Exploration of Data is Critical

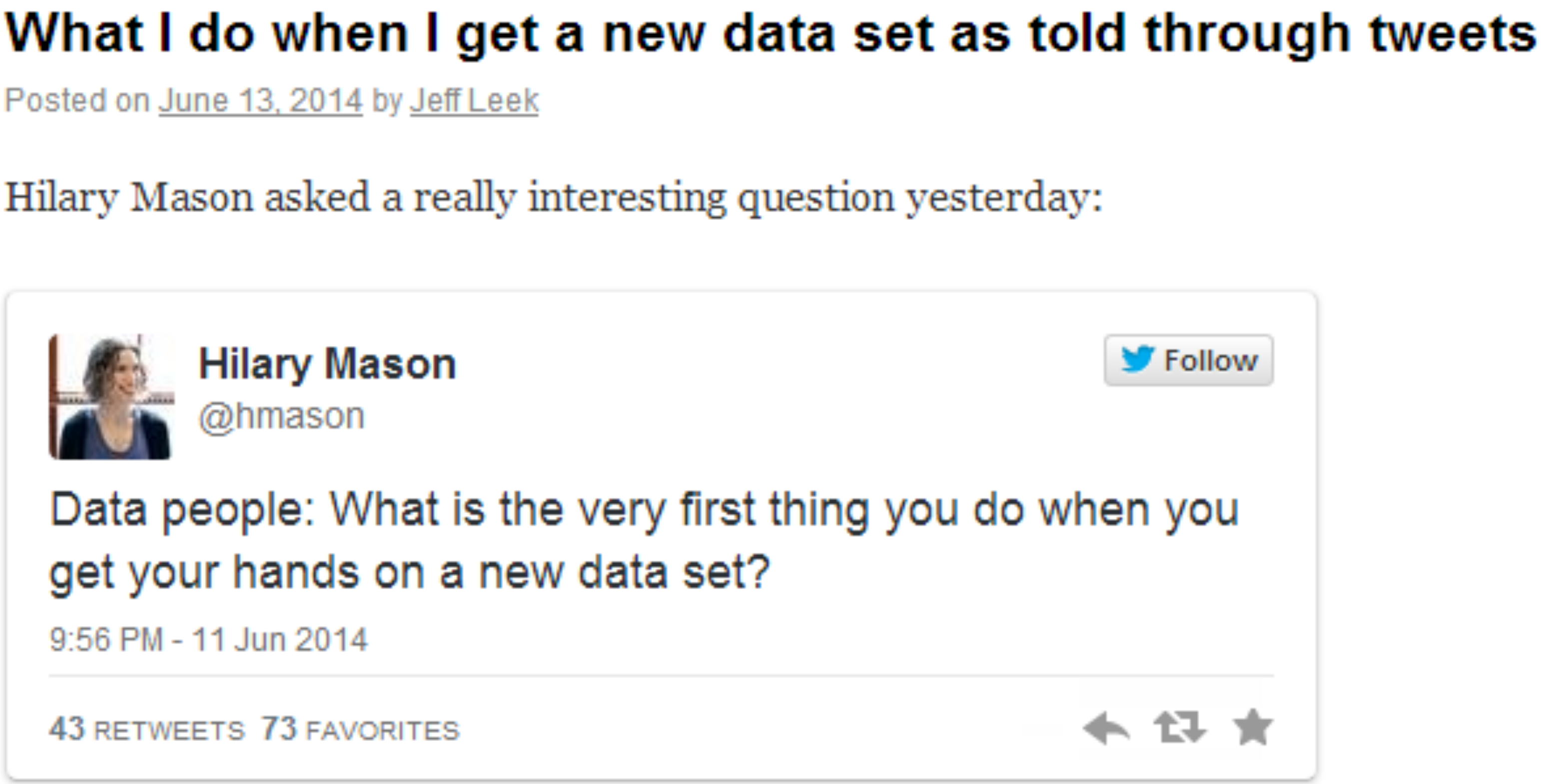
- EDA should take place before more rigorous statistical analysis. 1st part of larger process

- First line of defense against bad data, use to check assumptions about data.

- Maybe lead to new insights, new questions or feed into process of building predictive models

What do with a new dataset? – Jeef Leek

Blog: http://simplystatistics.org/



What I do when I get a new data set as told through tweets

Posted on June 13, 2014 by Jeff Leek

Hilary Mason asked a really interesting question yesterday:

Hilary Mason
@hmason

Data people: What is the very first thing you do when you get your hands on a new data set?

9:56 PM - 11 Jun 2014

43 RETWEETS 73 FAVORITES

What do with a new dataset? – Jeff Leek

- **Step 0: Figure out what I'm trying to do with the data**
  - "Look, Stop, Think…",
  - Check-in with person generating data

- **Step 1: Learn about the elephant**
  - figure out what the data set "looks" like: head(), tail, sapply(df, class)
  - look for NA,
  - check for personally identifiable information
  - colnames(), summary(), str() etc

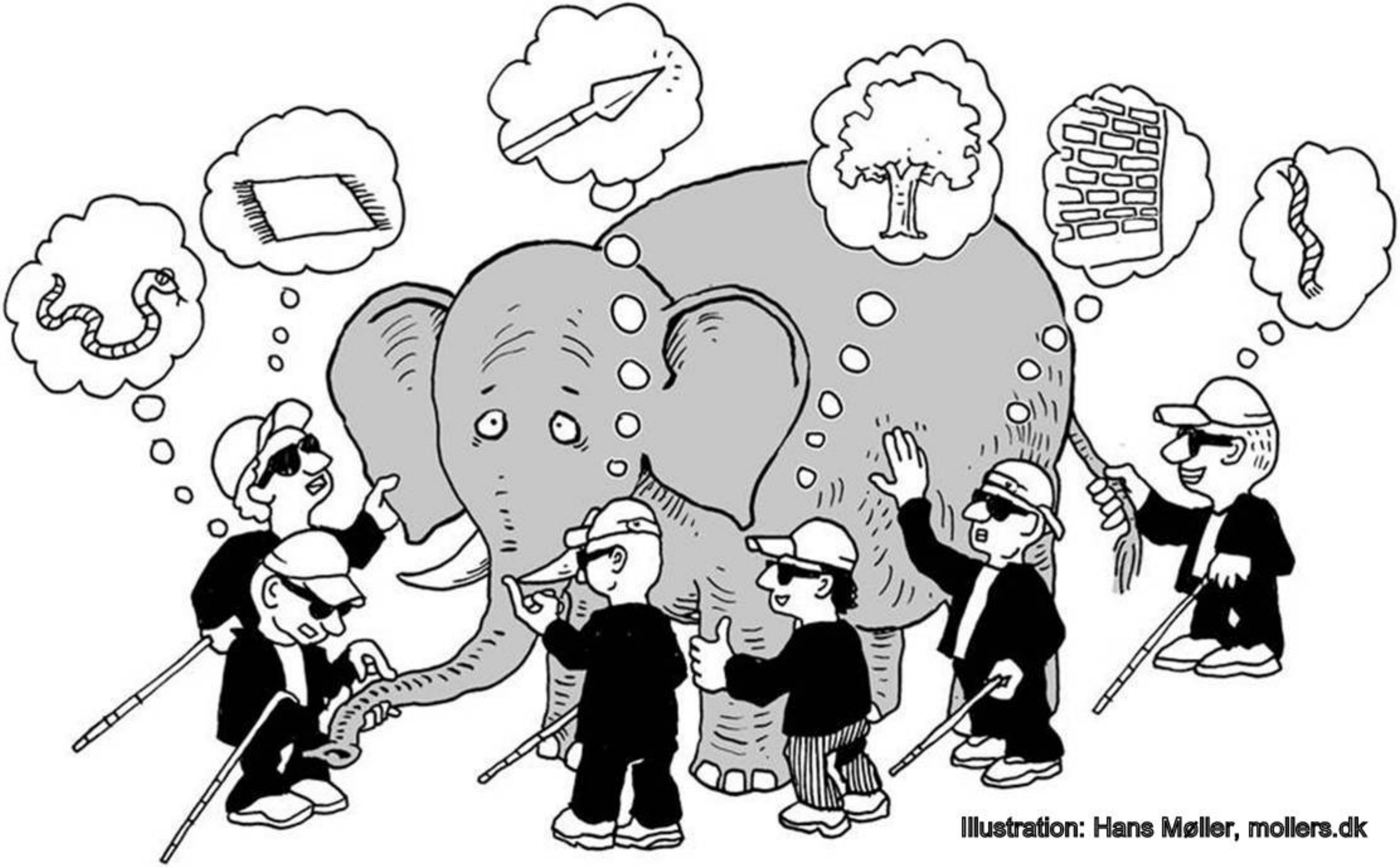


Illustration: Hans Møller, mollers.dk

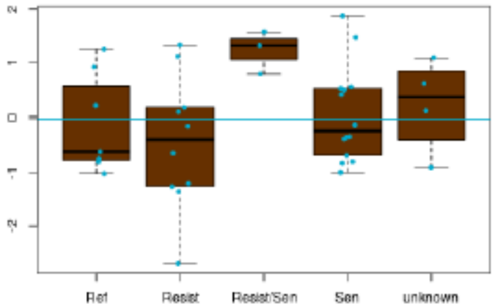
What do with a new dataset? – Jeff Leek

- Step 0: Figure out what I'm trying to do with the data
- Step 1: Learn about the elephant
- **Step 2: Clean/organize**
  - Fix data, NA,
  - swear a lot

