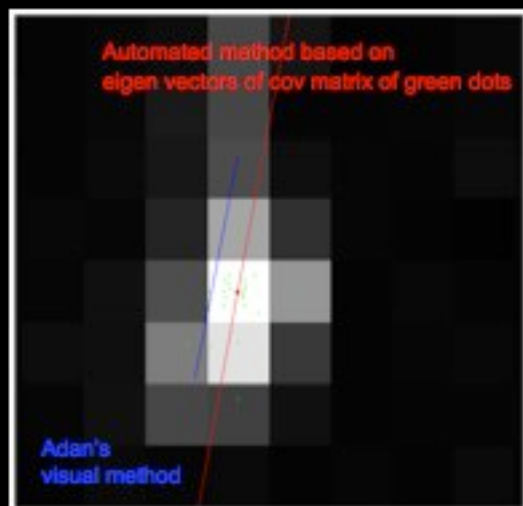
**Applications** to image analysis, and alignment of structures



*Strongylocentrotus purpuratus*



uncaged speract response  
200 frames/s



Automated method based on  
eigen vectors of cov matrix of green dots

Adan's  
visual method

## Which translation and rotation ?

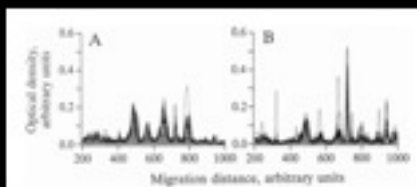


Table 1. Total variance of reactivities of light of newborns and adults with self- and non-self-antigens calculated by PCA.

Tissue extract	Individuals	Total variance	Cumulative variance, %	
			Factor 1	Factor 2
Kidney	Neonates	29.32	27	86
	Adults	71.22	56	77
Liver	Neonates	45.76	80	86
	Adults	622.06	62	75
Muscle	Neonates	117.86	56	80
	Adults	97.01	58	78
Stomach	Neonates	97.29	63	86
	Adults	267.87	59	77
Thymus	Neonates	25.04	49	68
	Adults	60.13	60	78
<i>P. aer.</i>	Neonates	40.54	75	87
	Adults	280.91	59	76
<i>E. coli</i>	Neonates	7.71	48	74
	Adults	125.06	49	67
<i>E. sh.</i>	Neonates	6.13	67	84
	Adults	46.91	64	82
<i>Bs. amp.</i>	Neonates	26.28	60	76
	Adults	431.68	59	84

Reactivities of light of newborns and of adults with antigens in extracts of five human tissues, of the bacteria *P. aeruginosa* (*P. aer.*), *E. coli* (*E. coli*), and *E. shigellae* (*E. sh.*) and of the plant *Thymus amplexicaulis* (*Th. amp.*) were analyzed separately in a 30- to 36-dimensional vector space. Total and cumulative variances of factor 1 and 2 for reactivities of light of newborns and adults are shown.

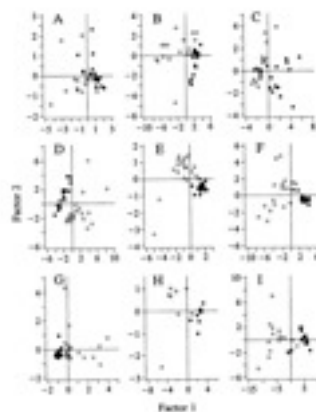
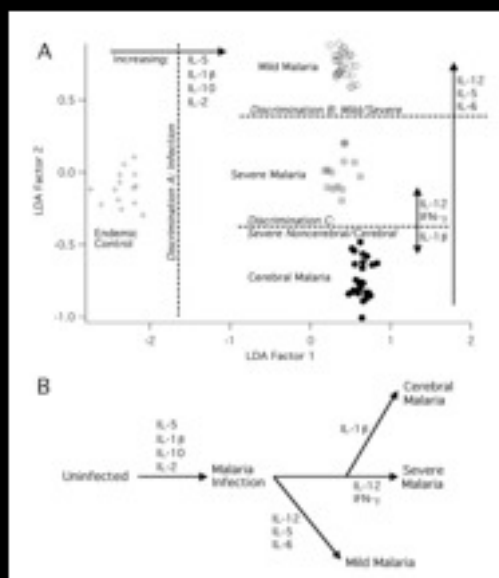


Fig. 3. PCA. Calculated areas under the curve corresponding to peaks of reactivity of light of newborns (■) and healthy young men (□) with self-antigens (*A-E*) and non-self-antigens (*P-E*) were subjected to PCA. The results are those obtained with light from the 15 neonates and 16 healthy men depicted in Figs. 1-4. Antigens were from kidney (*A*), liver (*B*), muscle (*C*), stomach (*D*), thymus (*E*), *P. aeruginosa* (*P*), *E. coli* (*E. coli*) (*E*), *E. shigellae* (*E. sh.*) (*E*), and *Thymus amplexicaulis* (*Th.*). Each individual is represented by a dot. The data relating to the 10 individuals were analyzed in a 30- to 36-dimensional vector space, depending on the tissue extract, and fitted within a two-dimensional linear subspace (factors 1 and 2) corresponding to 65-94% of the variance.



Prakash, Veer 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) ?

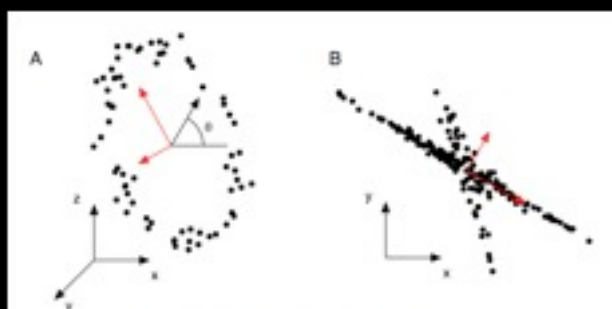


FIG. 6 Example of when PCA fails (red lines). (a) Tracking a person on a ferris wheel (black dots). All dynamics can be described by the phase of the wheel  $\theta$ , a non-linear combination of the naive basis. (b) In this example data set, non-Gaussian distributed data and non-orthogonal axes causes PCA to fail. The axes with the largest variance do not correspond to the appropriate answer.

**A word about  
Clustering  
and  
Multidimensional Scaling**

(analysis of the distances)

**Support Vector Machines  
(SVM)**