MULTIVARIATE ANALYSIS OF METABONOMICS DATA

Chris Ambrozic

Umetrics Inc.

www.umetrics.com

UMETRICS

Research & Development involves, among others:

- **Ideas** ← Creativity, Knowledge, Insight
- **Checking ideas** ← Experimentation, Measurements
  Analysis of Data and Interpretation

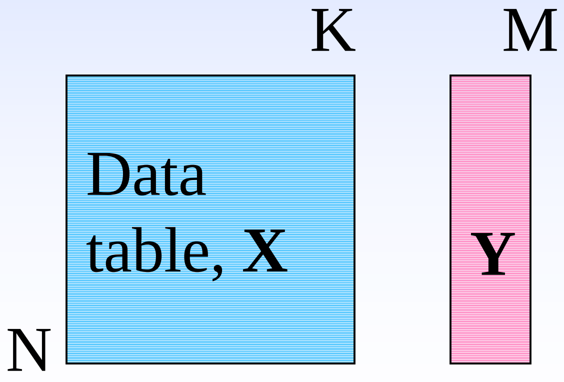
Modern instrumentation – spectrometers (NMR, X-Ray, MS, IR, ….)
chromatography, EF, gene-arrays, …,

and samples, genes, proteins, cells, urine, blood,…….

provide LOTS of data highly multidimensional (K > 1000)

Mega and Giga-variate

**Pull out information from data,
but not more, and not less**



Software issues

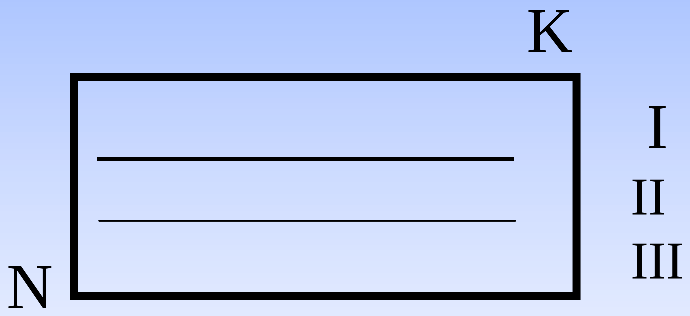
- Software packages are an integral part of metabonomics analysis
- Integrated part of tools, not separate issue
- Subject to 21CFRpt 11& regulatory concerns
- Calculations must be understandable
- And science based
- Results must be interpretable
- And quantitative
- And reproducible

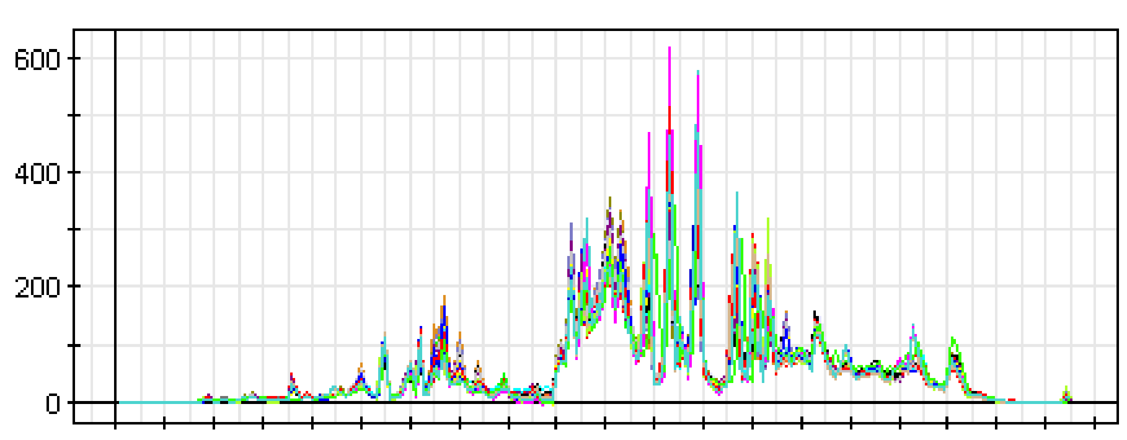
Metabonomics Analysis implementation needs the following:

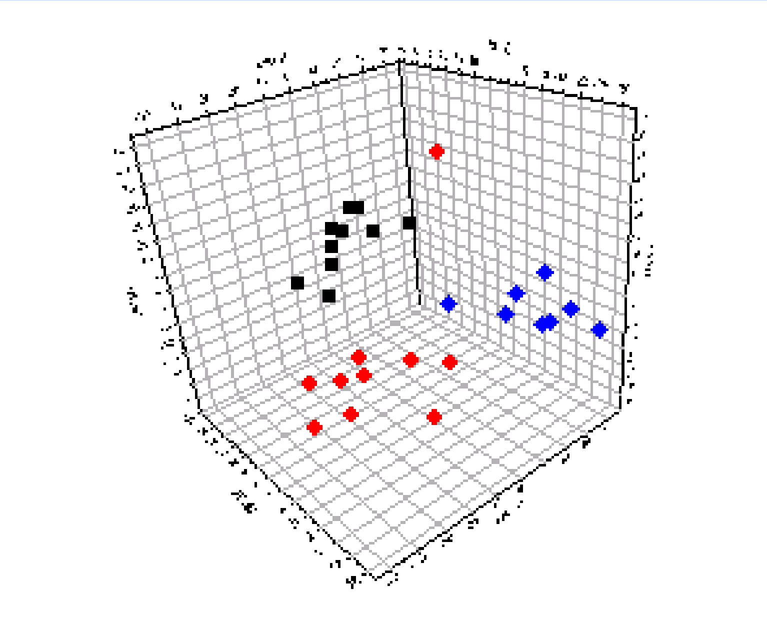
- Planning & Organization
- Process knowledge – what and where to measure
- Hardware
- Software
- Education & Training
  - operators
  - engineers & scientists
  - managers & executives
  - regulatory agencies (++)
  - academic community (- -)

Ex.1 Classification of rats (Sprague-Dawley) **controls** vs exposed to amiodarone or chloroquine using metabonomic profiling. (Data from Eriksson, Antti, Holmes and Johansson, Tox Met, 2003)

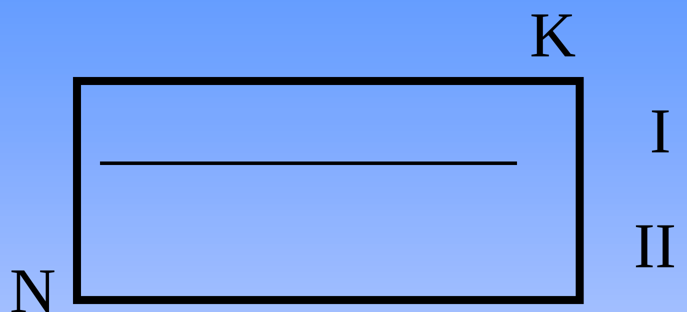
- N=28
- K=197
- G=3







Traditional analyses; COST, cross-tab, t-tests, regression, *inadequate and misleading*. Why ?



Risk for ***spurious results*** when testing K times, e.g., for group differences, or for correlations

risk = $1-0.95^K$

K =	1	10	30	100
risk=	.05	.40	.79	.994

Basic Assumption:

independent variables

- –absurd when K > 10-20
- –spurious results when tested independently
- –information about complicated systems sits in ***combinations*** of variables !