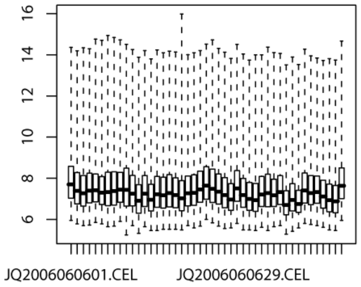
- **Step 3: Plot. That. Stuff**
  - look at variables one by one like
    - Histograms,
    - scatterplots,
    - density plots,
    - jittered 1d plots
  - If data are multivariate ,get a feel for high dimensional structure
    - dimension reduction : principal components,
    - hierarchical clustering analysis

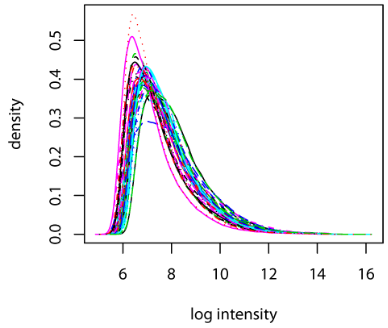
Visualize variables one by one

- Histograms, density plots
- Boxplots
- Scatterplots
- MA plots (variance v mean)
- Pairs
- Correlations, Corrplot



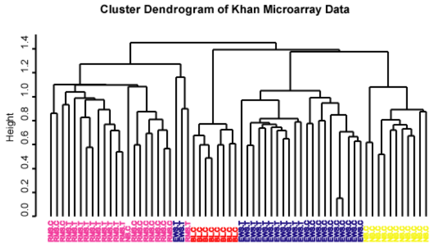
distEisen(exprs(dataset))
hclust (*, "average")



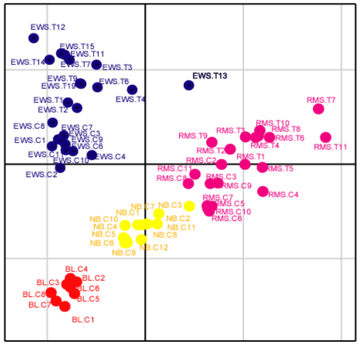


Get a feel for high dimensional structure of multivariate data

- Clustering
  - Hierarchical
  - Flat (k-means)



- Matrix Factorization/ Dimension Reduction
  - Principal Component analysis, Correspondence analysis
  - etc

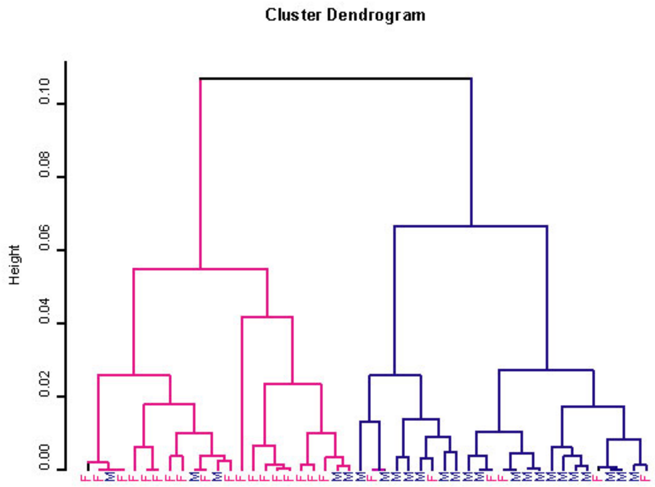


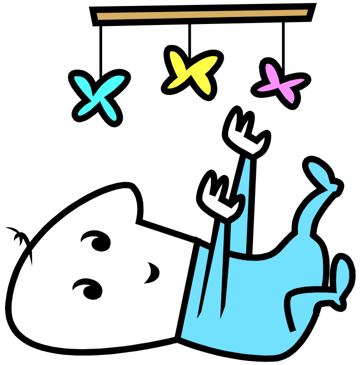
What do with a new dataset? – Jeff Leek

- Step 0: Figure out what I'm trying to do with the data
- Step 1: Learn about the elephant
- **Step 2: Clean/organize**
- **Step 3: Plot. That. Stuff**
- **Step 4: Get a quick answer to the question from Step 1**
  - quick and dirty answer to the question;
  - simple predictive model or a really basic regression model.
  - Check back with person generating data.

Limitations of hierarchical clustering

- Samples compared in a pair wise manner
- Hierarchy forced on data
- Sometimes difficult to visualise if large data
- Overlapping clustering or time/dose gradients ?





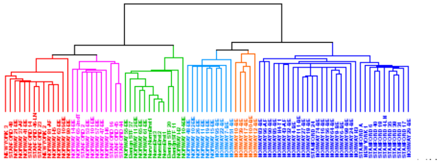
Matrix Factorization

- Also refers to as
  - Latent variable analysis, Dimension reduction, Ordination

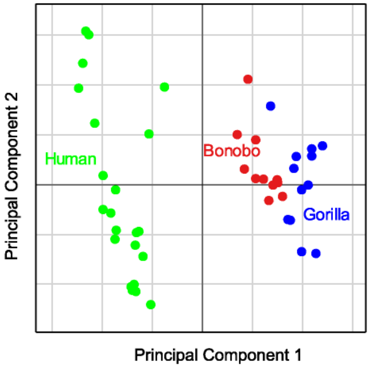
- Aim:

  Find axes onto which data can be project so as to explain as much of the variance in the data as possible

Complementary methods



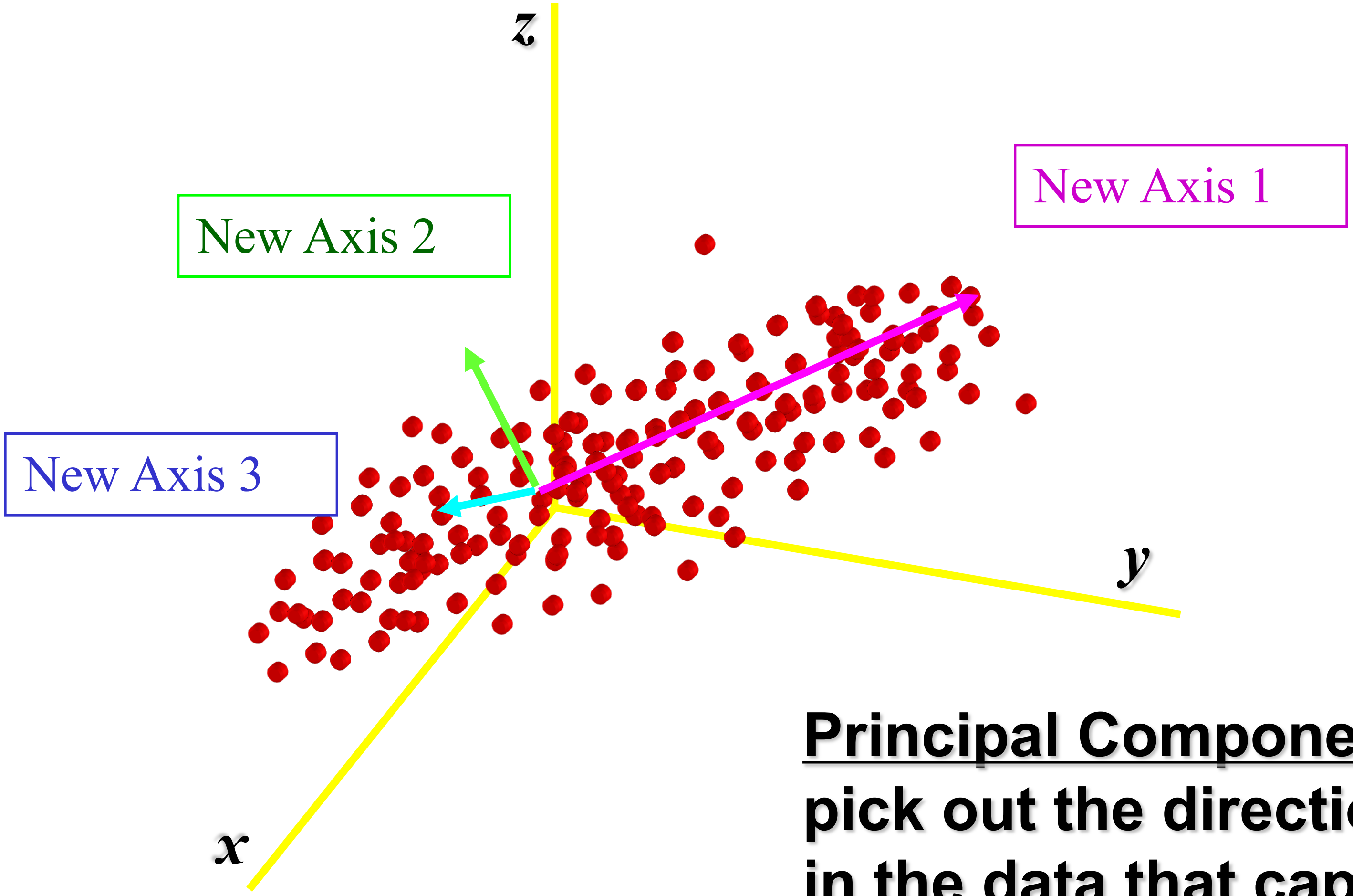
Cluster analysis generally investigates pairwise distances/similarities among objects looking for fine relationships



