Multivariate Data Analysis in Omics Research

Diverging Alternative Splicing Fingerprints Identified in Thoracic Aortic Aneurysm

Sanela Kjellqvist, PhD

WABI RNAseq course

2015-10-22

Enabler for Life Sciences

Karolinska Institutet
KTH Royal Institute of Technology
Stockholm University
Uppsala Universitet
SciLifeLab

Outline

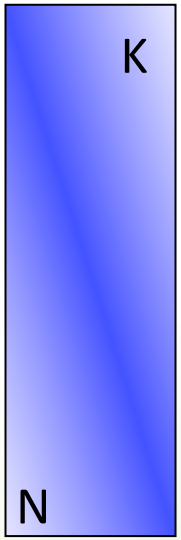
- Why multivariate data analysis?
- Multivariate statistics
  - Different analyses
  - Data preprocessing
- Alternative splicing in thoracic aortic aneurysm
  - Thoracic aortic aneurysm
  - Study setup
  - Aim of the study
  - Results
  - Summary
- Today's exercise

WHY MULTIVARIATE DATA ANALYSIS?

Development of Classical Statistics – 1930s

- Multiple regression
- Canonical correlation
- Linear discriminant analysis
- Analysis of variance

Tables are long and lean

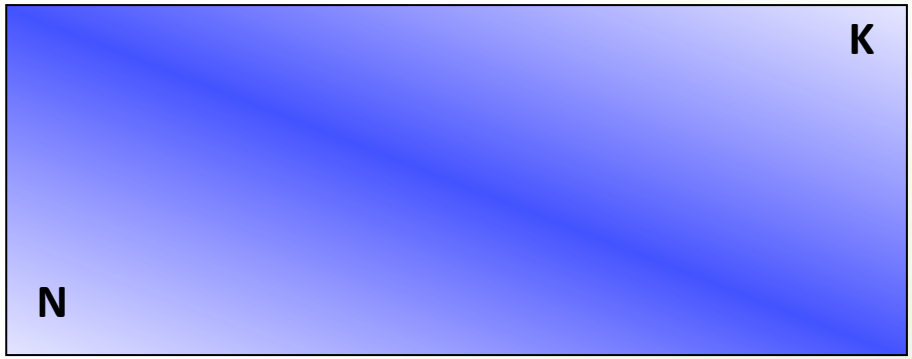


Assumptions:

- Independent X variables
- Many more observations than variables
- Regression analysis one Y at a time
- No missing data

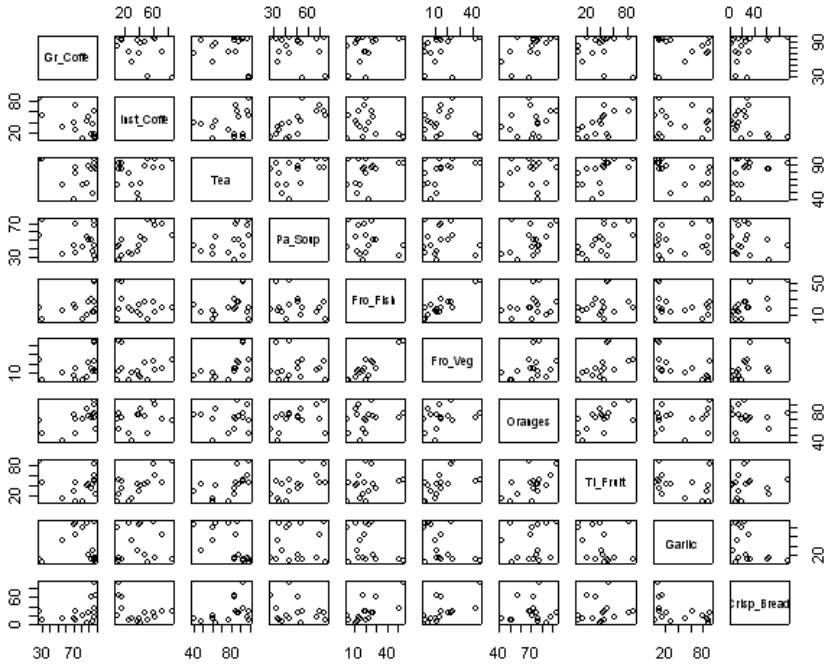
Today's data

- RNASeq, Array, LC-MS/MS, GC/MS or NMR data
- Problems
  - Many variables
  - Few observations
  - Noisy data
  - Missing data
  - Multiple responses
- Implications
  - High degree of correlation
  - Difficult to analyse with conventional methods
- Data ≠ Information
  - Need ways to extract information from the data
  - Need reliable, predictive information
  - Ignore random variation (noise)

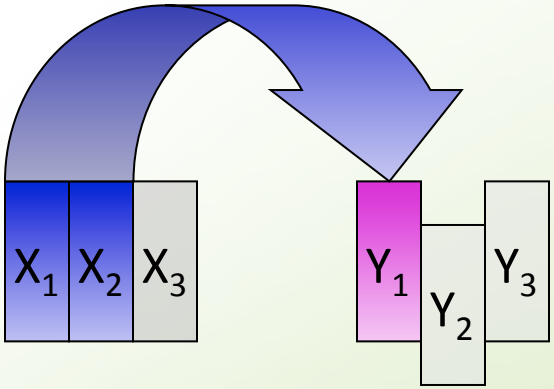


Poor Methods of Data Analysis

- Plot pairs of variables
  - Tedious, impractical
  - Risk of spurious correlations
  - Risk of missing information

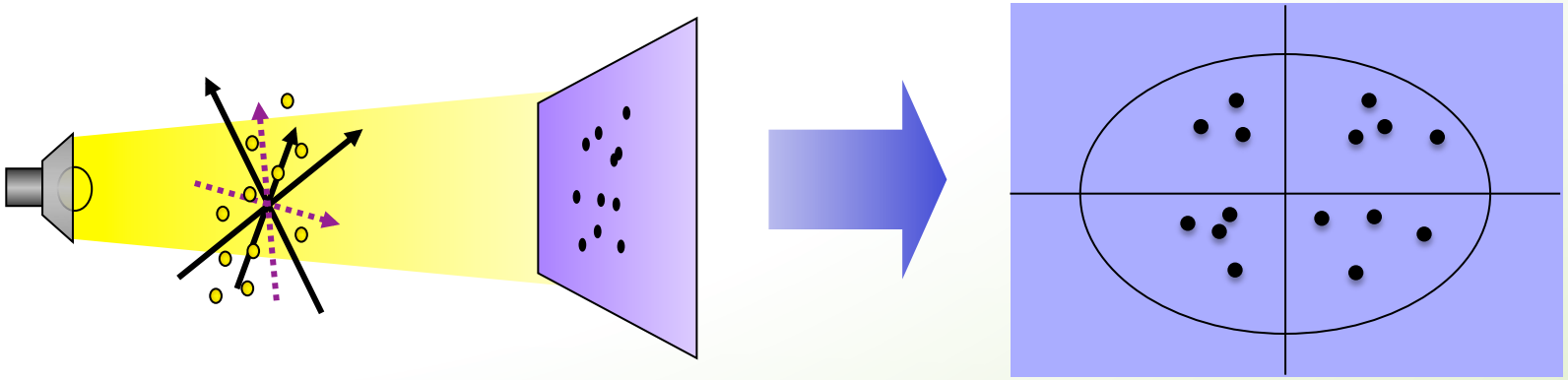


- Select a few variables and use MLR
  - Throwing away information
  - Assumes no 'noise' in X
  - One Y at a time

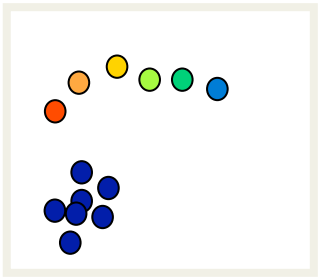


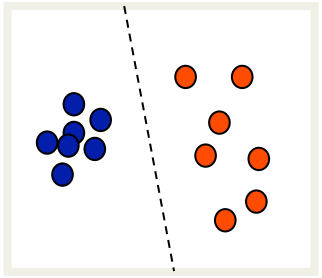
A Better Way…

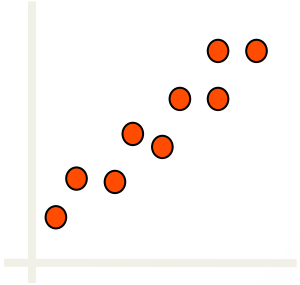
- Multivariate analysis by Projection
  - Looks at ALL the variables together
  - Avoids loss of information
  - Finds underlying trends = “latent variables”
  - More stable models



Fundamental Data Analysis Objectives







Overview	Discrimination	Regression
Trends Outliers Quality Control Biological Diversity Patient Monitoring	Discriminating between groups Biomarker candidates Comparing studies or instrumentation	Comparing blocks of omics data Metab vs Proteomic vs Genomic Omic vs medical Prediction