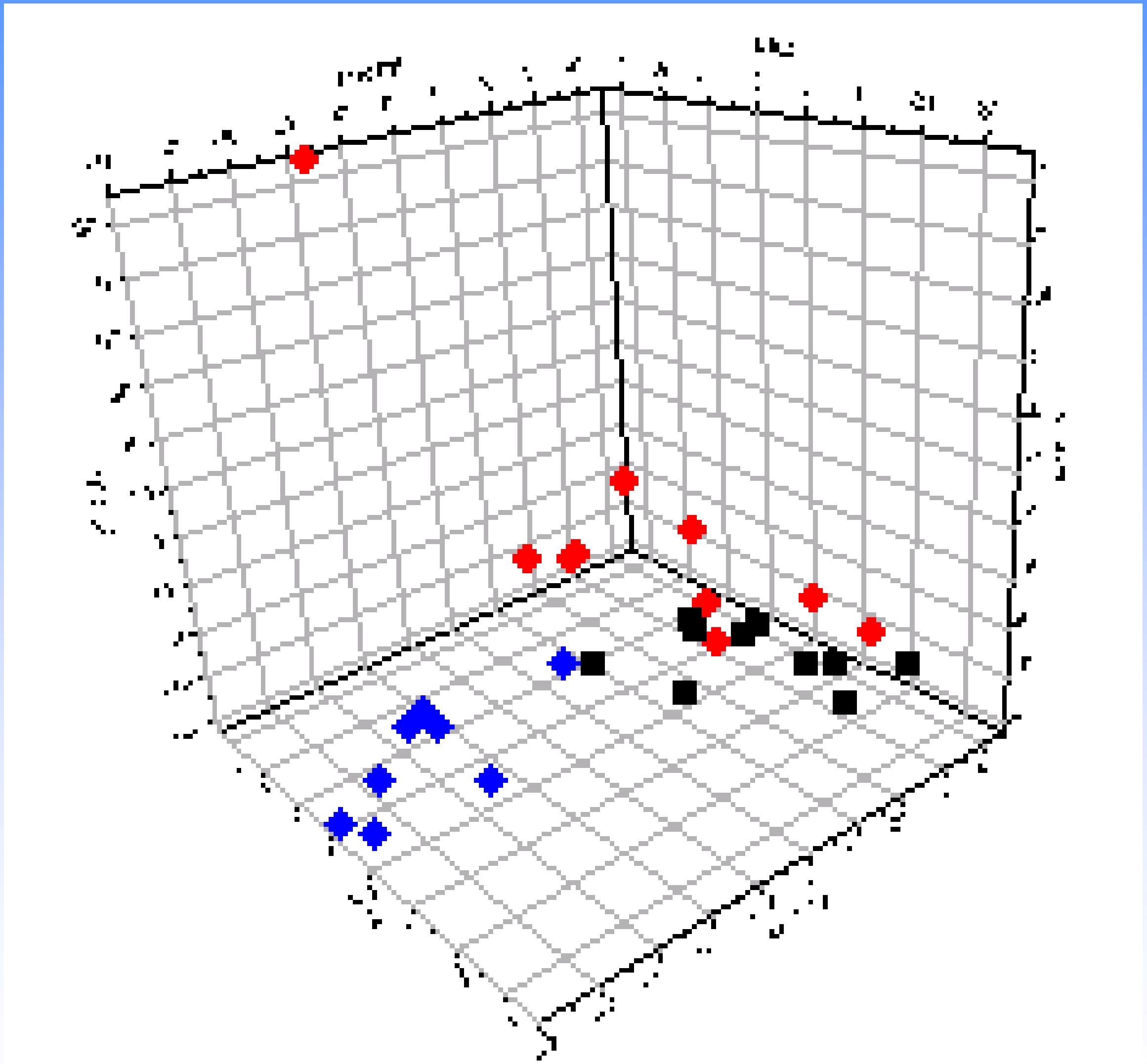
COST approach does not give your research ideas a fair chance !

Data from complicated systems (David Botstein, 2002)

- Correlated patterns more robust than individual measurements
  - Look at all variables together
- Patterns based on ALL data
  - Look at all observations (samples, cases) together
- Importance ≠ Significance
  - Have separate criteria for importance and significance
- Open access to data ⇒ reanalysis
  - Desirable redundancy and reliability

1. Not one variable at a time (confusion, false positive)
But, PCA of normalized data matrix (N=28 x K=197)

PC scores,
t_1 & t_2 & t_3
(optimal
summaries),
show *some*
separation.



Convincing,

but….

2. A more efficient class separation by PLS-DA

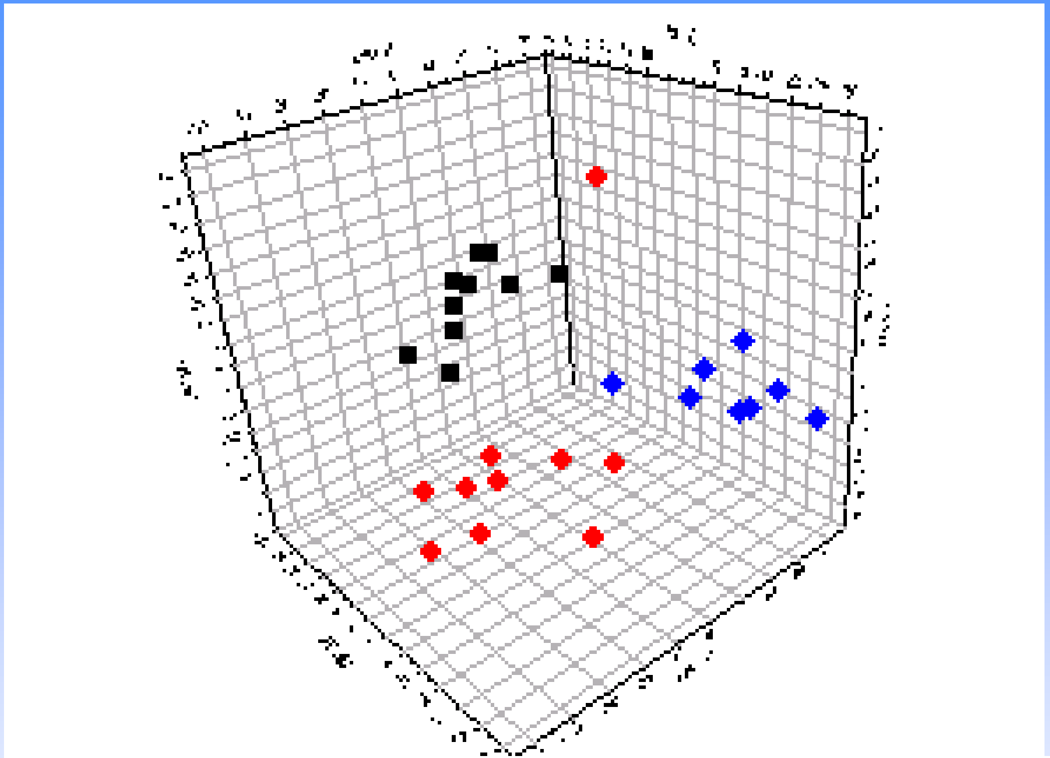
PLS-DA scores, t_1 & t_2 & t_3 show a clear separation between the three classes