Ordination in reduced space considers the variance of the whole dataset thus highlighting general gradients/patterns

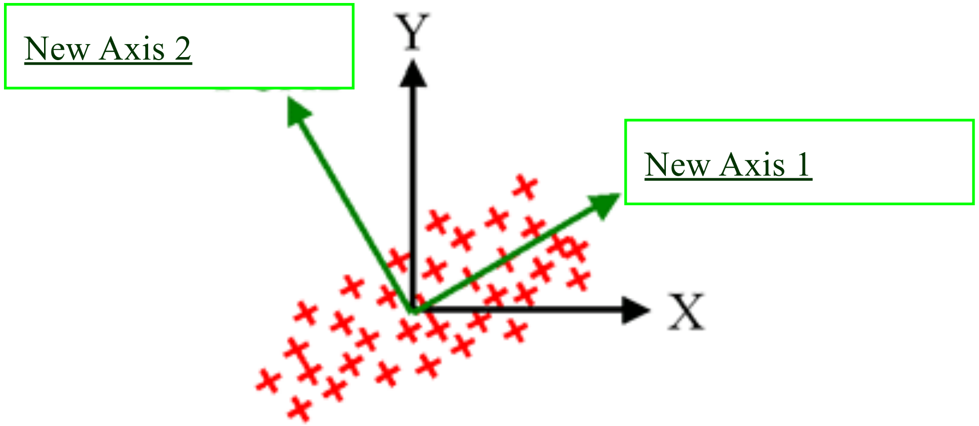
(Legendre and Legendre, 1998)

Dimension Reduction (Ordination)



Principal Components pick out the directions in the data that capture the greatest variability

Representing data in a reduced space


New Axis 2

New Axis 1

The first new axes will be projected through the data so as to explain the greatest proportion of the variance in the data.

The second new axis will be orthogonal, and will explain the next largest amount of variance

Interpreting an Ordination

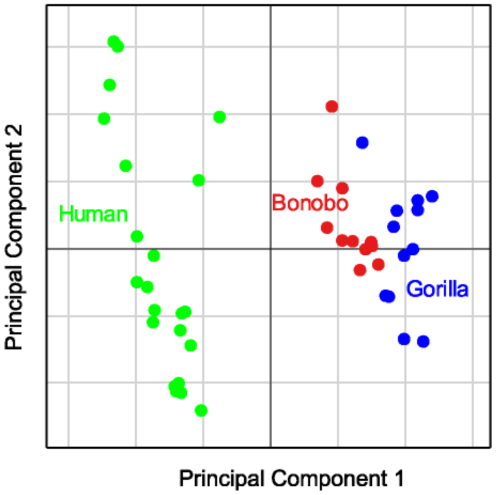
Each axes represent a different "trend" or set of profiles

The further from the origin

Greater loading/contribution

(ie higher expression)

Same direction from the origin



Principal Axes

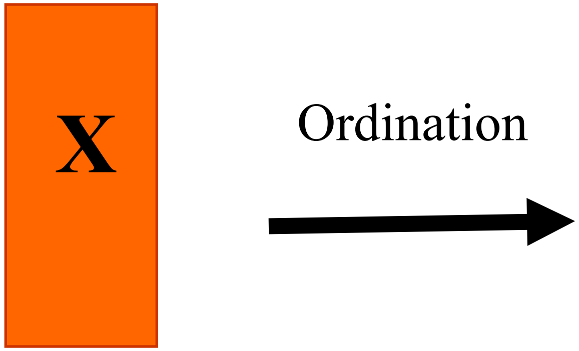
- Project new axes through data which capture variance. Each represents a different trend in the data.

- Orthogonal (decorrelated)

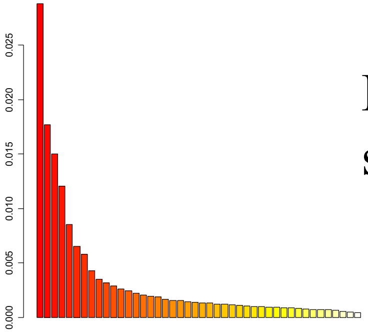
- Typically ranked: First axes most important

- Principal axis, Principal component, latent variable or eigenvector

Typical Analysis


X

Ordination



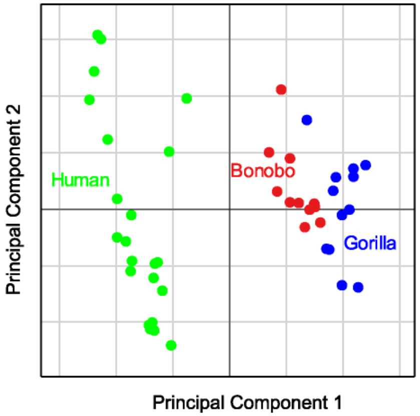
Plot of eigenvalues,
select number.

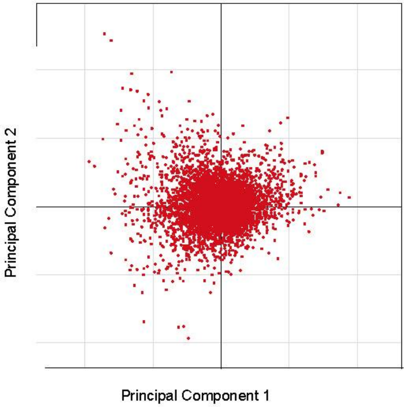
Array Projection

Gene Projection

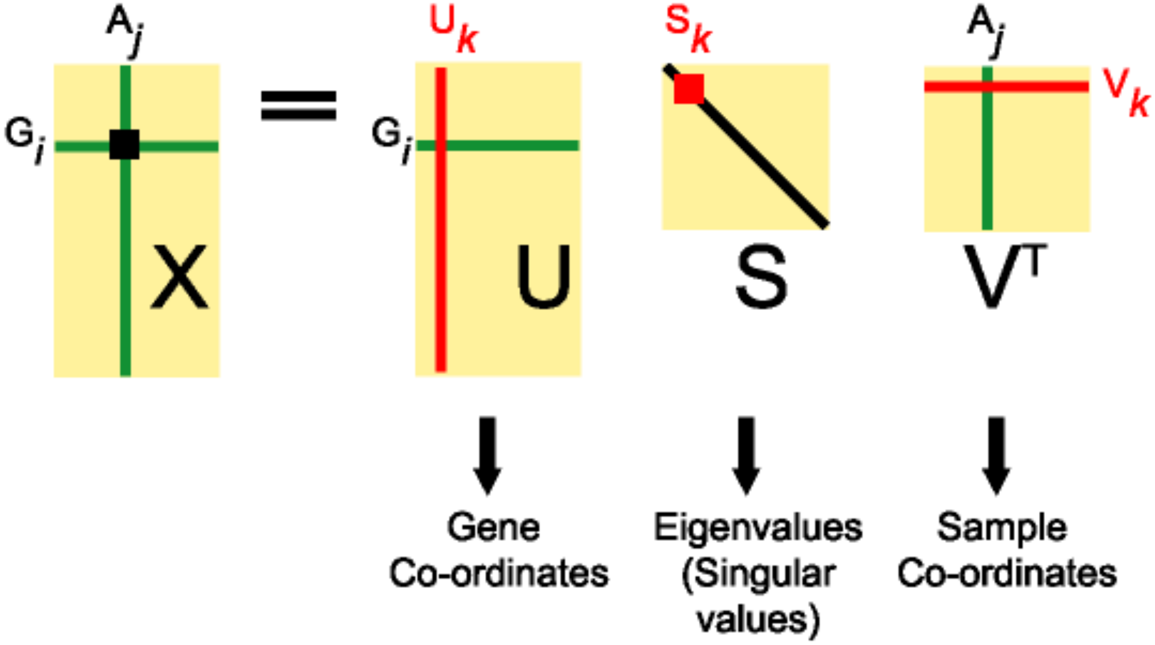
Plot PC1 v PC2

etc





Singular Value Decomposition $X=USV^T$



Gene Co-ordinates

Eigenvalues (Singular values)