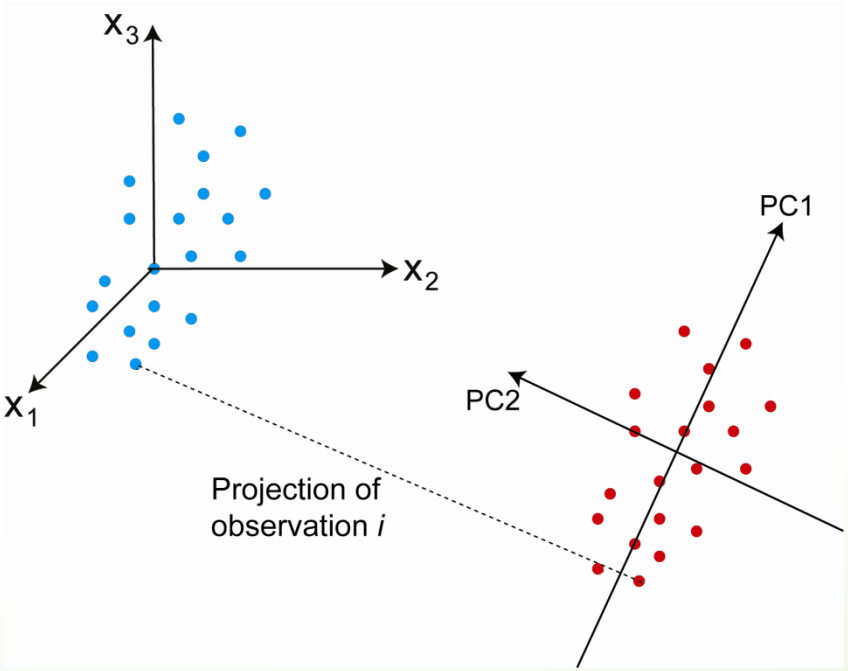
MULTIVARIATE STATISTICS

Different methods

- Principal component analysis (PCA)
- Partial least squares to latent structures analysis (PLS)
- Orthogonal partial least squares to latent structures analysis (OPLS)
- PLS-DA
- OPLS-DA
- K-means clustering
- Hierarchical clustering
- Biplot analysis
- Canonical correlation analysis

What is a projection?

Principal component analysis (PCA)

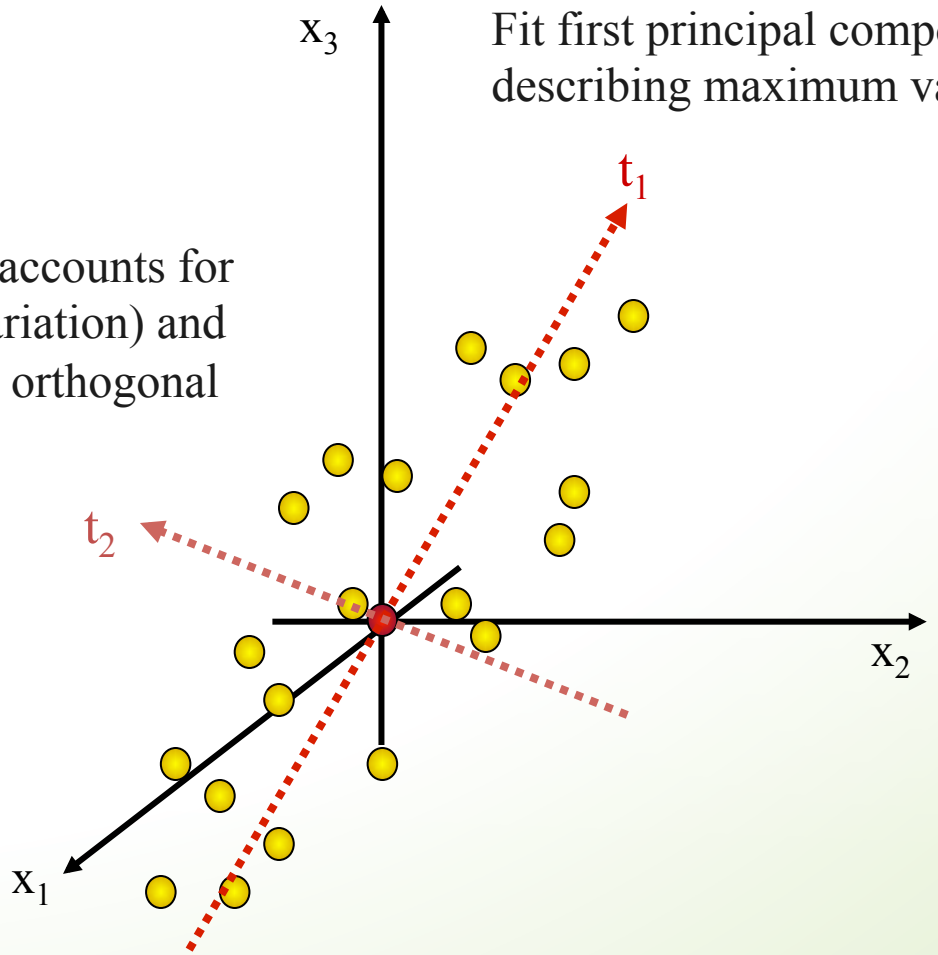


- Algebraically
  - –Summarizes the information in the observations as a few new (latent) variables
- Geometrically
  - – The swarm of points in a K dimensional space (K = number of variables) is approximated by a (hyper)plane and the points are projected on that plane.

PCA - Geometric Interpretation

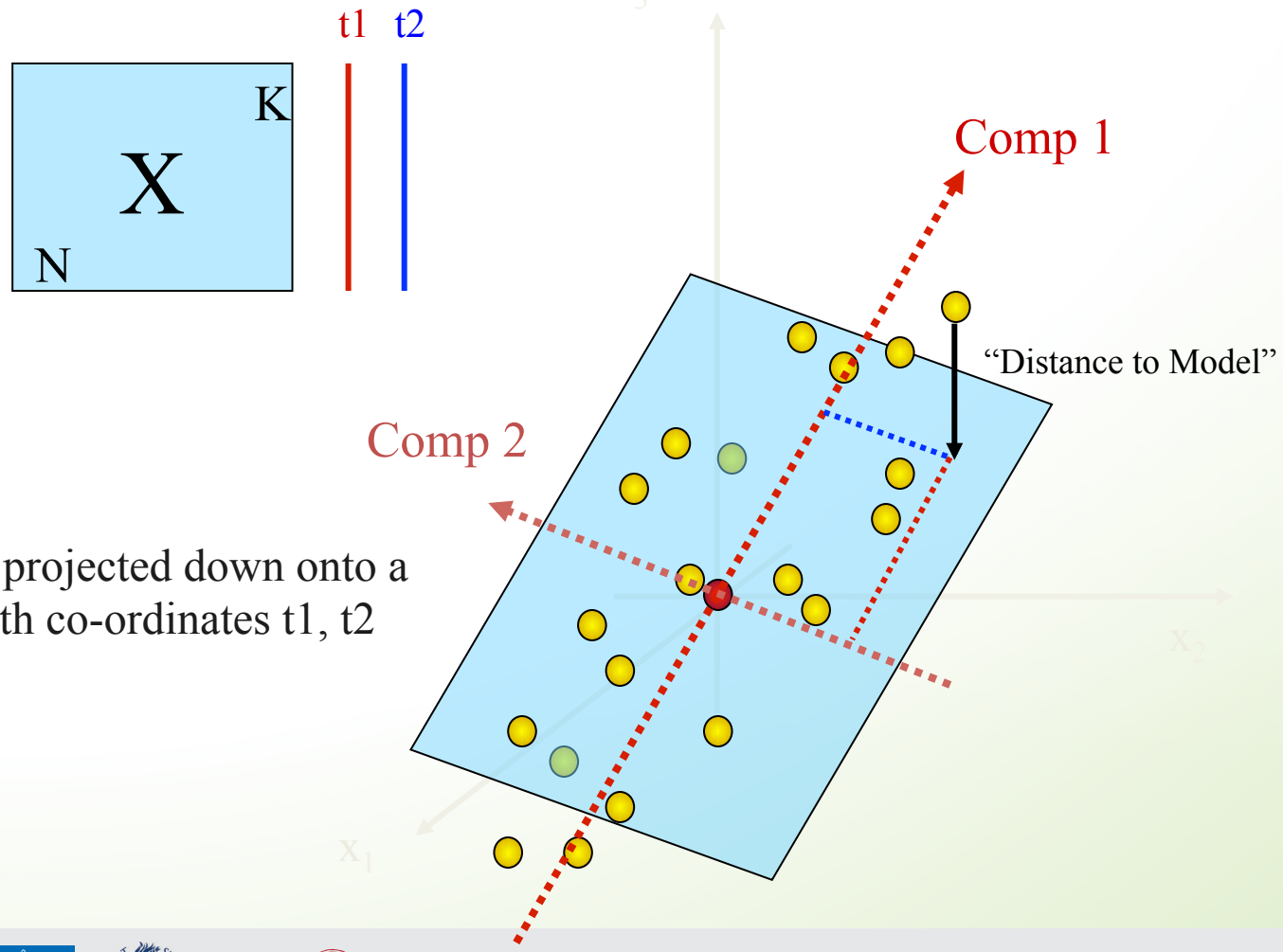
Fit first principal component (line describing maximum variation)