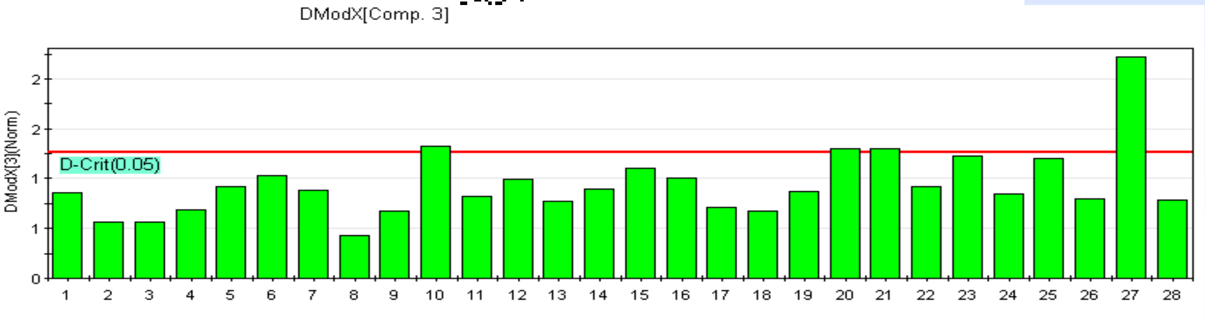
Ctrl

S_chlorquine

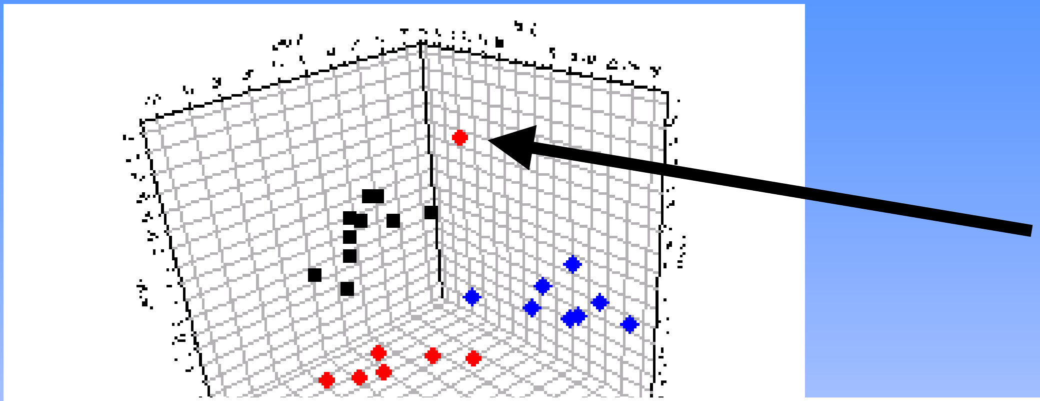
S_amiodarone

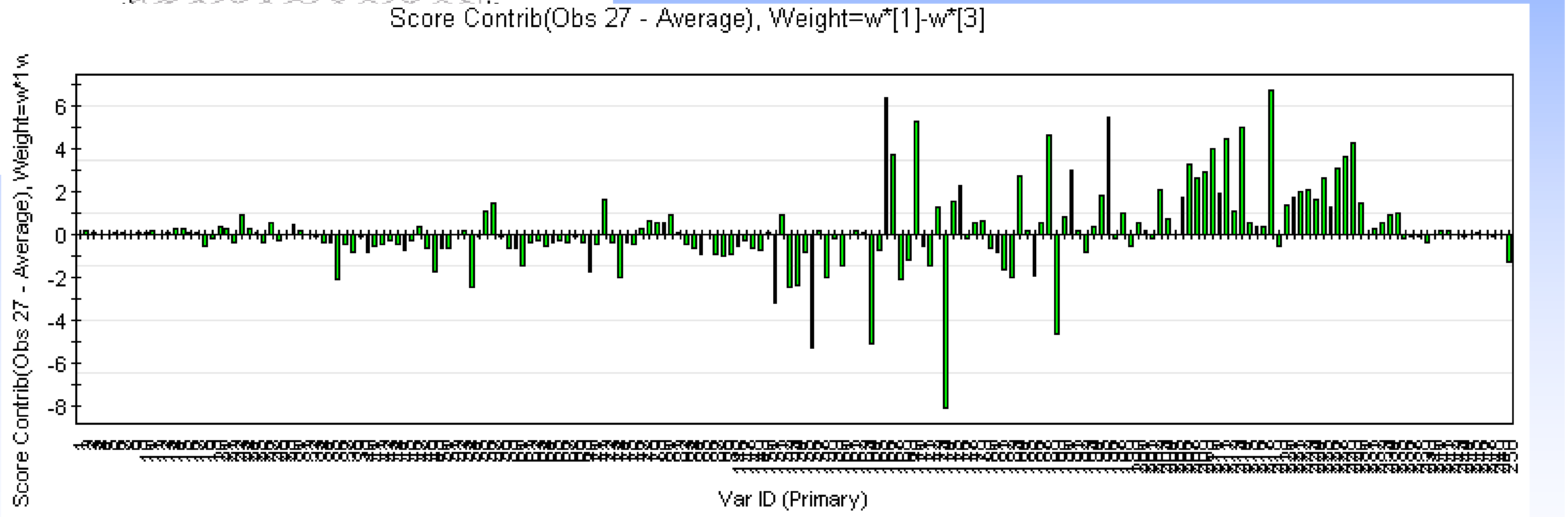
# 27 is out, also in DModX (lower plot)