Sample Co-ordinates

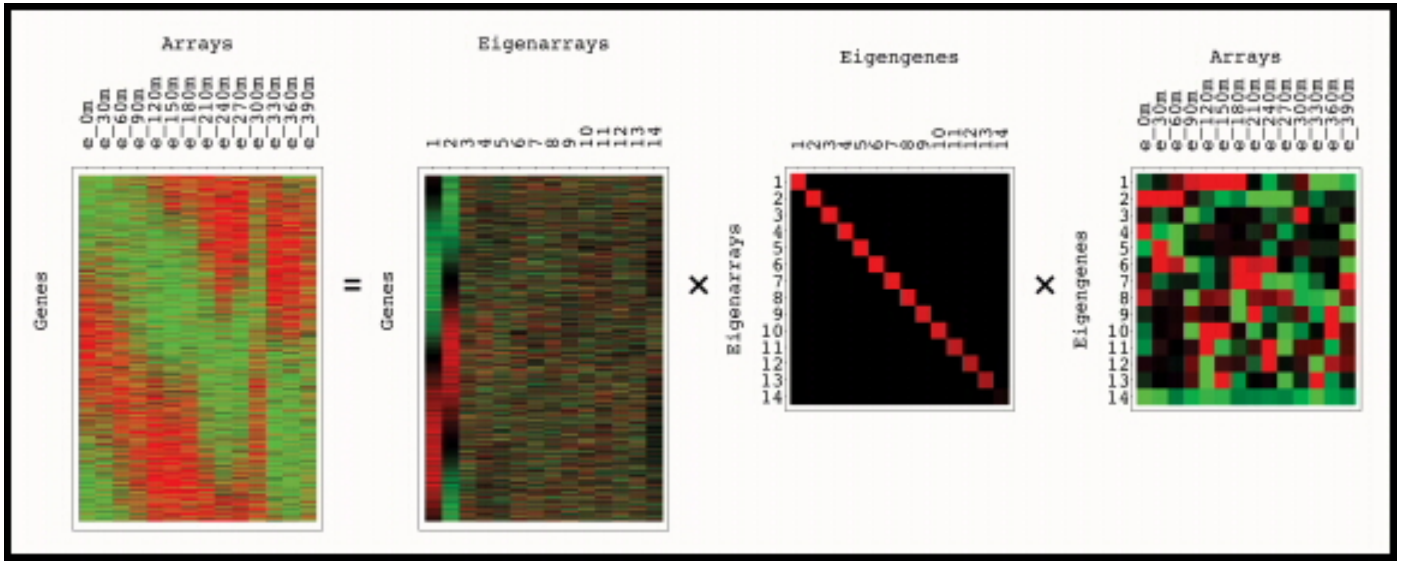


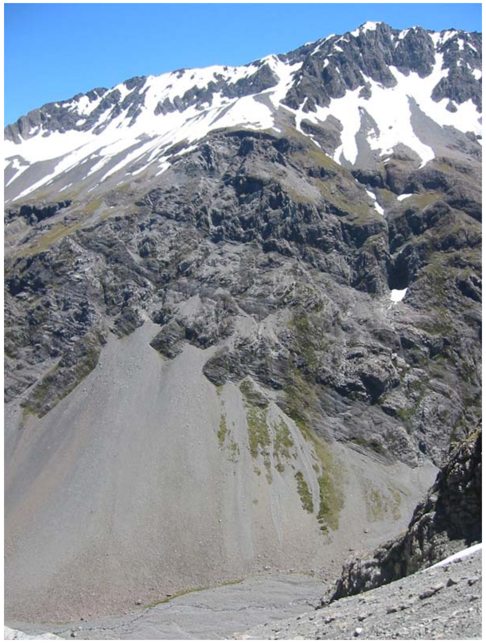
Image from

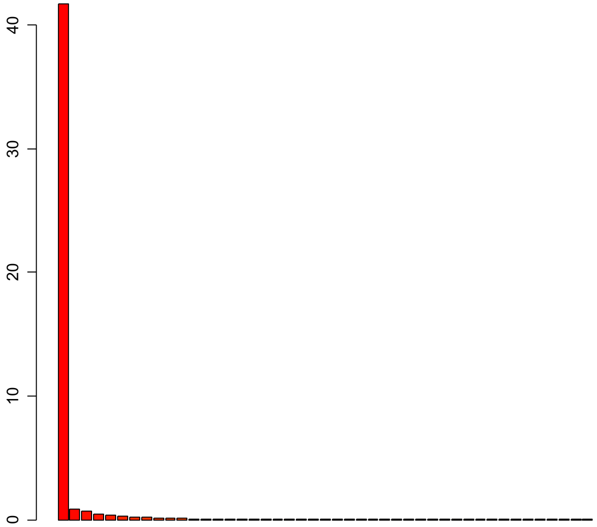
http://genome-www.stanford.edu/SVD/

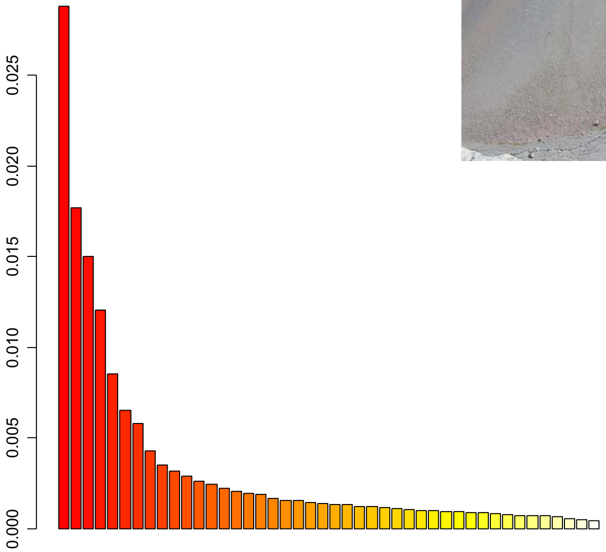
Eigenvalues

- Describe the amount of variance (information) in eigenvectors
- Ranked. First eigenvalue is the largest.
- Generally only examine 1st few components
  - scree plot

Choosing number of Eigenvalues: Scree Plot







Maximum number of Eigenvalues/Eigenvectors = min(nrow, ncol) -1

Ordination Methods

- Most common :
  - Principal component analysis (PCA)
  - Correspondence analysis (COA or CA)
  - Nonmetric multidimensional scaling (NMDS, MDS)
  - Principal co-ordinate analysis (PCoA)

Relationship

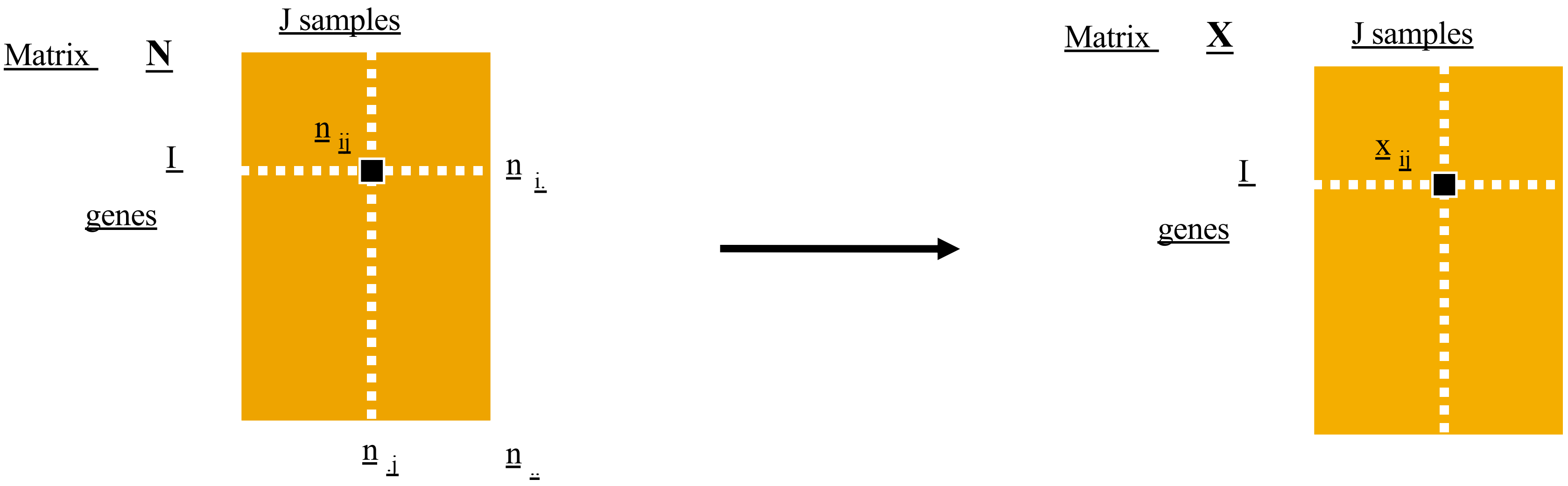
- PCA, COA, etc can be computed using Singular value decomposition (SVD)
- SVD applied to microarray data (Alter et al., 2000)
- Wall et al., 2003 described both SVD, PCA (good paper)

Principal Component Analysis (PCA)

- Probably most popular ordination method
- Eigenanalysis of a covariance matrix (most common) or correlation matrix
- Dates back to Pearson (1901)
- Applied to quantitative data.
- First applied to microarray data by Raychadhuri et al., 2000.
- PCA: prcomp(stats), princomp(stats) dudi.pca (ade4).