Add second component (accounts for next largest amount of variation) and is at right angles to first - orthogonal

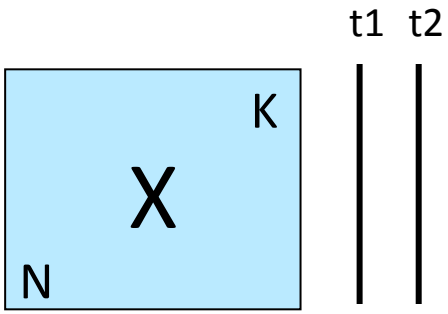


Each component goes through origin

PCA - Geometric Interpretation

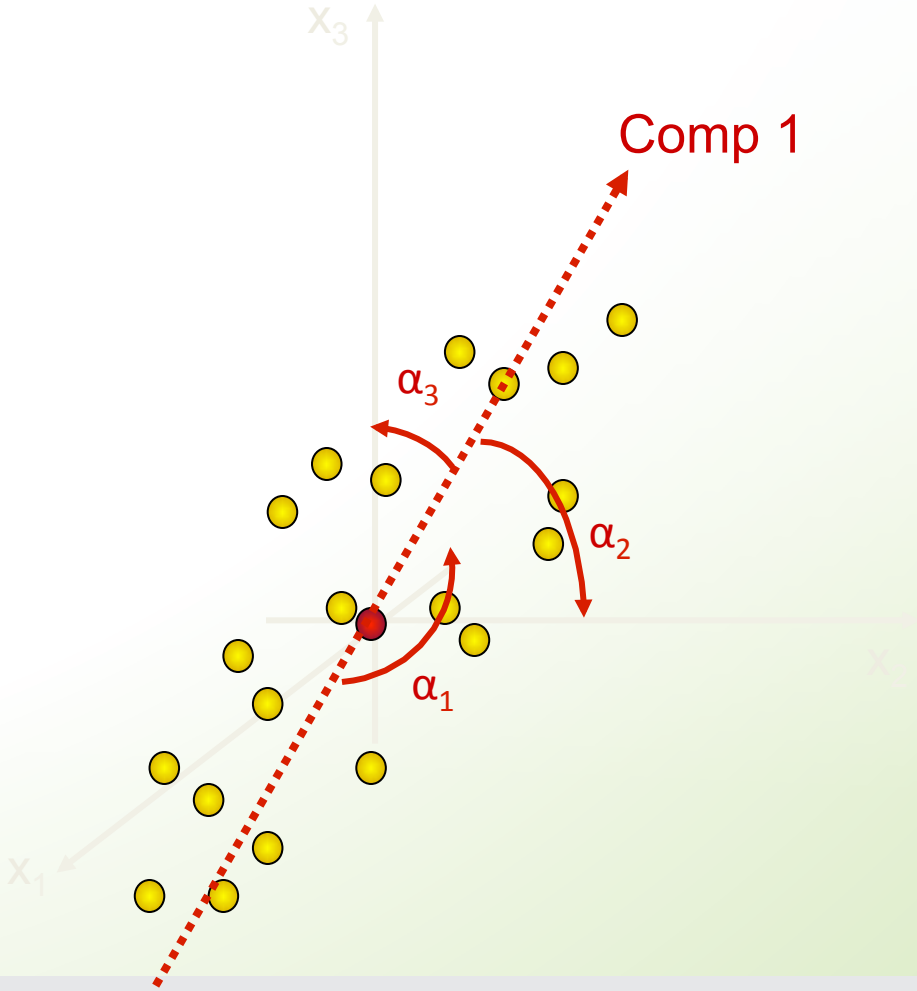


Loadings

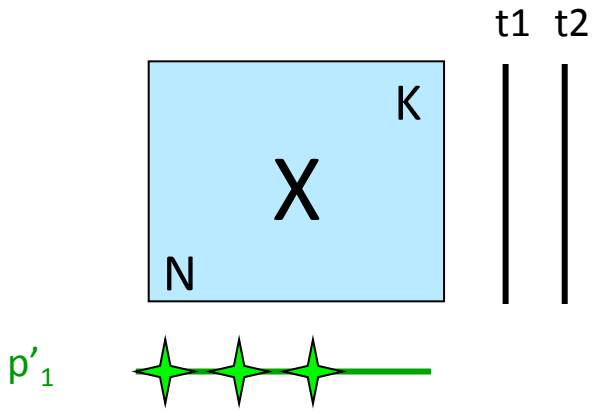


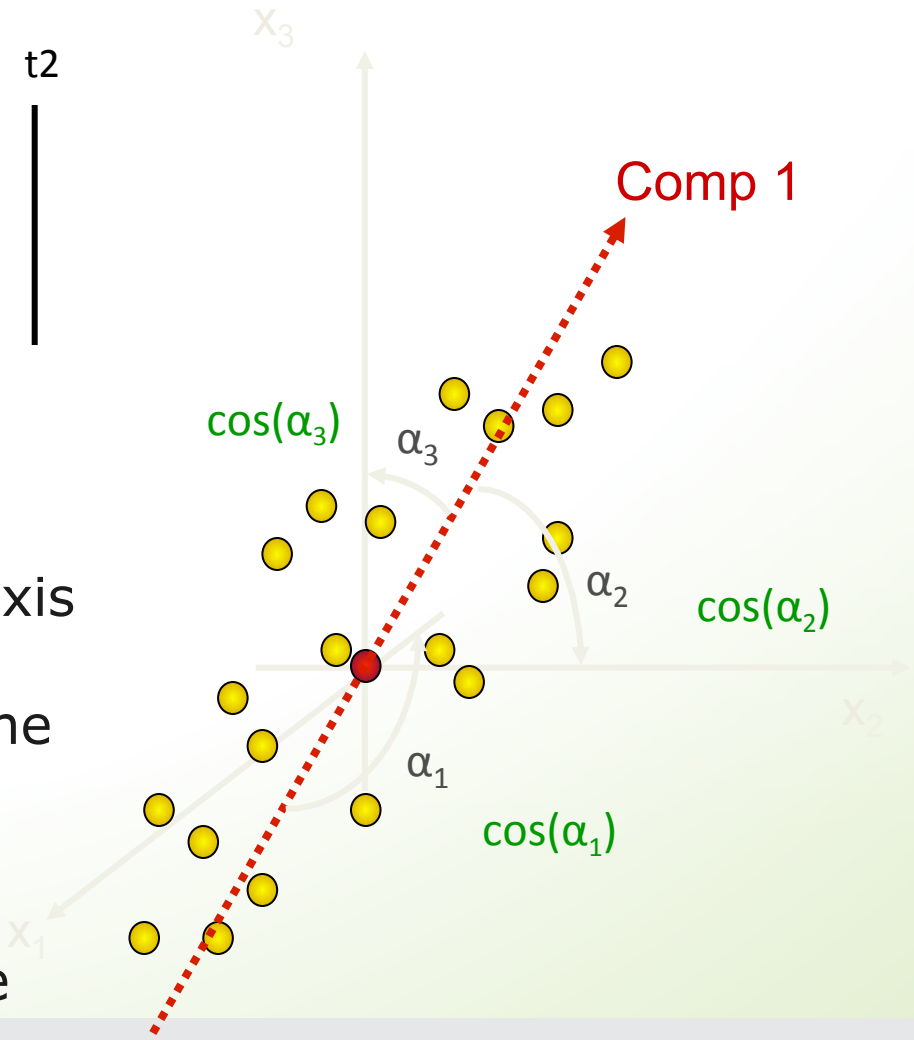
How do the principal components relate to the original variables?