Why is # 27 an outlier ?   Contribution Plot



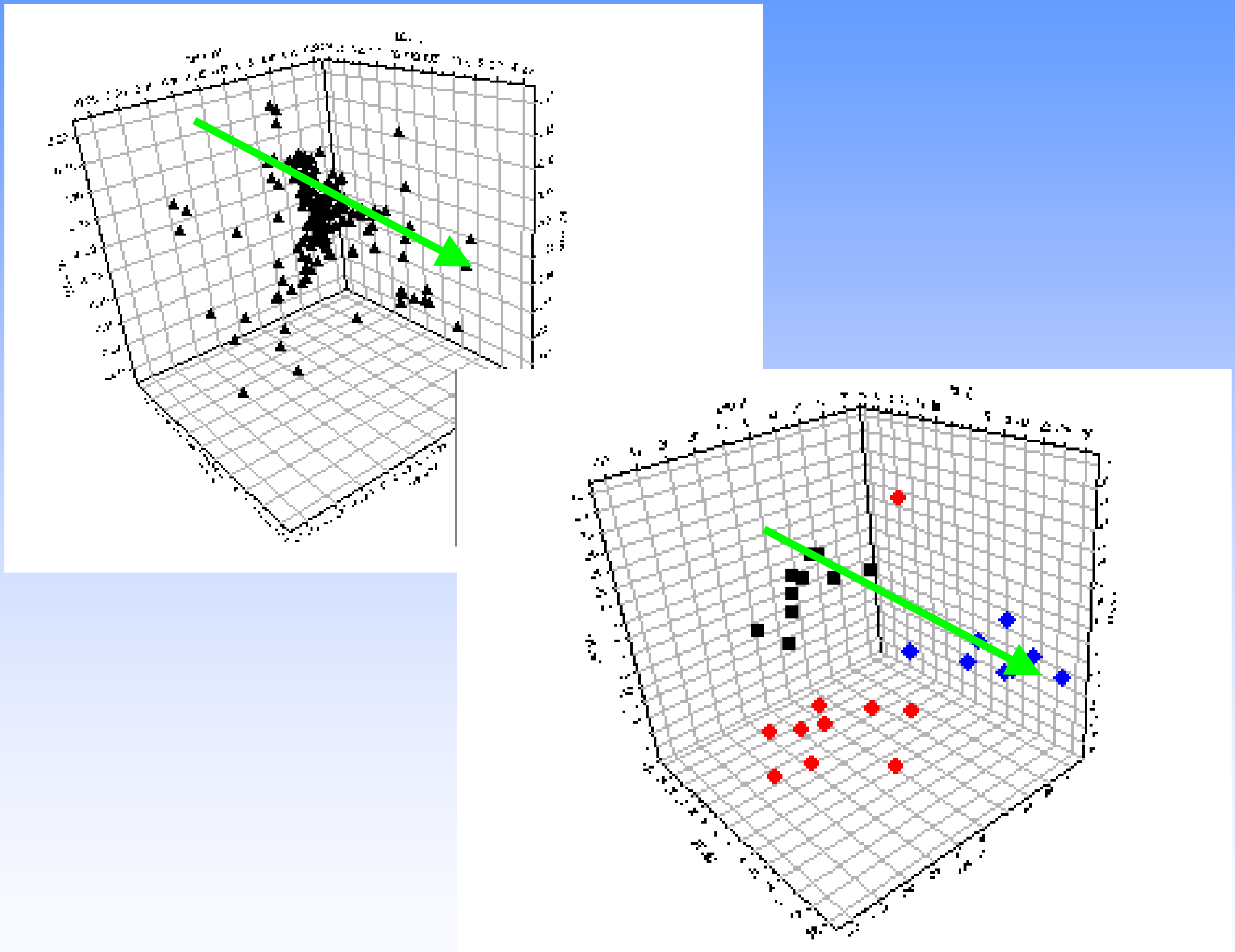


UMETRICS

2b. The PLS-weights (w_1 & w_2 & w_3) indicate which variables that together separate the classes

Each point in the plot marks

a **variable**.



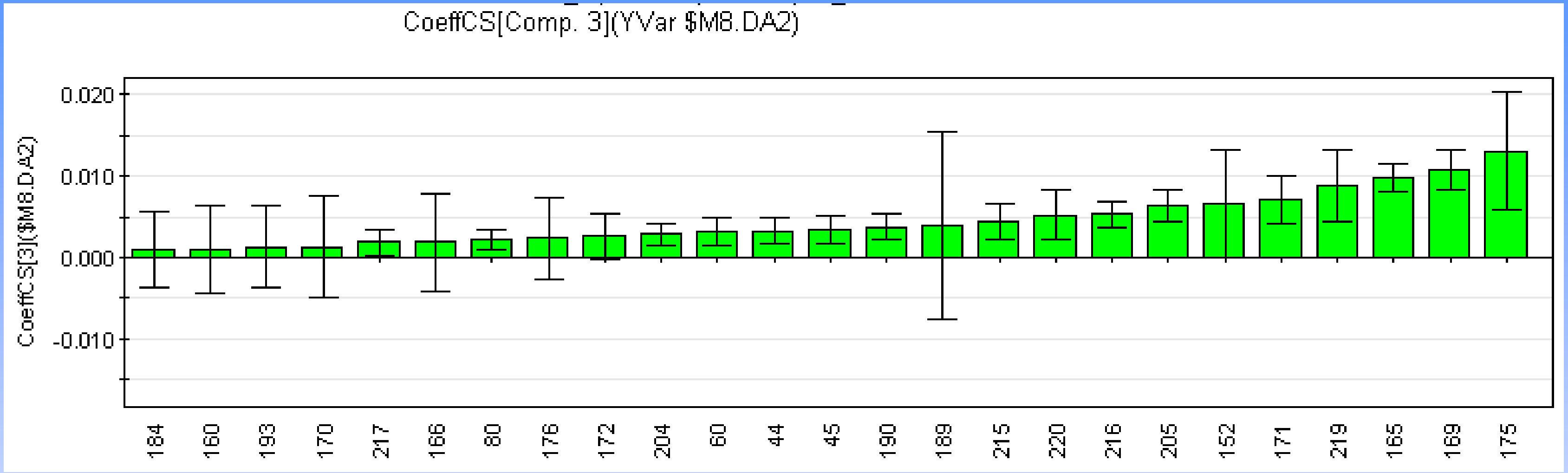
Directions in score plot ***correspond*** to

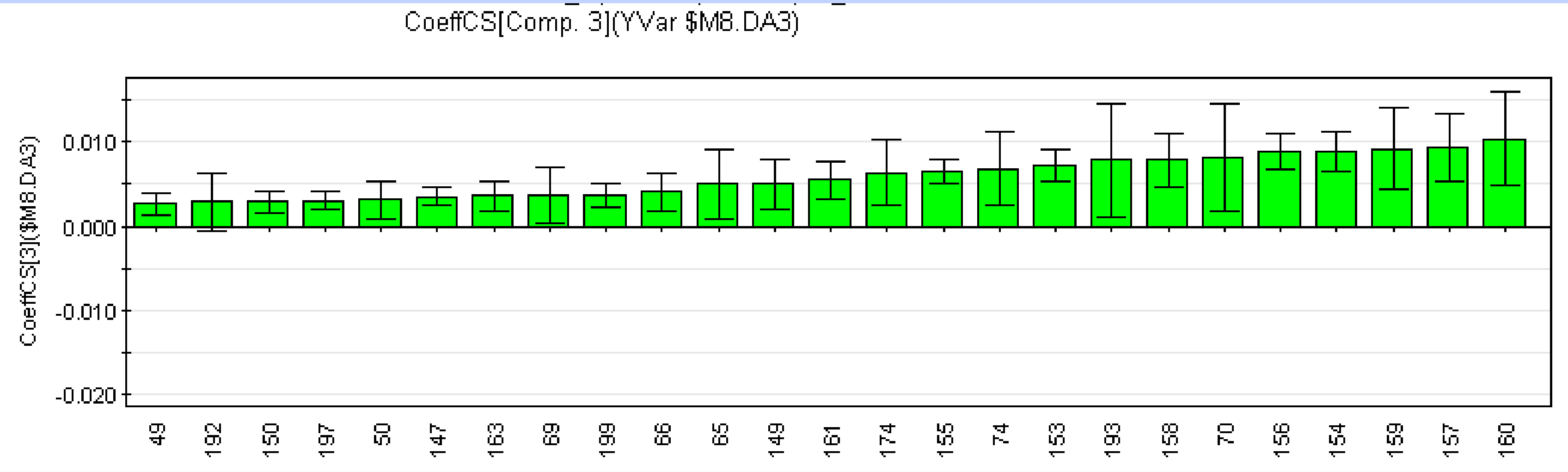
directions in weight plot

(loading plot)

25 Largest Discriminant Coefficients s_c

size ↔ importance; error bar ↔ significance





We need *tools and models* (simplifications); intuition is not a sufficient basis for data analysis.