PCA: Initial data transformation

Column mean centred
(covariance PCA)



Matrix **N**

J samples

I

genes

n_{ij}

$n_{i.}$

$n_{.j}$

$n_{..}$

Matrix **X**

J samples

I

genes

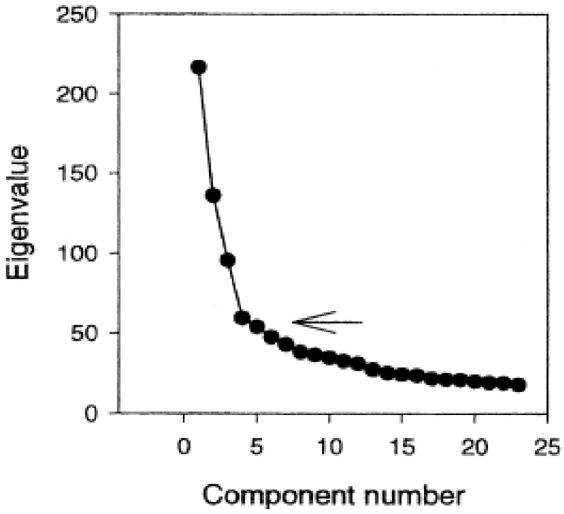
x_{ij}

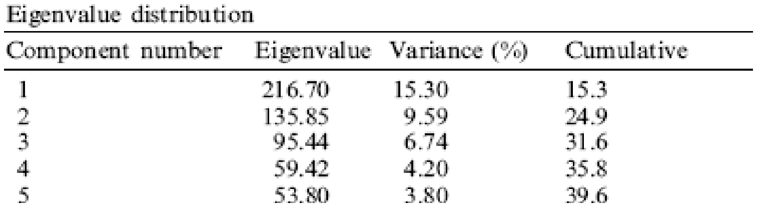
$$x_{ij} = n_{ij} - m_{j}$$

Where *mj* is the mean of column J

PCA of the NCI 60 cell lines

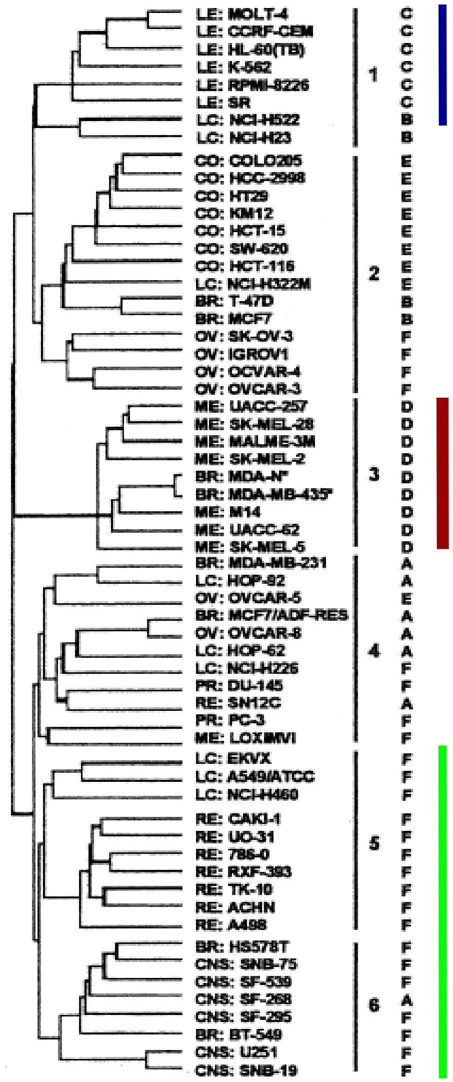
- Crescenzi and Giuliani, 2001
- Gene expression profiles of 60 cancer cell lines representing diverse cancers
- Performed PCA
- Found 5 PC's to contain most of the variance. Of which first 3 most interesting

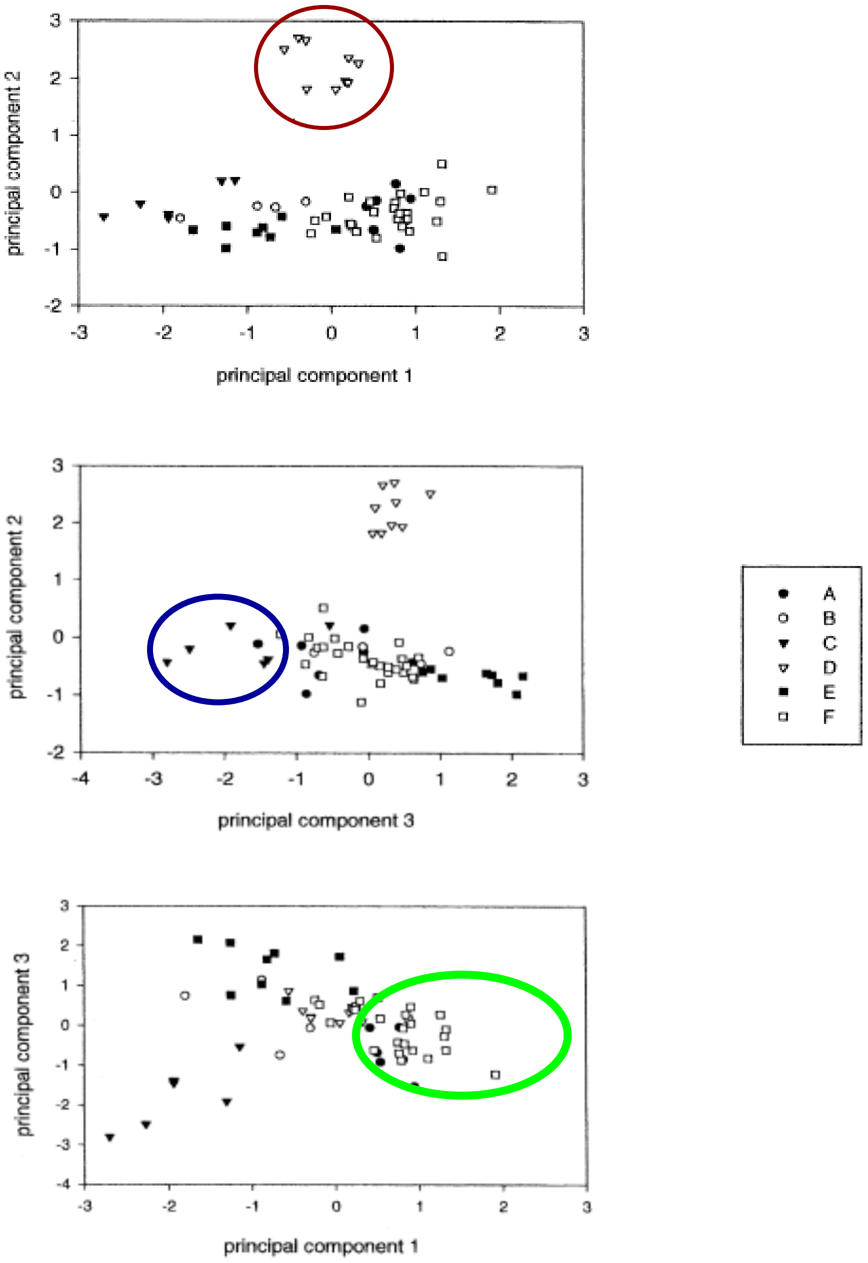




Eigenvalue distribution

Component number	Eigenvalue	Variance (%)	Cumulative
1	216.70	15.30	15.3
2	135.85	9.59	24.9
3	95.44	6.74	31.6
4	59.42	4.20	35.8
5	53.80	3.80	39.6