Look at the angles between PCs and variable axes



Loadings





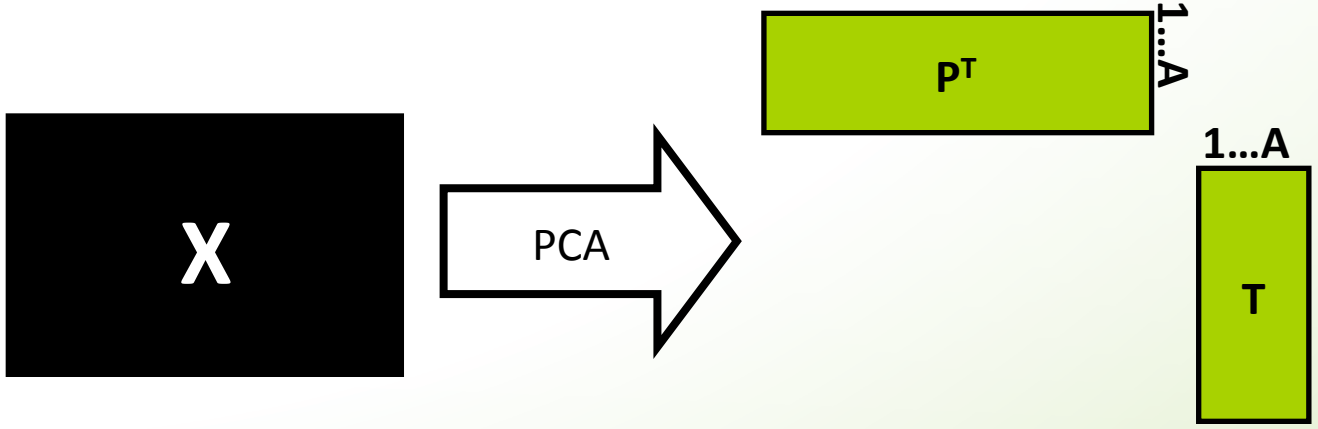
Take cos(α) for each axis

Loadings vector p' - one for each principal component

One value per variable

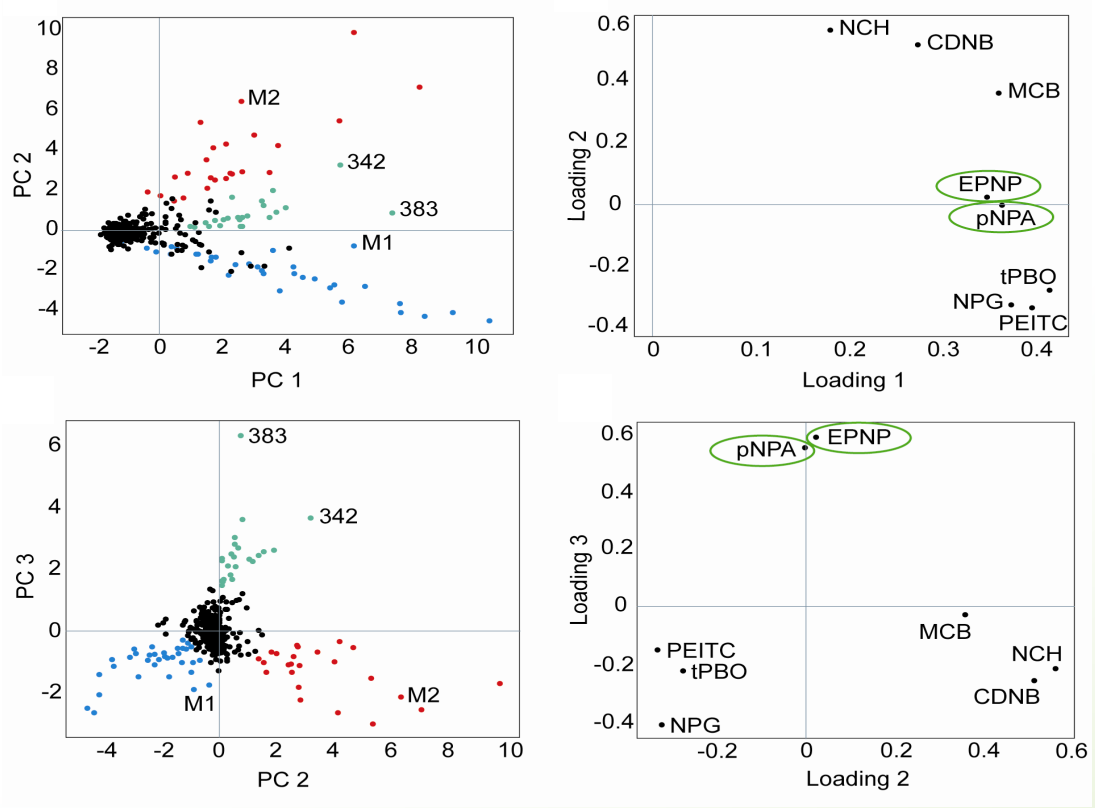
Principal component analysis (PCA)

- PCA compress the **X** data block into **A** number of orthogonal components
- Variation seen in the score vector **t** can be interpreted from the corresponding loading vector **p**



PCA Model $X = t_1p_1^T + t_2p_2^T + \ldots + t_Ap_A^T + E = TP^T + E$

Recognition of molecular quasi-species (evolving units) in enzyme evolution by PCA

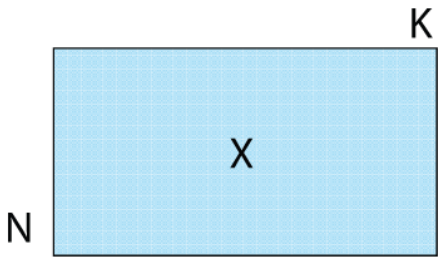


Emrén, L., **Kurtovic, S.**, Runarsdottir, A., Larsson, A-K., & Mannervik, B. (2006) Proc Natl Acad Sci U S A, 103, 10866-10870

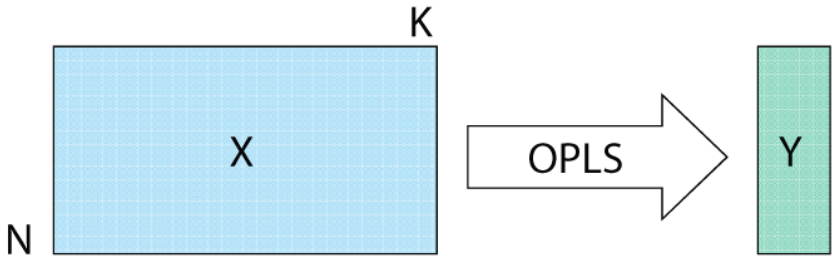
Kurtovic, S, & Mannervik B (2009) Biochemistry, 48, 9330-9339

Orthogonal partial least squares to latent structure – Discriminant analysis (OPLS-DA)

From PCA to OPLS



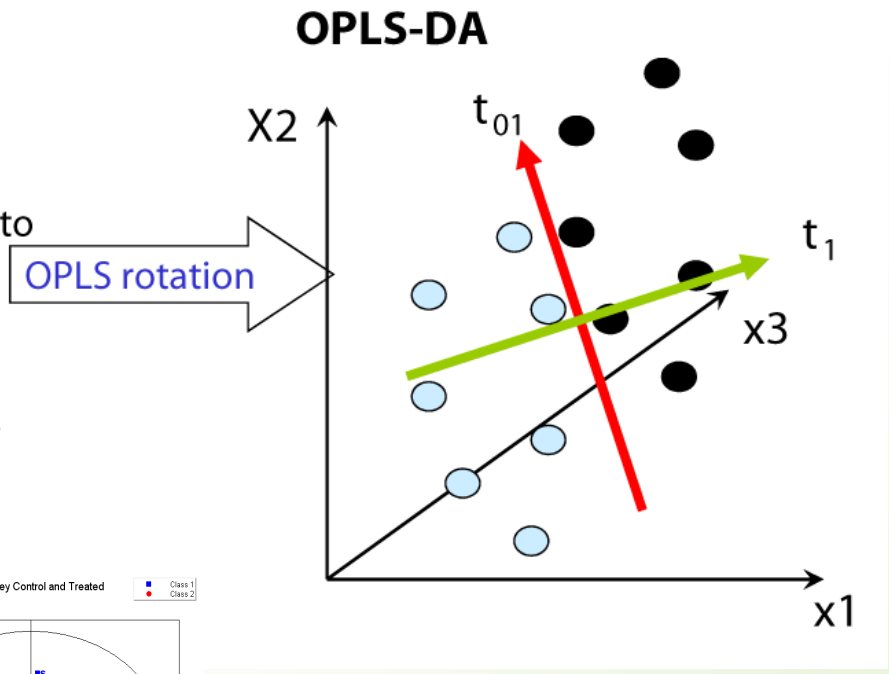
Unsupervised
PCA on X will find the maximal variation in the data. PCA is the basis of all multivariate modelling.