“If our brains were simple enough for us to understand them, we’d be so simple that we couldn’t.”

Jack Cohen and Ian Stewart: The Collapse of Chaos.

Hofstadter, Wiener, Gödel, Schrödinger, Heisenberg, Bohr, …

Postulate: This generalizes to all biological systems

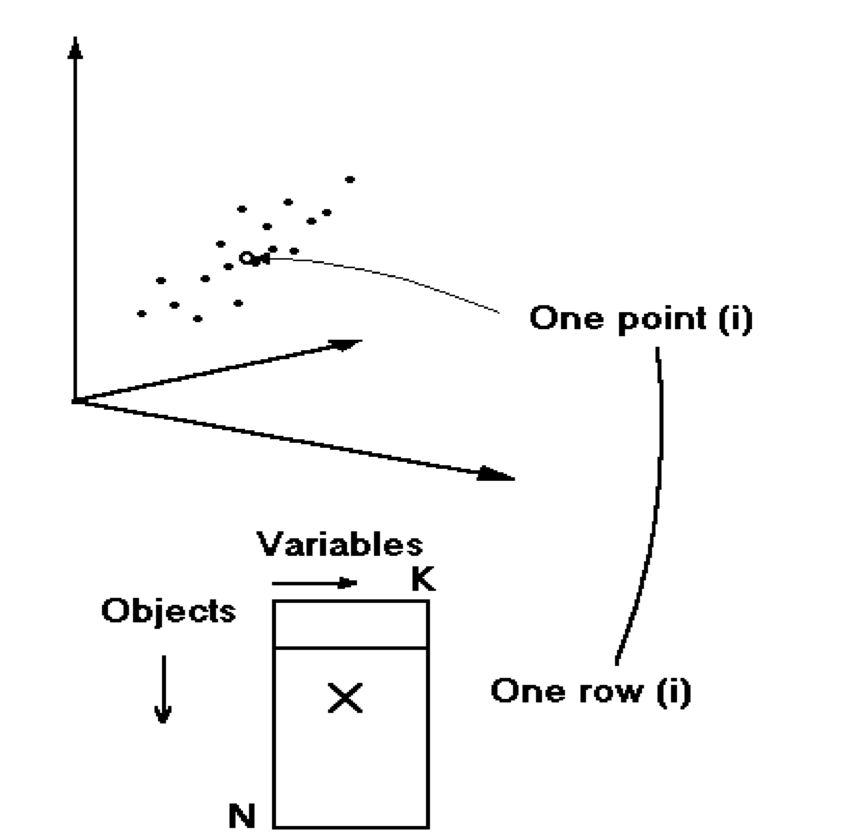
Consequence: Our brains alone are not sufficient for the analysis of these systems

Metabonomics, xxx-omics

- Each sample (tissue, blood, urine, cell, ….) is characterized by LOTS of data, typically 200 to 20000 numbers (variables, peaks, …), ***multivariate profiles, “finger prints”***
- No good theory how (and if) the profiles are related to the current question / problem
- The data contain patterns NOT related to the current question, and also various types of noise.
- Questions: **Classification** and/or ***Quantitative relationships***
- One desires quantitative results including
  - dominating variables (peaks) in relation to questions
  - similarities / dissimilarities of samples.
  - estimates of signal /noise, etc., reliability, precision, …
  - understandable displays

Tools: Multivariate analysis by means of projections

(data often are noisy, collinear, and incomplete)



- Data shaped as a table, **X**

- Space with K axes (K-space)

  K = number of variables (col.s)

  Each obs. (process time point)

  is a point in this space

- Multivariate analysis

  –finding structures in M-space

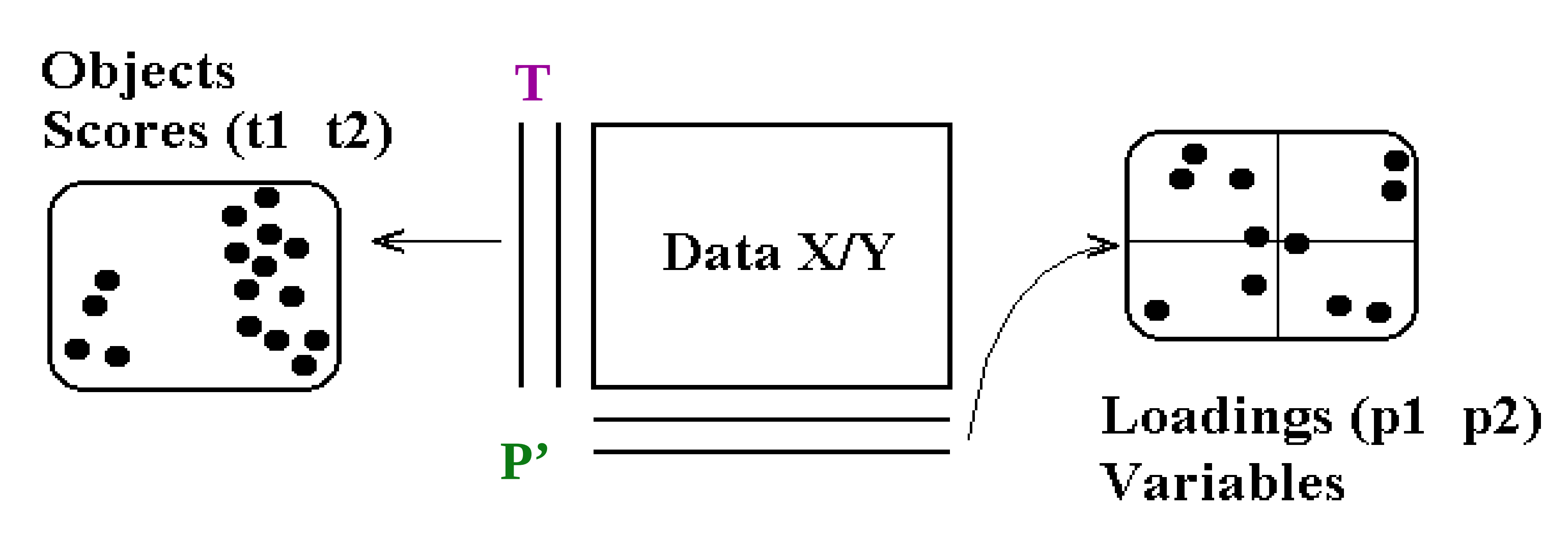
  –describing them (math & stat)

  –using them for problem solving

  –and for predictions

Data tables X approximated (summarized) as: X = T P' + E

Columns of T ↔ score plot. Rows of P' ↔ loading plot



Directions in score plot, correspond to directions in loading plot,

The scores, t_a, are optimal summaries, ***weighted averages*** of the variables

PCA: best summary of X
Principal Components Analysis

PLS: T also predicts Y
Projection to Latent Structures

Projection methods (PCA, PLS, ….) apply to: (analysis & predictions)