LE: MOLT-4 C
LE: CCRF-CEM C
LE: HL-60(TB) C
LE: K-562 C
LE: RPMI-8226 C
LE: SR C
LC: NCI-H522 B
LC: NCI-H23 B
1
CO: COLO205 E
CO: HCC-2998 E
CO: HT29 E
CO: KM12 E
CO: HCT-15 E
CO: SW-620 E
CO: HCT-116 E
LC: NCI-H322M E
BR: T-47D B
BR: MCF7 B
OV: SK-OV-3 F
OV: IGROV1 F
OV: OCVAR-4 F
OV: OVCAR-3 F
2
ME: UACC-257 D
ME: SK-MEL-28 D
ME: MALME-3M D
ME: SK-MEL-2 D
BR: MDA-N* D
BR: MDA-MB-435* D
ME: M14 D
ME: UACC-62 D
ME: SK-MEL-5 D
3
BR: MDA-MB-231 A
LC: HOP-92 A
OV: OVCAR-5 E
BR: MCF7/ADF-RES A
OV: OVCAR-8 A
LC: HOP-62 A
LC: NCI-H226 F
PR: DU-145 F
RE: SN12C A
PR: PC-3 F
ME: LOXIMVI F
4
LC: EKVX F
LC: A549/ATCC F
LC: NCI-H460 F
RE: CAKI-1 F
RE: UO-31 F
RE: 786-0 F
RE: RXF-393 F
RE: TK-10 F
RE: ACHN F
RE: A498 F
5
BR: HS578T F
CNS: SNB-75 F
CNS: SF-539 F
CNS: SF-268 A
CNS: SF-295 F
BR: BT-549 F
CNS: U251 F
CNS: SNB-19 F
6


principal component 2
3
2
1
0
-1
-2
-3
-2
-1
0
1
2
3
principal component 1
principal component 2
3
2
1
0
-1
-2
-4
-3
-2
-1
0
1
2
3
principal component 3
principal component 3
3
2
1
0
-1
-2
-3
-4
-3
-2
-1
0
1
2
3
principal component 1
● A
○ B
▼ C
▽ D
■ E
□ F

Considerations when applying PCA

- Distance – Euclidean
- Robust, but designed for analysis of multi-normal distributed data
  - if very skewed data, the first few axes will only separate a few objects with extreme values instead of displaying main axes of variation
- Generally for microarray analysis: Row centre.
  - Eliminate size effect. Low abundance genes can be detected
- Problems which Correspondence analysis handles better
  - If lots zero
  - Unimodal or non-linear trends. Get distortion or artifact in plot, in which the second axis is an arched function of the first axis. Called horseshoe effect in PCA.

If I'd known they wanted me to use all this info - I would never have asked for it!
Julie ®

Correspondence Analysis

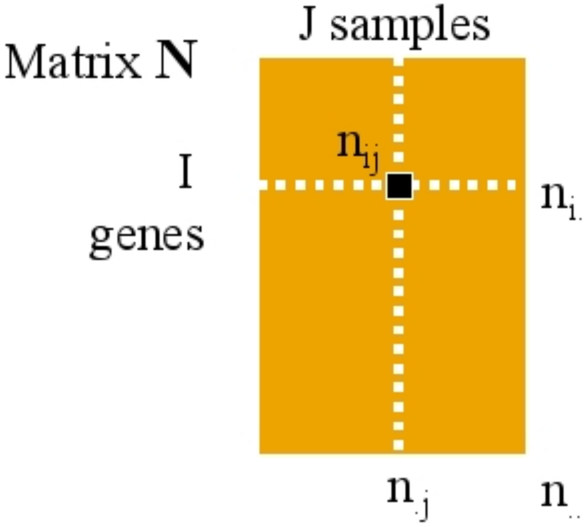
- **COA** (or CA) is an eigenanalysis of a Chi-square distance matrix.

- Measures the “strength” of association between an up-regulated gene and an array sample.

- Developed by numerous authors, also known as reciprocal averaging/ordering, dual scaling etc.

- Initially designed for analysis of 2-way contingency tables (frequency counts). Thus assumes matrix counts positive integers or zeros.

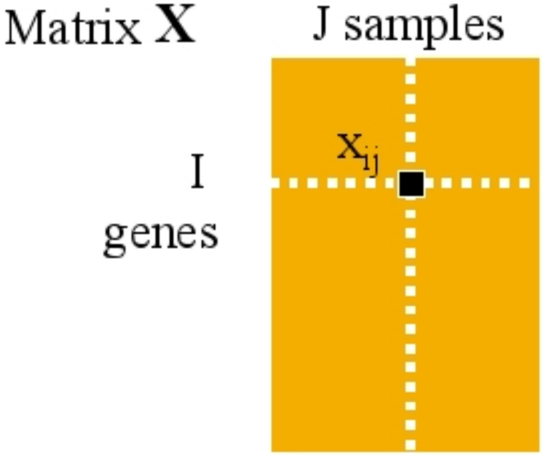
COA: Initial Transformation



$$c_j = n_{.j}/n_{..}$$

$$r_i = n_{i.}/n_{..}$$

$$p_{ij} = n_{ij}/n_{..}$$



$$x_{ij} = (p_{ij} - r_i c_j)/ \sqrt{\ } \ r_i c_j \ )$$

Pearson chi-square statistic O_{ij}-E_{ij} / $\sqrt{E_{ij}}$

COA of the Yeast Cell-Cycle Data

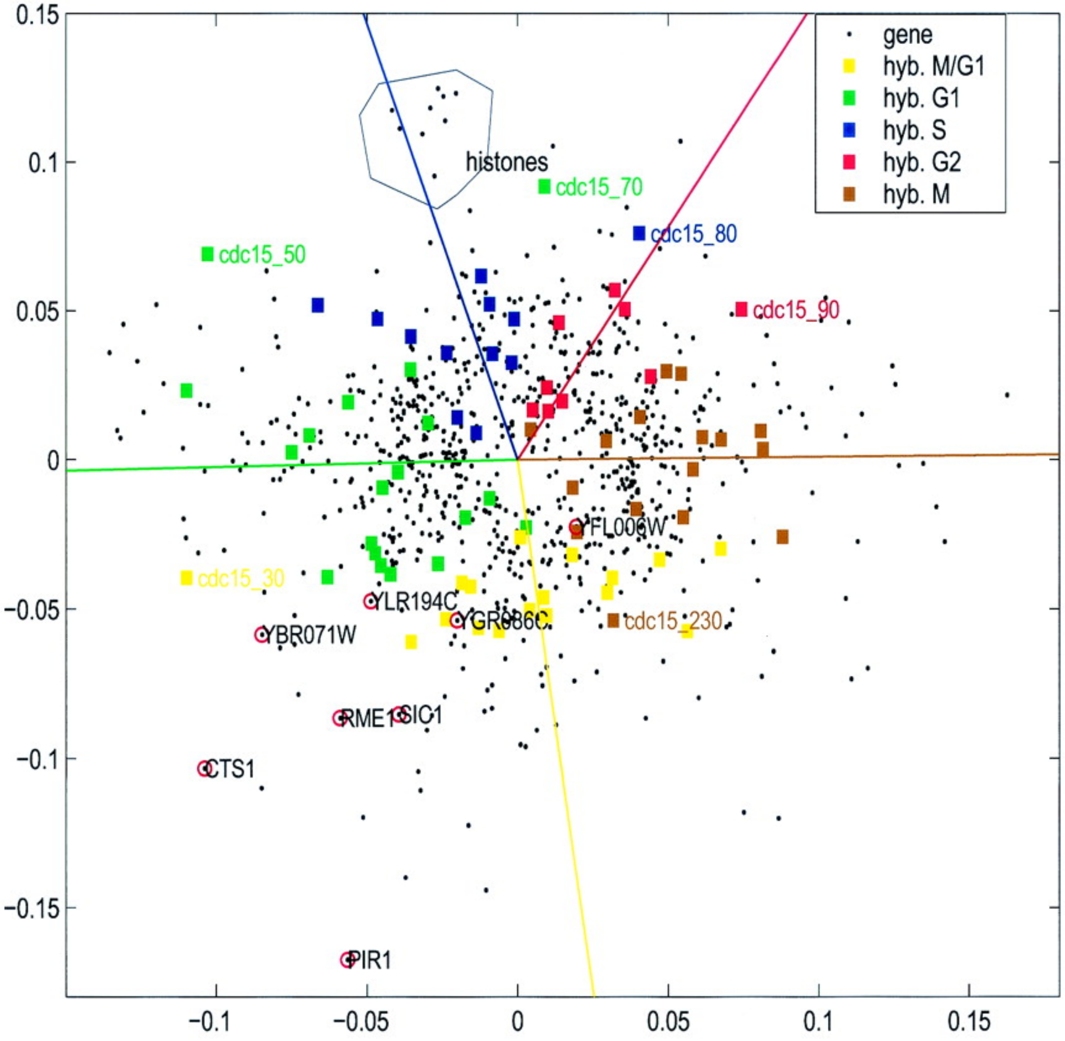
Fellenberg et al., 2001

- Dataset:
  - Gene Expression of *S. cerevisiae* arrested during cell cycle by 4 methods
    - alpha factor-, *CDC15*-, *CDC28*-based blocking and elutriation. (Spellman et al., 1998)
- COA
- Visualised using biplot of genes and arrays