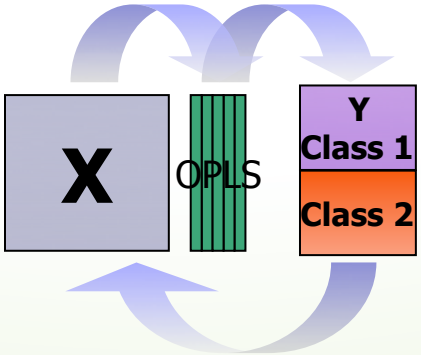


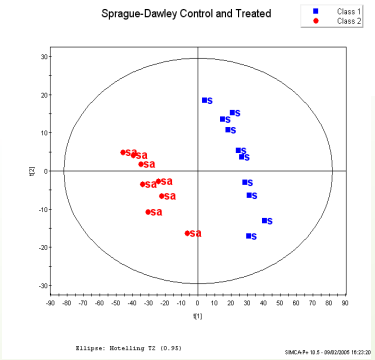
Supervised
OPLS is a prediction and regression method that finds information in theX data that is related to known information, theY data.

Orthogonal partial least squares to latent structure – Discriminant analysis (OPLS-DA)

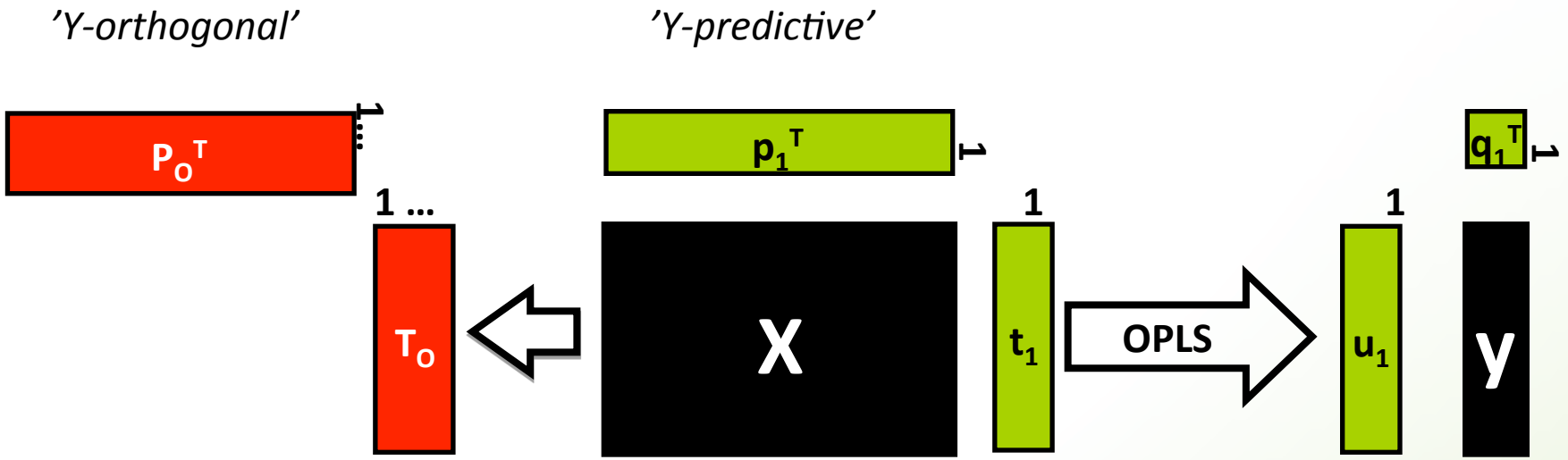
- OPLS-DA finds the variation in X that is correlated the Y variable
- This is done by a rotation towards the direction of Y
- At the same time OPLS-DA finds components that are uncorrelated to Y but systematic in X
- As in PCA, the data should first be scaled and centred
- Each observation is represented by a point in multi-dimensional space







OPLS with single Y / modelling and prediction

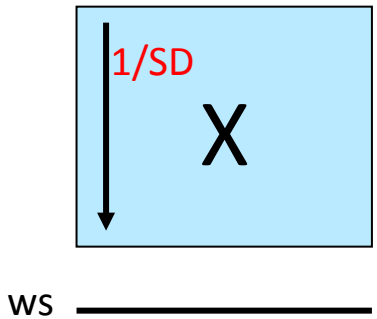


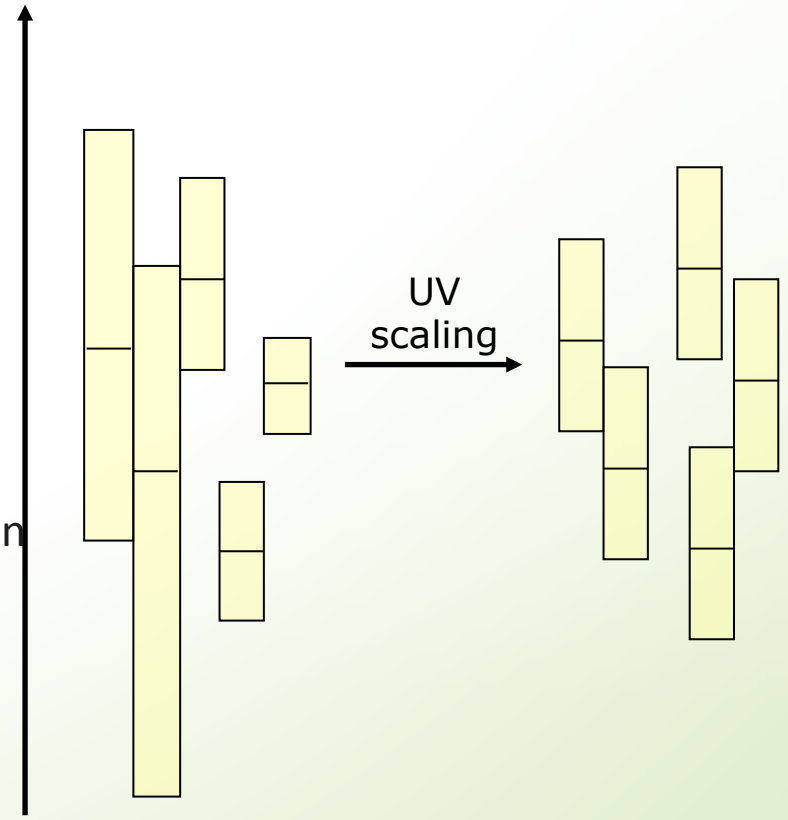
OPLS Model

$$\begin{cases} X = t_1 p_1^T + T_O P_O^T + E \\ Y = t_1 q^T_1 + F \end{cases}$$

Data Preprocessing – Scaling

- PCA and other methods are scale dependent
  - Is the size of a variable important?

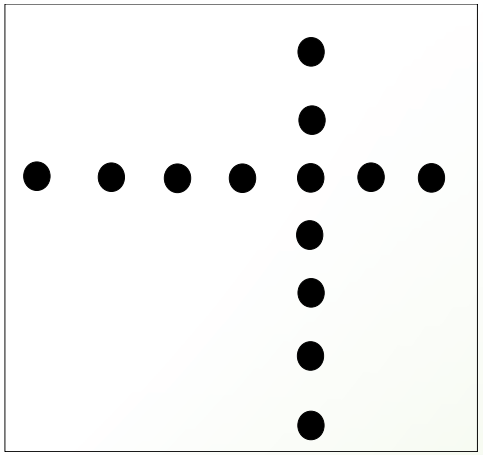




- Scaling weight is 1/SD for each variable i.e. divide each variable by its standard deviation – Unit Variance Scaling
- Variance of scaled variables = 1
- Many other kinds of scaling exist

Cross-Validation

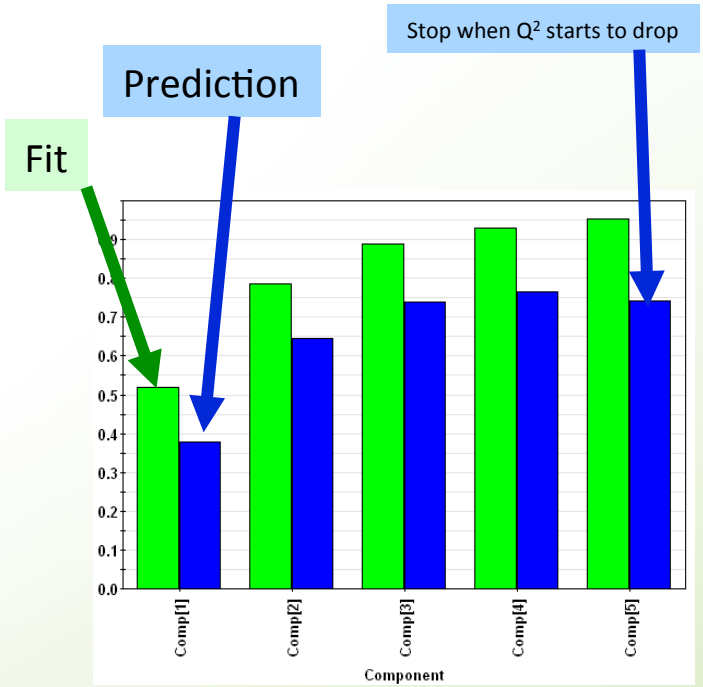
- Data are divided into G groups (default in SIMCA-P is 7) and a model is generated for the data devoid of one group
- The deleted group is predicted by the model ⇒ partial PRESS (Predictive Residual Sum of Squares)
- This is repeated G times and then all partial PRESS values are summed to form overall PRESS
- If a new component enhances the predictive power compared with the previous PRESS value then the new component is retained