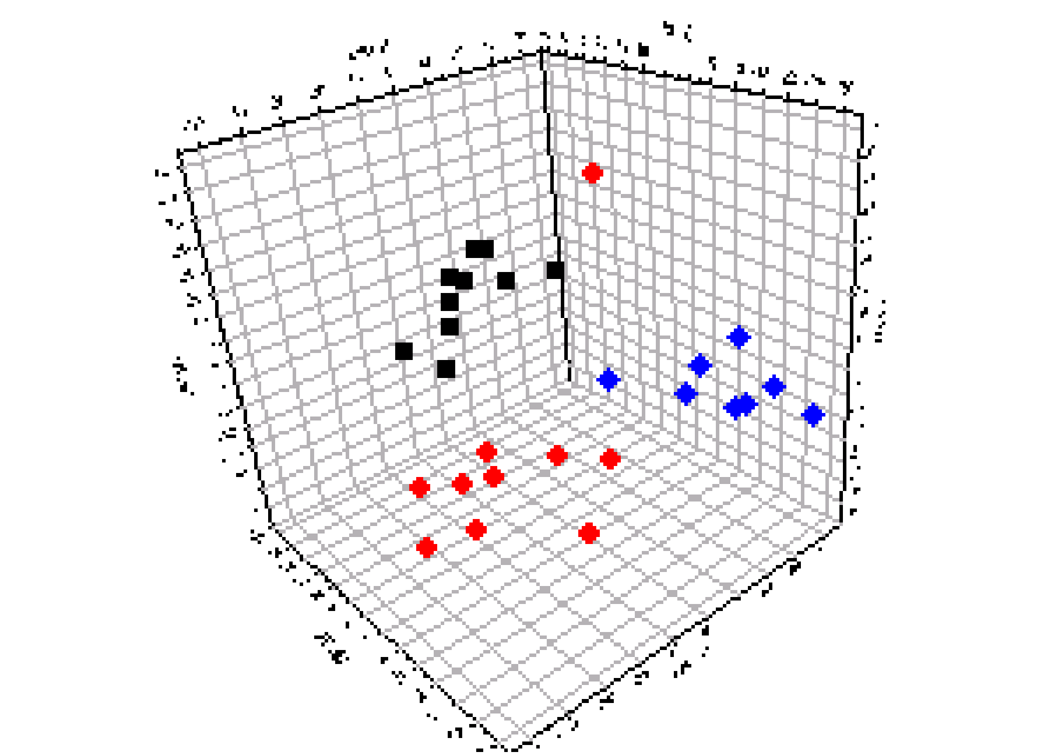
• Data set overview	PCA
• Identification	PCA or PLS
• Classification & Discriminant Analysis	PCA_Class or PLS-DA
• Variation (PC ANOVA)	PCA + ANOVA
• Relationships	PLS
• Dynamics	PLS, y=time, Batch PLS
• Cluster Analysis	in PC or PLS scores
• Visualization	**T** & **P** + **color** + connect
• Parsimonious models	sel-PLS
• Structure	Hierarchical models
• Expert Systems	Scores + DModX
• MV Design, …..	Design in scores

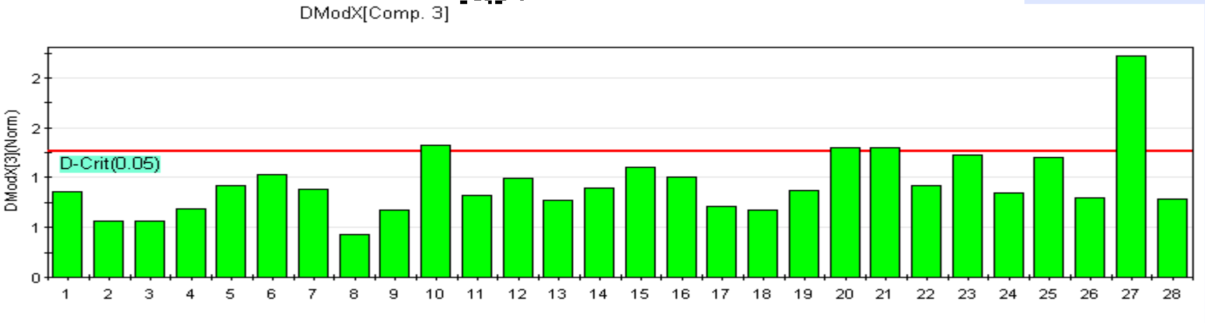
2. A more efficient class separation by PLS-DA

PLS-DA scores,
t_1 & t_2 & t_3
show a clear
separation
between the
three classes

Ctrl
S_chlorquine
S_amiodarone

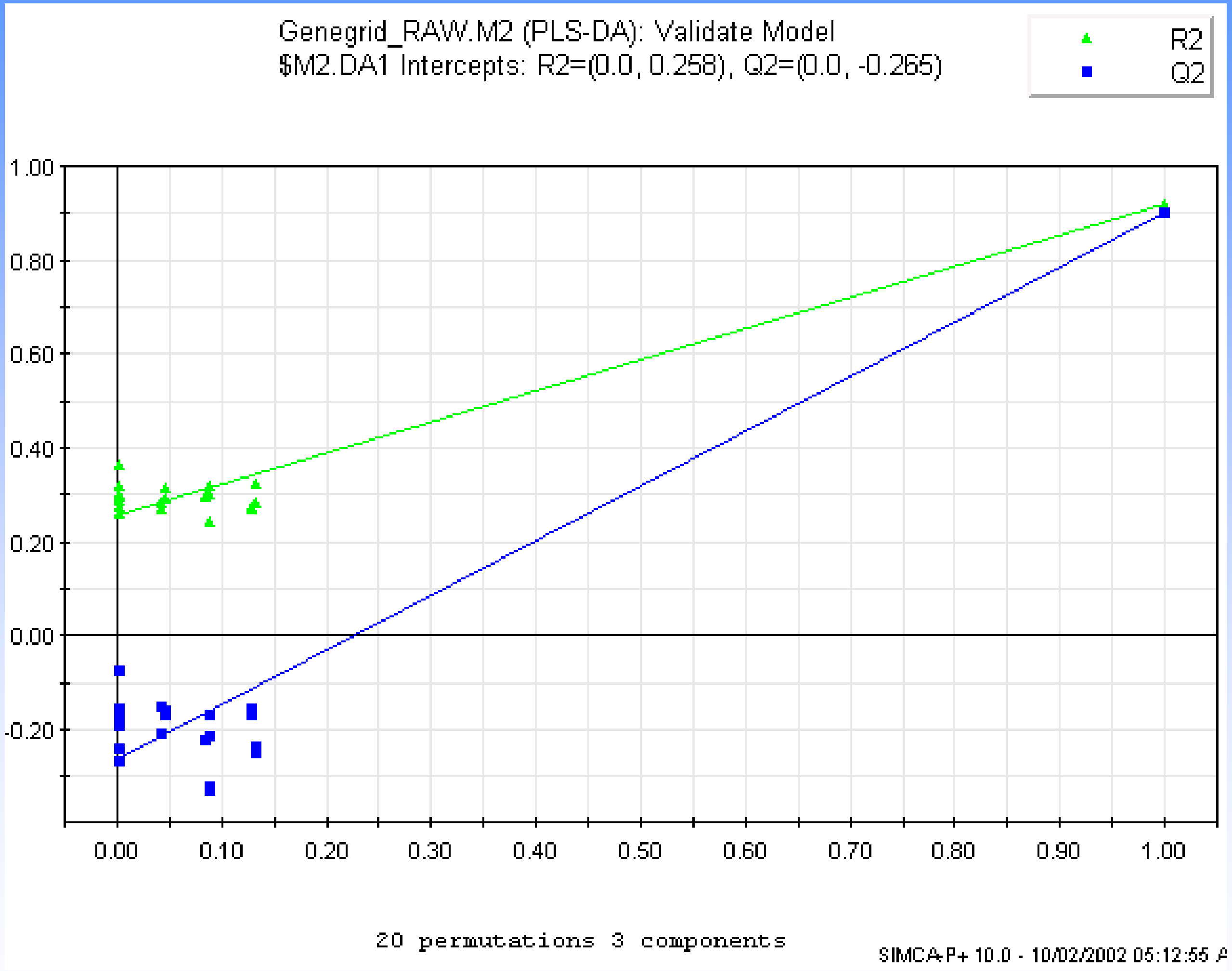
**# 27 is out,
also in
DModX
(lower plot)**