Correspondence Analysis

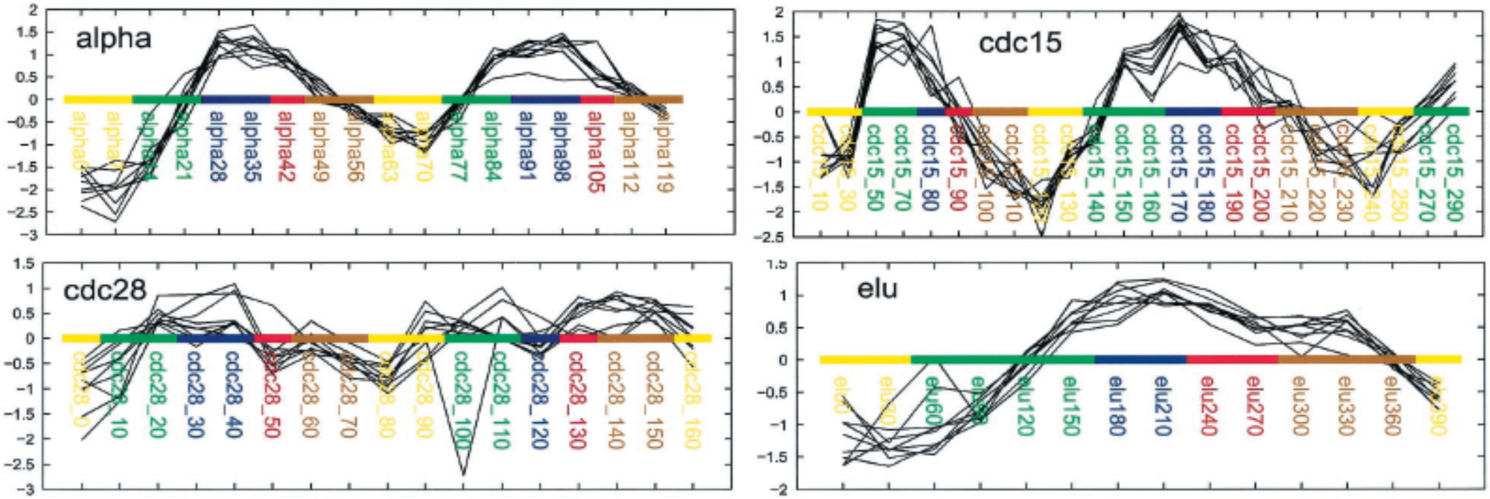
Biplot

Arrays and genes with a strong association (correspondence) are projected in the same direction from the origin.



Fellenberg et al., (2001) PNAS 98(19):10781-10786

Gene Expression of 9 histone genes



See gene expression of histones ↑ during ■ S phase

Consideration when applying COA

- Data must be in same units (so they can be added)
- Data must be non-negative or made position by translation (scalar addition)
- In case of steep gradients (many zero) COA should produce better results than PCA
- Data is dual (column and row) scaled.
- Unimodal or non-linear trends may be represented as arch (2nd axis). Less serious than PCA's horseshoe effect.

Among the other related methods

Independent Component Analysis

- generalization of PCA
- does not constrain the axes to be orthogonal
- attempts to place them in the directions of statistical dependencies in the data.
- Lee & Batzoglou 2003 and Saidi et al., 2004 show ICA outperformed PCA

Spectral map analysis

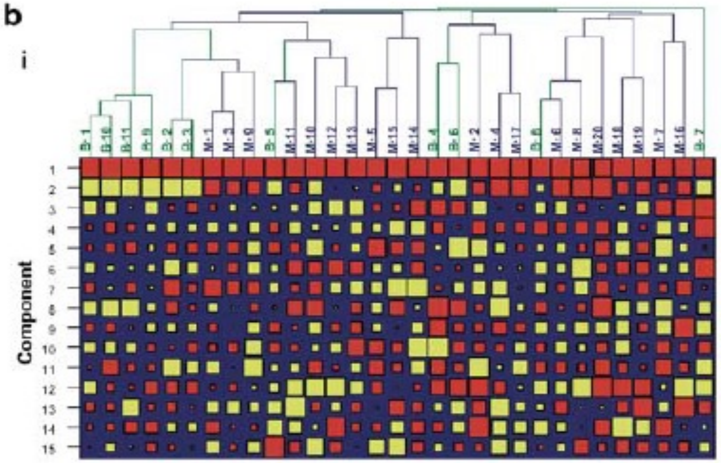
- related to COA (dual scaling of both rows + columns)
- not limited to contingency tables and cross-tabulations. possibility to use other weighting factors
- Wouters et al., 2003 showed SMA outperformed PCA, comparable to COA.

MDS

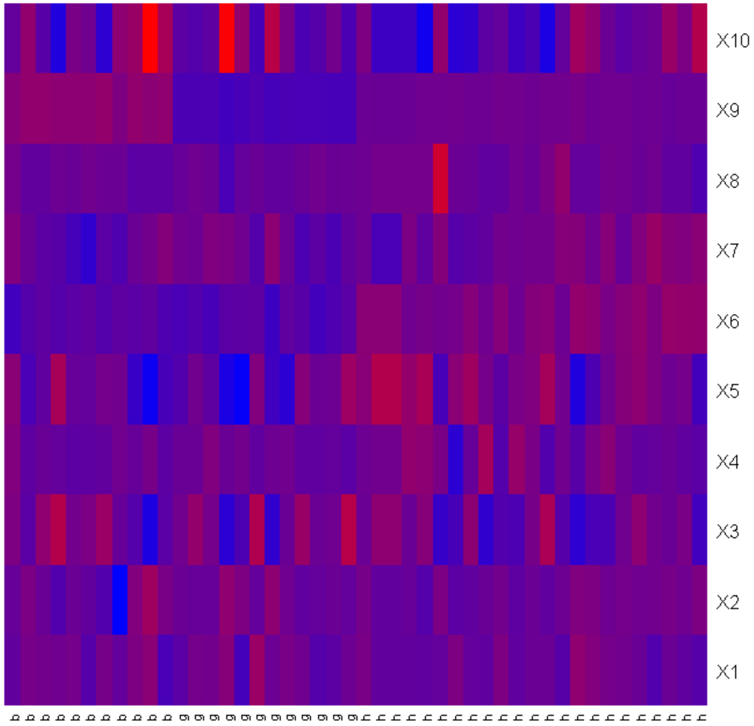
- Multidimensional scaling
  - metric and non-metric.
- Input distance matrix
- Classical MDS is identical to principal coordinates analysis (PCoA). R code cmdscale (stats), dudi.pco (ade4)
- NMDS. Iterative. isoMDS (MASS), sammon (MASS).

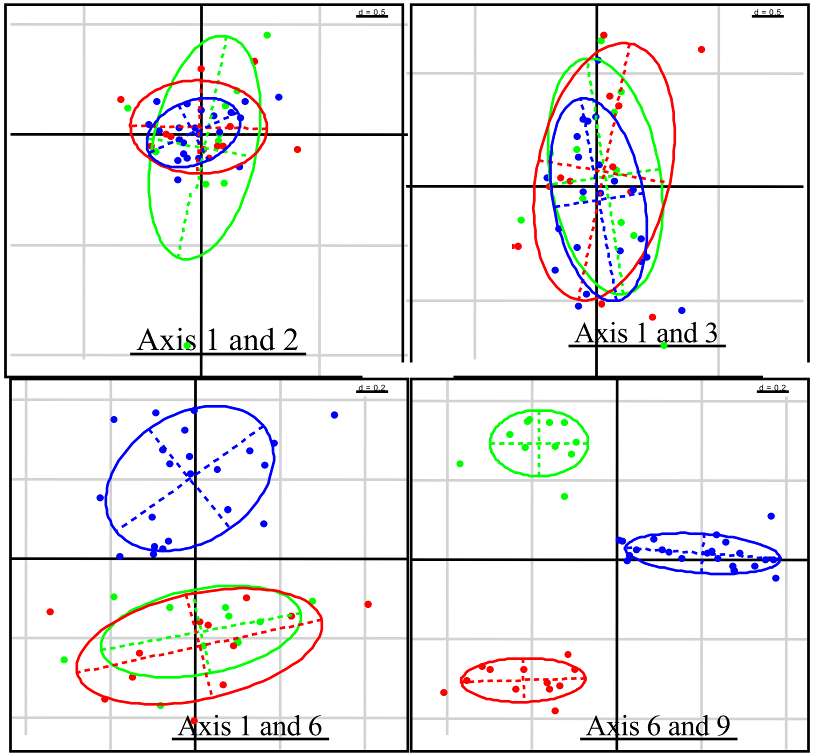
Independent Component Analysis Saudi et al., 2003