- PCA cross-validation is done in two phases and several deletion rounds:
  - first removal of observations (rows)
  - then removal of variables (columns)

Model Diagnostics

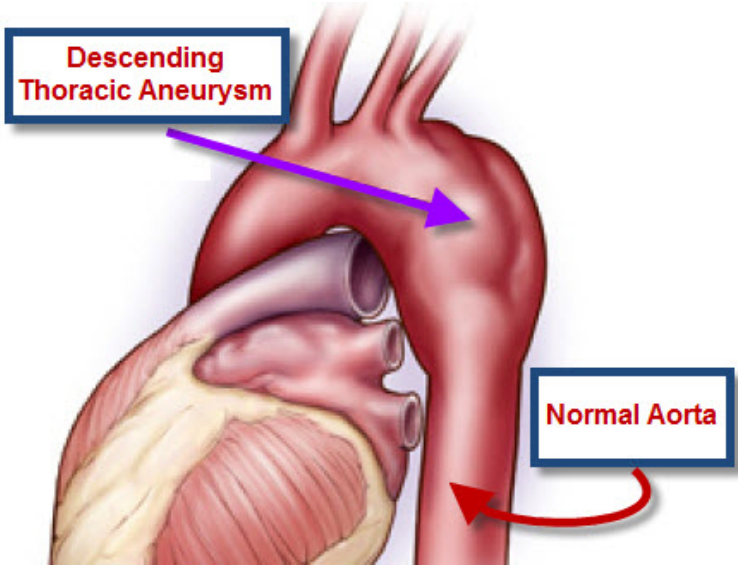
- **Fit or R^2**
  - Residuals of matrix E pooled column-wise
  - Explained variation
  - For whole model or individual variables
  - RSS = Σ (observed - fitted)^2
  - R^2 = 1 - RSS / SSX
- **Predictive Ability or Q^2**
  - Leave out 1/7th data in turn
  - 'Cross Validation'
  - Predict each missing block of data in turn
  - Sum the results
  - PRESS = Σ (observed - predicted)^2
  - Q^2 = 1 – PRESS / SSX

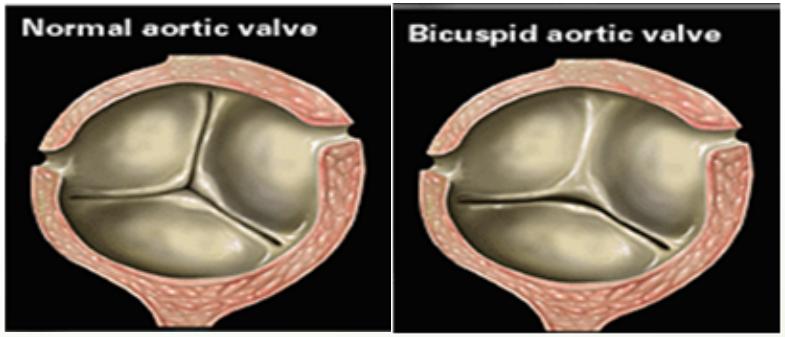


Kurtovic, Paloschi, Folkersen, Gottfries, Franco-Cereceda, Eriksson (2011) Molecular Medicine, **17**; 665-675

ALTERNATIVE SPLICING IN THORACIC AORTIC ANEURYSM

Thoracic aortic aneurysm (TAA)

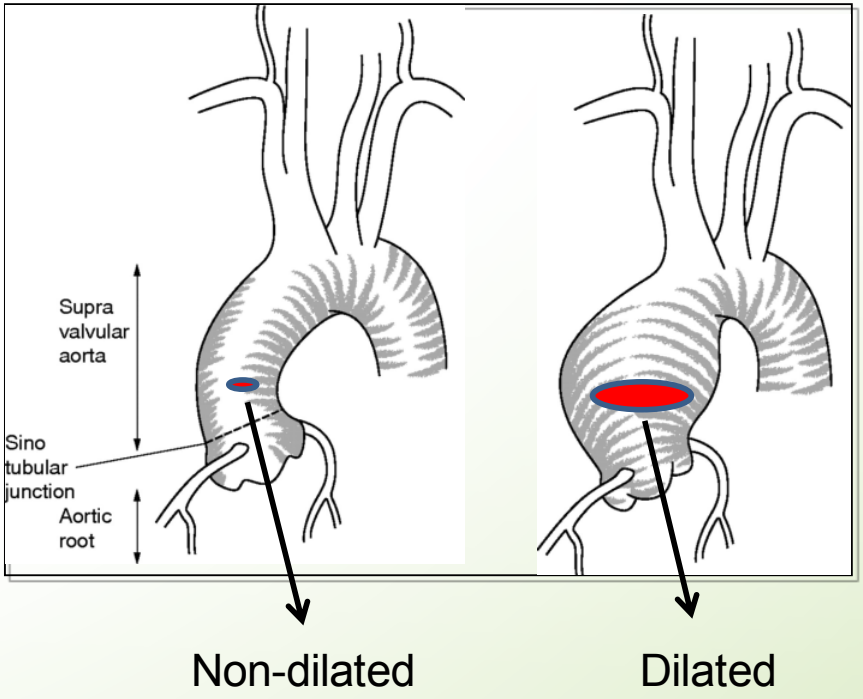




- Monogenic
  - Marfan syndrome
  - Loeys Dietz
- Aneurysm associated with bicuspid aortic valve (BAV)
- Idiopathic thoracic aortic aneurysm

Outline of the study

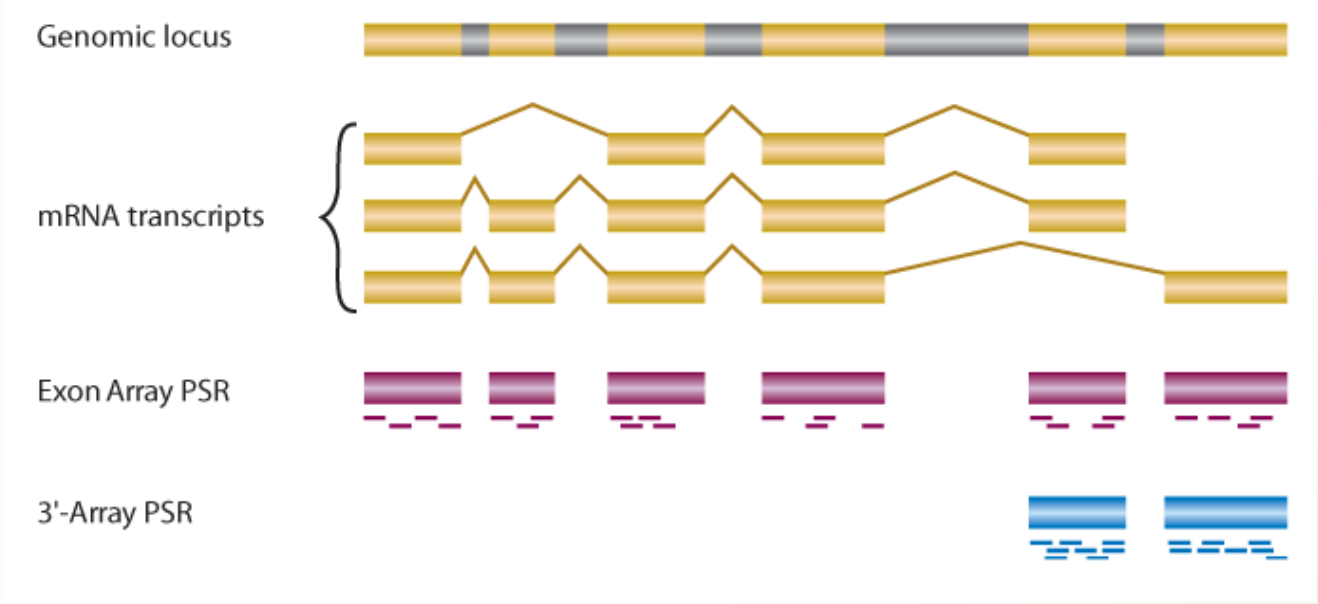
- Biopsies are collected from both non-dilated and dilated aorta during valve replacement surgery and reconstruction of the dilated aorta respectively
- Media from ascending aorta
- RNA
  - Affymetrix human exon 1.0 ST microarrays (in this study 81 patients)
  - RNAseq (30 patients)
- Protein
  - HiRiEF iTRAQ LC-MS/MS
  - 2D gel electrophoresis followed by iTRAQ LC-MS/MS



Aim of the study

- Alternative splicing in transforming growth factor-β (TGFβ) signaling pathway
- TGFβ pathway is known to be important in aortic aneurysm
- Are there any alternatively spliced genes in the TGFβ pathway?
- Is alternative splicing an important mechanism in thoracic aortic aneurysm (TAA)?
- How do we analyze alternative splicing?