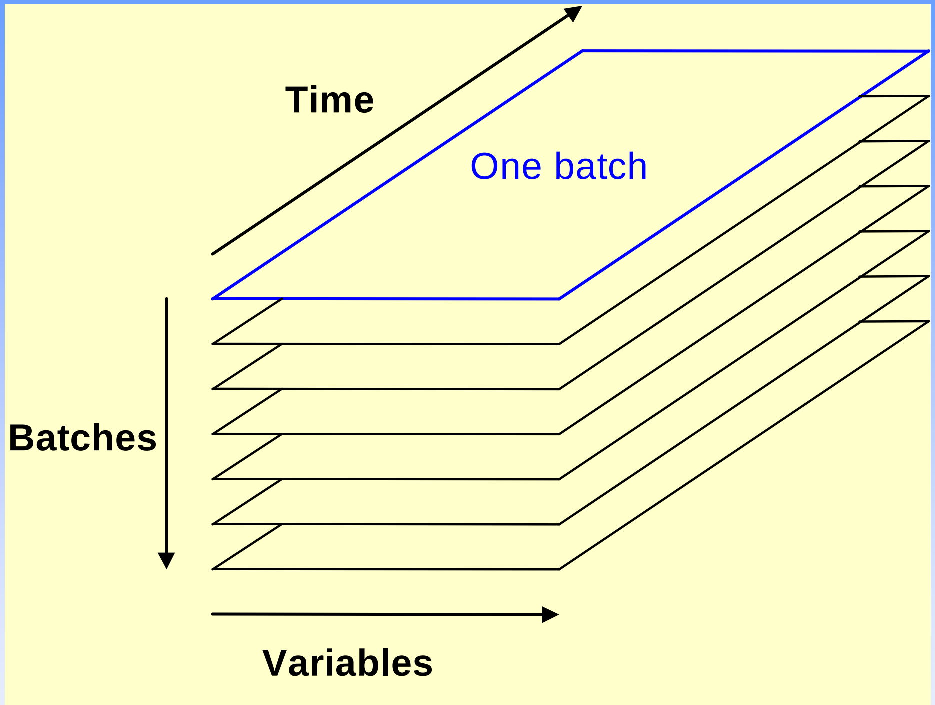


UMETRICS

(c) PLS-DA + permutation test



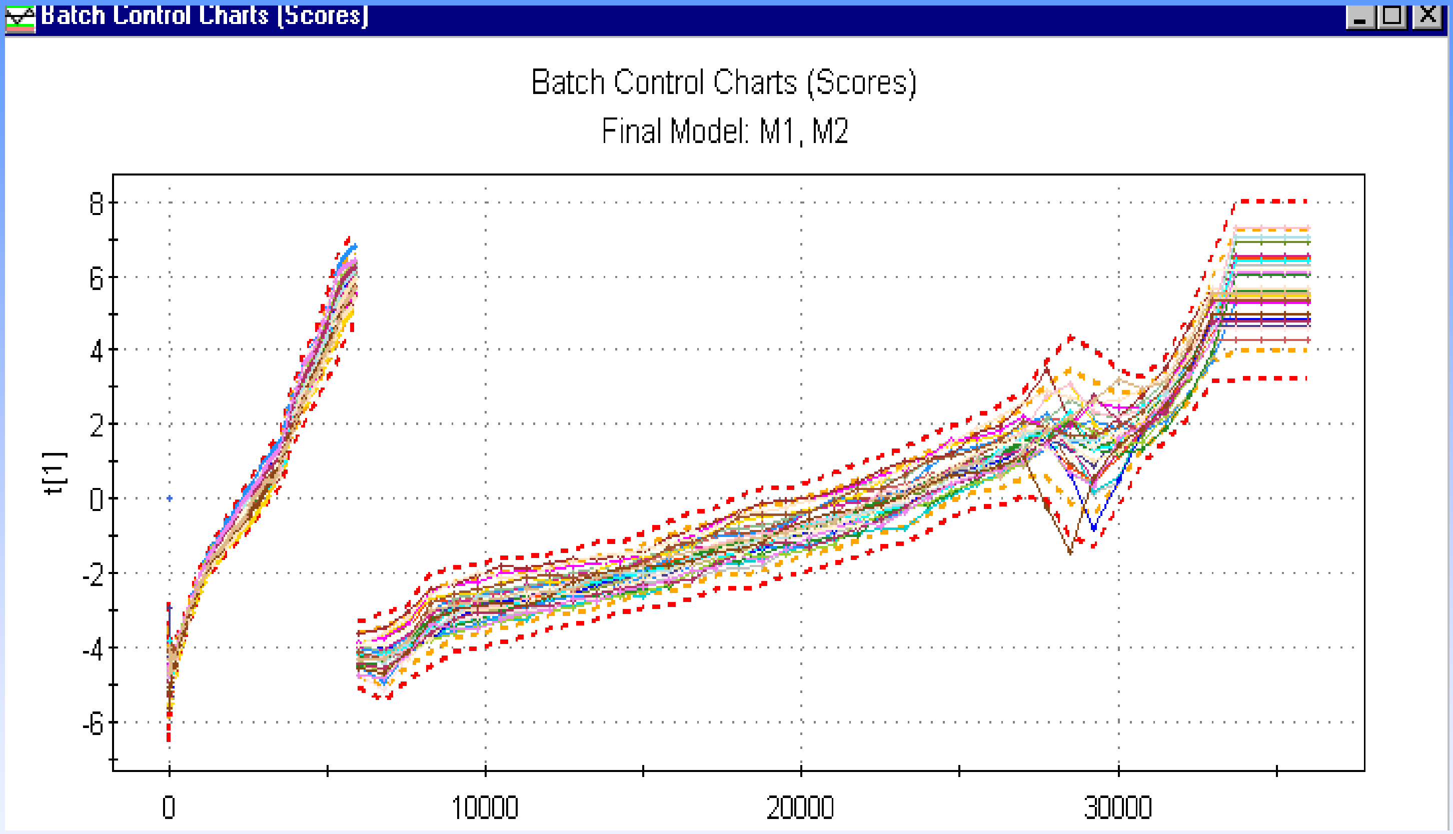
Nature of Batch Data, e.g., individuals evolving with time



- The data structure is a 3-way matrix
- Batches can have different lengths
- Additional tables with (for each batch)
  - – initial conditions
  - –quality measurements

- Multivariate batch analysis models the dynamic correlation structure(s) in the 3-way data
- Participating variables (coefficients, confidence intervals)
- Predictions
- Plots

Control Charts of score 1 (t1) vs. time (chip production, IBM Burlington)



Can address maturity concerns, etc.