- Axes/eigenvectors are not constrained to be orthogonal
- May detect more subtle patterns in data
- More complex interpretation of axes

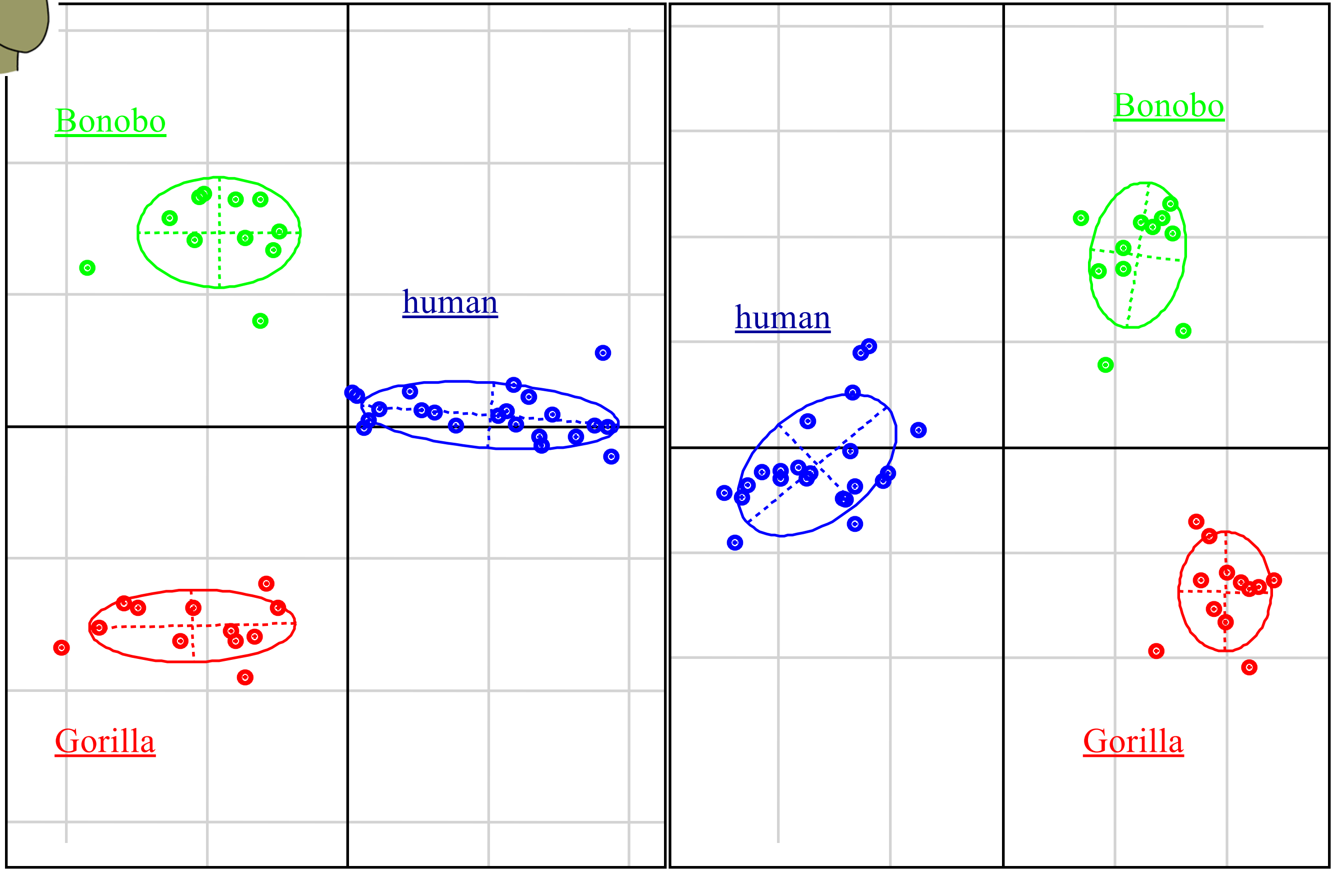


Fast ICA of Karaman Ape Data





Results from fastICA and COA



fastICA, Axis 6 and 9

COA, Axis 1 and 3

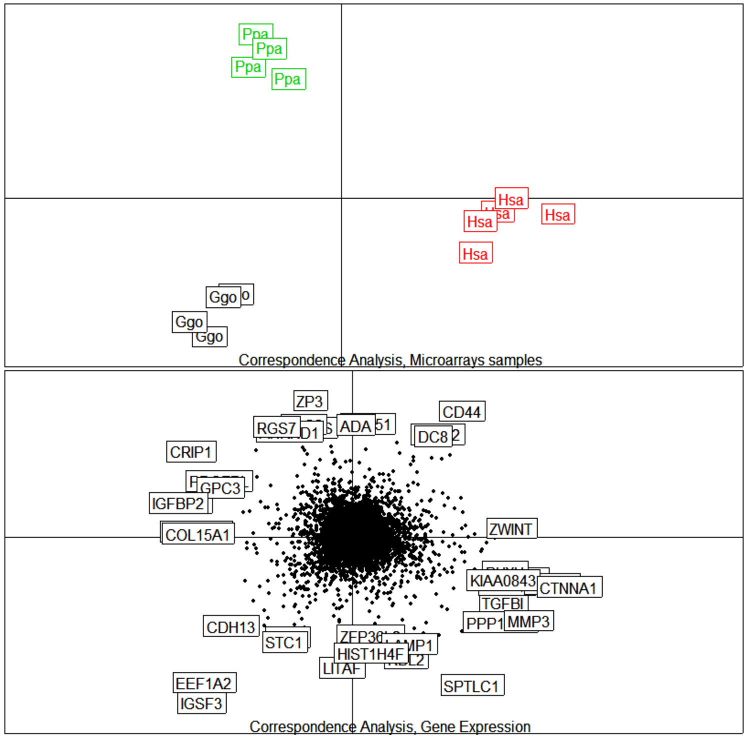
Summary: Exploration analysis using Ordination

- SVD = straightforward dimension reduction
- PCA = column mean centred +SVD
  - Euclidean distance
- COA = Chi-square +SVD
  - produces nice biplot

- Ordination be useful for visualising trends in data
- Useful complementary methods to clustering

Ordination in R

Ordination (PCA, COA)



- library(ade4)
- dudi.pca()
- dudi.coa()

- library(made4)
- ord(data, type=“pca”)
- plot()
- plotarrays()
- plotgenes()

Link to example 3d html file

Books/Book Chapters:

1. Legendre, P., and Legendre, L. 1998. *Numerical Ecology*, 2nd English Edition. ed. Elsevier, Amsterdam.
2. Wall, M., Rechtsteiner, A., and Rocha, L. 2003. Singular value decomposition and principal component analysis. In *A Practical Approach to Microarray Data Analysis*. (eds. D.P. Berrar, W. Dubitzky, and M. Granzow), pp. 91-109. Kluwer, Norwell, MA.

Papers:

1. Pearson, K. 1901. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* **2:** 559-572.
2. Hotelling, H., 1933. Analysis of a complex statistical variables into principal components. J. Educ. Psychol. 24, 417-441. Alter, O., Brown, P.O., and Botstein, D. 2000. Singular value decomposition for genome-wide expression data processing and modeling. *Proc Natl Acad Sci U S A* **97:** 10101-10106.
3. Culhane, A.C., Perriere, G., Considine, E.C., Cotter, T.G., and Higgins, D.G. 2002. Between-group analysis of microarray data. *Bioinformatics* **18:** 1600-1608.
4. Culhane, A.C., Perriere, G., and Higgins, D.G. 2003. Cross-platform comparison and visualisation of gene expression data using co-inertia analysis. *BMC Bioinformatics* **4:** 59.
5. Fellenberg, K., Hauser, N.C., Brors, B., Neutzner, A., Hoheisel, J.D., and Vingron, M. 2001. Correspondence analysis applied to microarray data. *Proc Natl Acad Sci U S A* **98**: 10781-10786.
6. Raychaudhuri, S., Stuart, J.M., and Altman, R.B. 2000. Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pac Symp Biocomput*: 455-466.
7. Wouters, L., Gohlmann, H.W., Bijnens, L., Kass, S.U., Molenberghs, G., and Lewi, P.J. 2003. Graphical exploration of gene expression data: a comparative study of three multivariate methods. *Biometrics* **59**: 1131-1139

Reviews

1. Quackenbush, J. 2001. Computational analysis of microarray data. *Nat Rev Genet* **2:** 418-427.
2. Brazma A., and Culhane AC. (2005) Algorithms for gene expression analysis. In *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics.* Dunn MJ., Jorde LB., Little PFR, Subramaniam S. (eds) John Wiley & Sons. London (download from http://www.hsph.harvard.edu/research/aedin-culhane/publications/)

Interesting Commentary

Terry Speed’s commentary on PCA download from http://bulletin.imstat.org/pdf/37/3