Affymetrix exon array design



Exons

Introns

PSR – probe selection region

Preprocessing of data

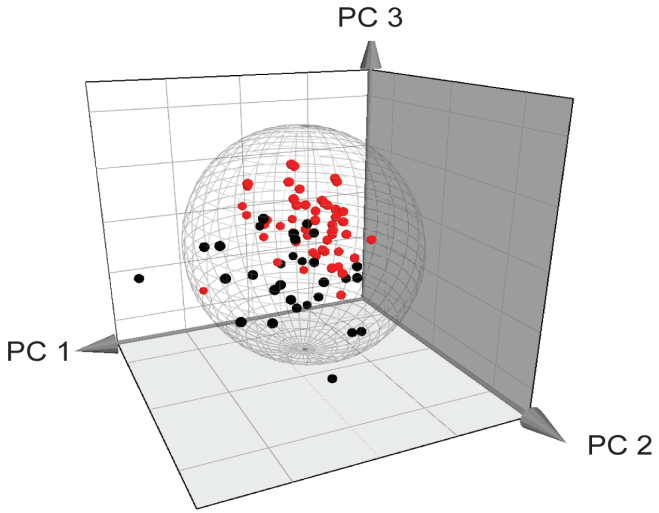
- Probe set core level
- Unique hybridization target
- Robust multichip average (RMA) normalized
- Splice Index calculated (in case of exon level analysis)

i = exon
j = sample
k = gene
e = exon signal
g = gene signal

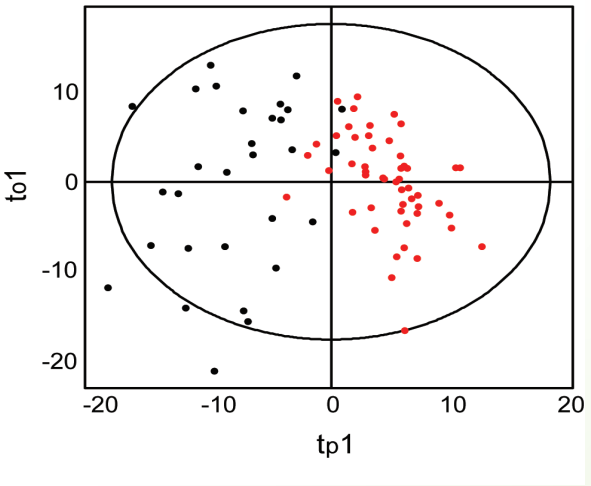
$$n_{i,j,k} = \frac{e_{i,j,k}}{g_{j,k}}$$

- Unit variance scaled and mean centered data prior to MVA

Alternative splicing pattern in the TGFβ pathway is different between dilated and non-dilated aorta



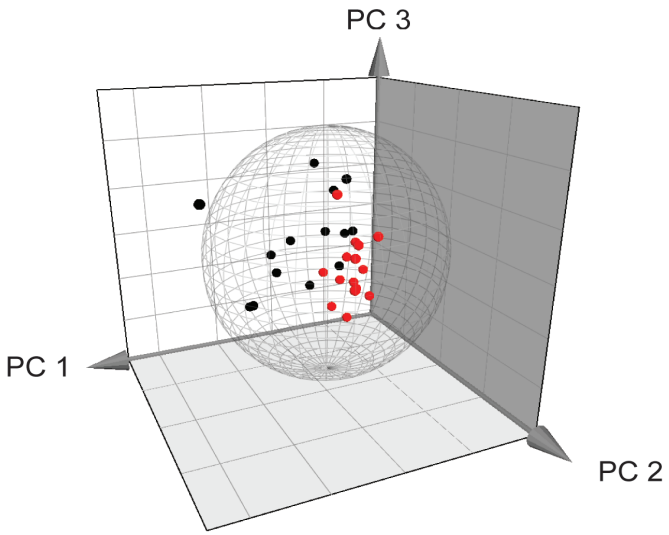
Non-supervised PCA



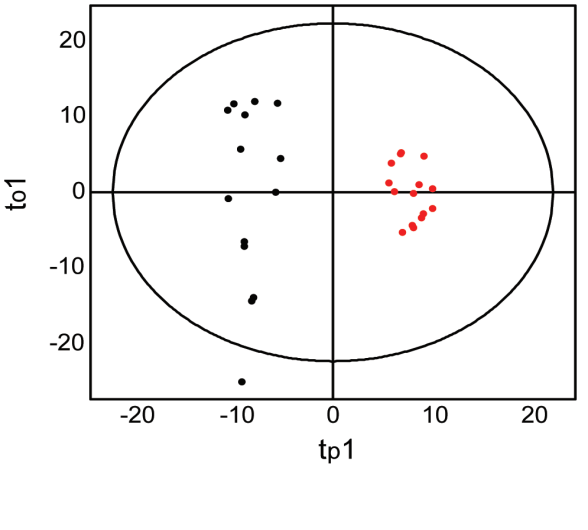
Supervised OPLS-DA

- **TAV and BAV together**
- 81 patients included
- 614 exons included
- Good model
- Good separation between the two groups

Alternative splicing pattern in the TGFβ pathway is different between dilated and non-dilated aorta



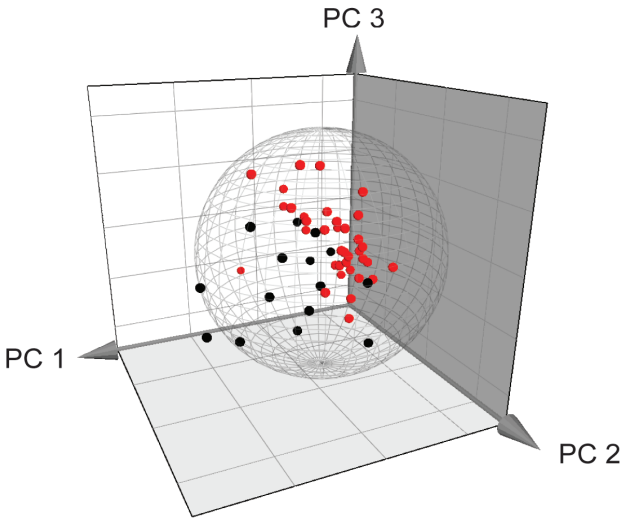
Non-supervised PCA



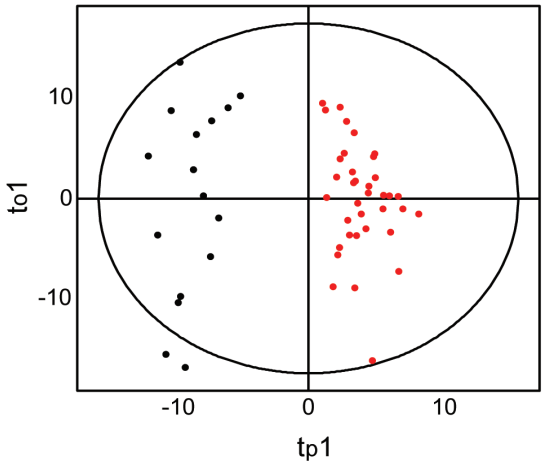
Supervised OPLS-DA

- **Only TAV patients**
- 29 patients included
- 614 exons included
- Good model
- Good separation between the two groups

Alternative splicing pattern in the TGFβ pathway is different between dilated and non-dilated aorta