Why multivariate projections (PCA & PLS & extensions)

- Based on all data
- Dimensionality problem
  - –can handle 1000’s of variables
  - –also K >> N
- Collinearities
- Missing data
- Noise in X and Y
- Models X, Y, and X ⇒ Y
- Graphical representation
  - –score plots of X, Y, & X ⇒ Y
  - –loading plots

The three basic applications

- Overview, Summary (PCA)
  - –maps
  - –trends, patterns, clusters
- Classification (Simca, PLS-DA)
  - –resolution of classes
  - –relevant variables
- Relationships X ↔ Y (PLS)
  - –interpretation
  - –predictions x → y
  - –optimization, y → x

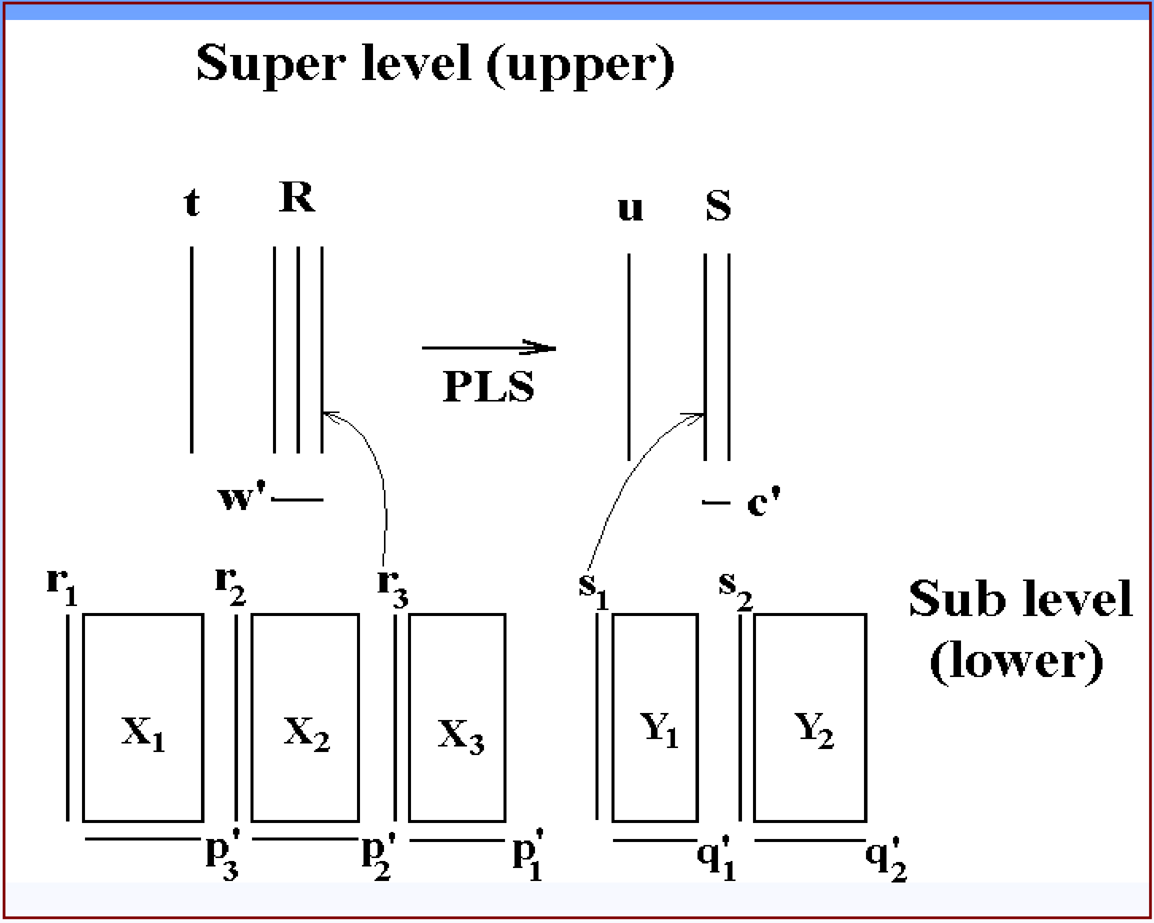
Some recent developments in chemometrics

- Hierarchical models (H-PCA and H-PLS)
  - Variables divided into meaningful blocks, that are modelled separately
  - The block scores (optimal summaries) are used as new variables on a higher level in the hierarchical model
  - Facilitates interpretation, lets us deal with very many variables
  - Analogous to clustering but of variables instead of observations (cases, samples)
- Orthogonal signal correction in PLS (Wold et al., 1998)
  - Filtering X data from secondary variation that is unrelated to Y
  - OPLS, O2PLS; Trygg, 2001- 2002
- Multivariate Batch modeling
  - Dynamics of batches (beer brewing, fermentation, patient data over time)

The block scores are variables in the “super” model

Many variants:

- No Y’s (hier PCA)
- Few Y’s; (H-PLS)
  Y unblocked
- Few X’s; (H-PLS)
  X unblocked
- Many X’s and Y’s
  X and Y blocked
  (H-PLS)



UMETRICS

MVA in Metabonomics - Give your ideas a fair chance !

- Much Data, especially in numbers of variables
- Possibilities
  - Overview, Classification, Relationships, Variation, Dynamics, …
- Types of results -- optimal summaries + deviations
  - Similarities, Dissimilarities between objects (samples, molecules, ...)
  - Relationships
  - Outliers
  - Variables related to these patterns
  - Feedback, Predictions
- The basis of Knowledge;
  - Representative cases (Design). ***Do NOT change one factor at a time***
  - Informative variables (Insight).
  - Adequate Analysis (Not one thing at a time).
  - **Understandable representation** of results, relationships, etc. **MODELS & PLOTS**
- Conclusions – what we can do, and what we can NOT do

Some references

- H.Martens and T.Naes. Multivariate Calibration. Wiley, N.Y., 1989.
- J.E. Jackson. A User's guide to principal components. Wiley, N.Y., 1991.
- L.Eriksson et al., Introduction to Multi and Megavariate Analysis, Umetrics 2000
- Nicholson, Holmes, Antti et al.
- WWW.umetrics.com
  - and links to Chemometrics Home Page,
  - Rasmus Bro's reference base
  - Umeå Univ. Chemometrics group
  - NAmICS (N. Amer. Ch. Int. Chemom. Soc)
- Chemometrics and Intell. Lab. Syst. (Elsevier),
- J. Chemometrics (Wiley)
- J.Med.Chem, QSAR, ….
- QSAR society

One last comment:

CHAMPS: CHemometrics Applied to Metabonomics, Proteomics & Systeomics, Sept 2004, Malmö, Sweden.
More info: anna@chemsoc.se

The End

Thanks for your attention