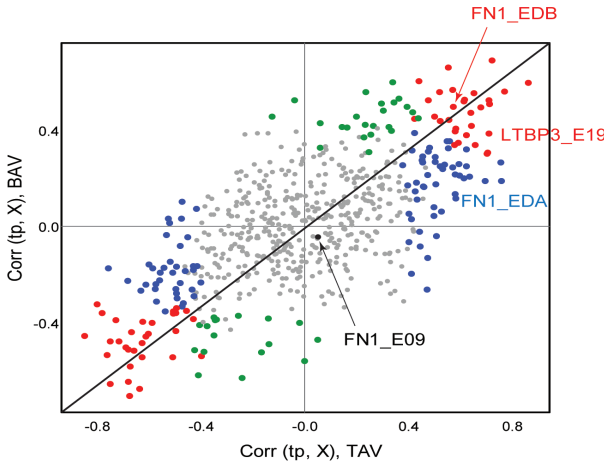
Non-supervised PCA

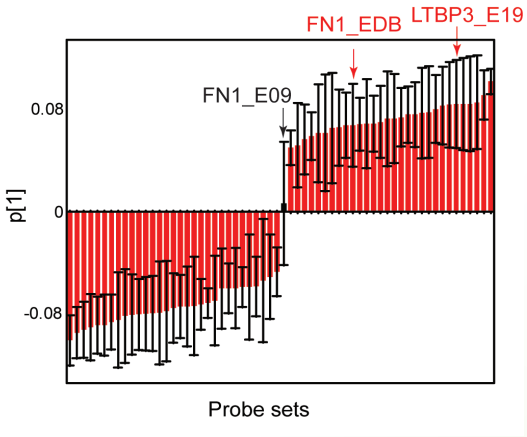


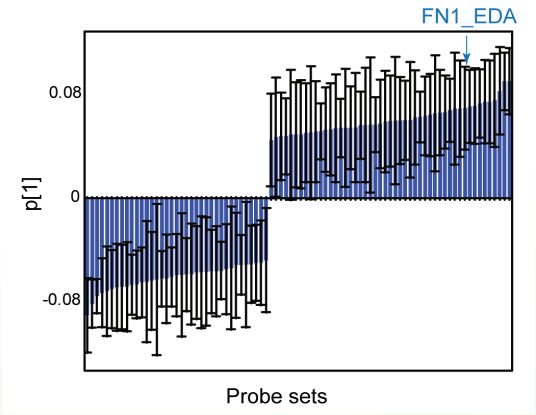
Supervised OPLS-DA

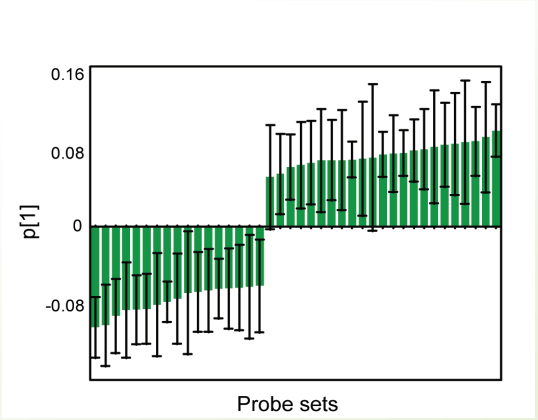
- **Only BAV patients**
- 52 patients included
- 614 exons included
- Good model
- Good separation between the two groups

Alternatively spliced exons are present in both TAV and BAV groups of patients

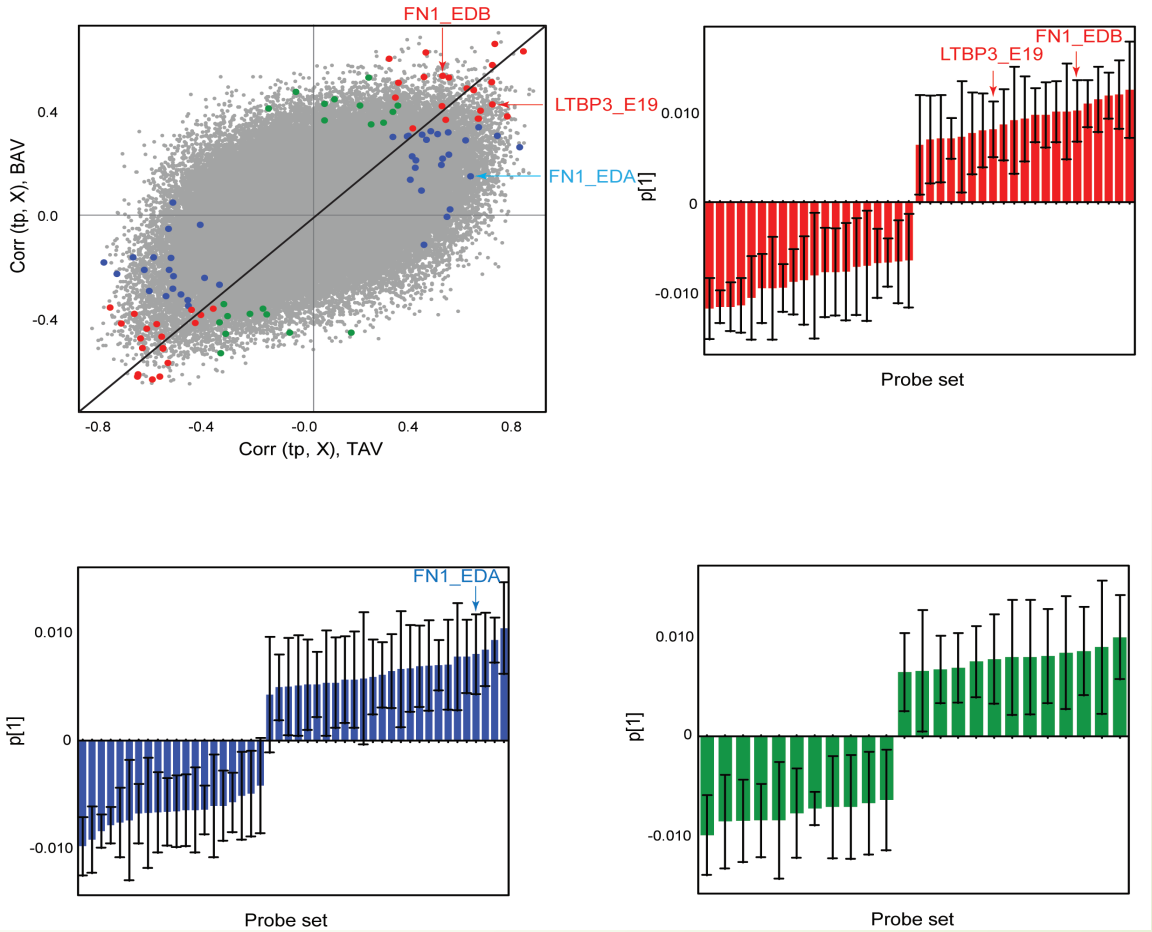




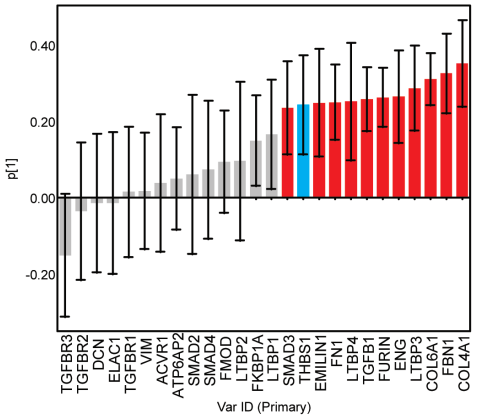


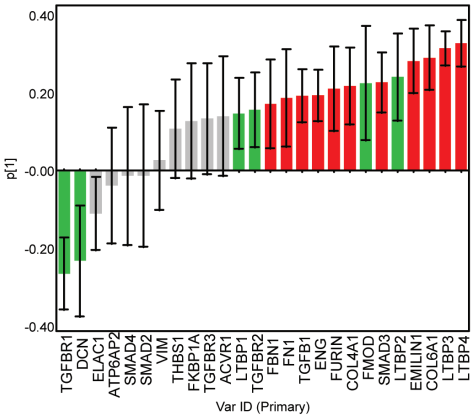


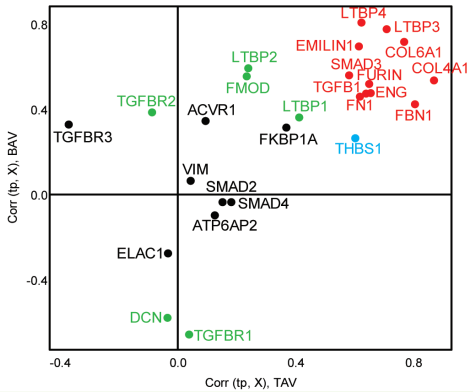
Alternative splicing analysis of all exons in the human genome reveals the importance of TGFβ pathway exons

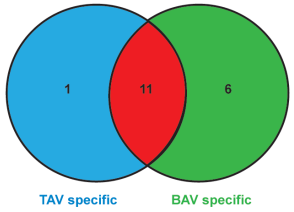


Gene expression patterns of differentially spliced genes









Summary

- TGFβ pathway exons clearly important according to an overall exon level analysis
- Dilated and non-dilated aortas show different alternative splicing patterns in dilated and non-dilated tissues with respect to TAV and BAV in TGFβ pathway
- Exons responsible for the diverging alternative splicing fingerprints in TGFβ pathway identified
- Implies that dilatation in TAV has different underlying molecular mechanisms compared to BAV patients
- New methods for analyzing array data

Today during the exercise

- PCA and OPLS-DA
- Thoracic aortic aneurysm data set
- Exon level expression Affymetrix arrays
- Compare two different phenotypes and subphenotypes