Cross model validation and multiple testing in latent variable models

Frank Westad
GE Healthcare
Oslo, Norway

2nd European User Meeting on Multivariate Analysis

Como, June 22, 2006

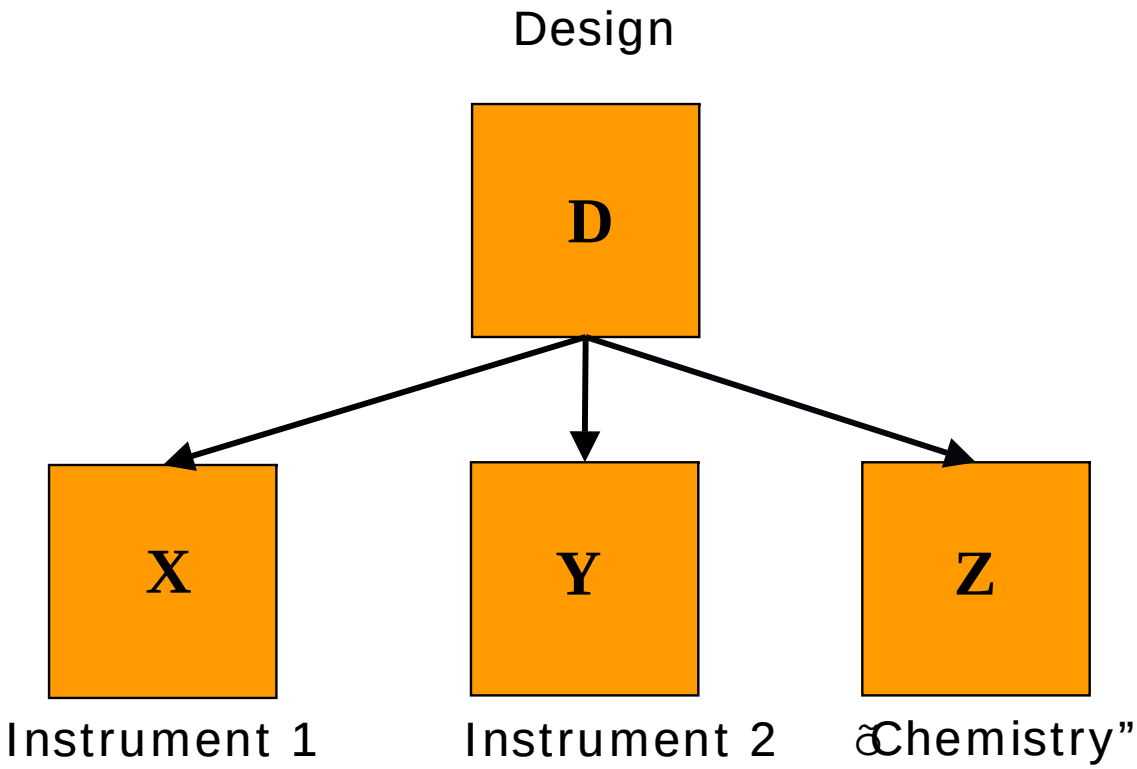
Outline

- Introduction
- Uncertainty estimates
- Cross model validation (CMV)
- Multiple testing
- Applications
- A useful trick: Passifying variables
- Summary

Introduction

- More and more applications involve “hundreds” of variables
- Examples:
  - Designed experiments:
    - X is some DoE, Y = multichannel (e.g. microarray)
    - D is some DoE, X and Y are multichannel data (instrument 1 and 2)
  - Spectroscopy (X = FT-IR, Y = proteins from 2D- gel electrophoresis)
  - Multiway data
- The objective is often to find the significant variables for interpretation, not necessarily for prediction
- How to solve the issue of multiple testing?
- How to visualise the design’s impact on the instrumental data?

A general data structure
- An experimental design is basis for the study



Uncertainty estimates from resampling

- Objectives
  - To estimate uncertainties in the model parameters
  - Reflect the *actual* data structure (outliers, skewness)

- Some approaches for estimation
  - Jackknifing/Cross validation
  - Bootstrapping from original data (unconditional)
  - Bootstrapping from model (Y = XB + F; conditional)

Pioneering work by Efron and Tibshirani

"Re-discovered" several times

Uncertainty estimates

The variance of the model parameters can be estimated by jack-knifing

Example: Regression coefficients, *b*

$$s^2b = \left(\sum_{m=1}^{M} (b - b_m)^2 \right) \left((M-1)/M \right)$$

Main model

Sub-model

M = the number of segments

$s^2(b)$ = estimated uncertainty (variance) of *b*

b = the regression coefficient at the cross validated Aopt components using all *N* objects

b_m = the regression coefficient at the rank Aopt using all objects except the object(s) left out in cross validation segment *m*.

A t-test can be performed to find significant variables

Partial Least Squares Regression (PLSR)

The structure model is:

$$X = TP^T + E_A$$

$$Y = TQ^T + F_A$$

X = Predictor variables
Y = Response variables
T = Score matrix
P = Loadings matrix for X
Q = Loadings matrix for Y
E_A = X-residual matrix
F_A = Y-residual matrix
W = max(covariance(E_A,F_A))

May also estimate uncertainty
For W, P and Q

$$B = W(P^TW)^{-1}Q^T$$

As a linear regression model:

$$Y = XB + F$$

How can we know if the estimates are "correct"?

- Choose data where we have the "truth"
  - Example: Factorial design or central composite design
  - ANOVA-PLSR as additional method to traditional ANOVA
  - Only 1 PLSR component when there is one response variable
  - Compare ANOVA with PLSR for jack-knife and bootstrap
- NB! The objective is not to replace ANOVA with PLSR for basic Design of Experiments (DoE)!

Data “Helicopter”
Box *et al* - Experimental design

Students construct paper-helicopters with different dimensions to estimate effect on “flying time”. Experiments performed in two blocks.

Results

Variable	ANOVA	JK PC1	BS PC1
Block	0.613	0.671	0.583
Wing Area	0.960	0.970	0.962
Wing ratio	0.005	0.024	0.004
Body Width	0.880	0.913	0.888
Body Length	0.001	0.008	0.001

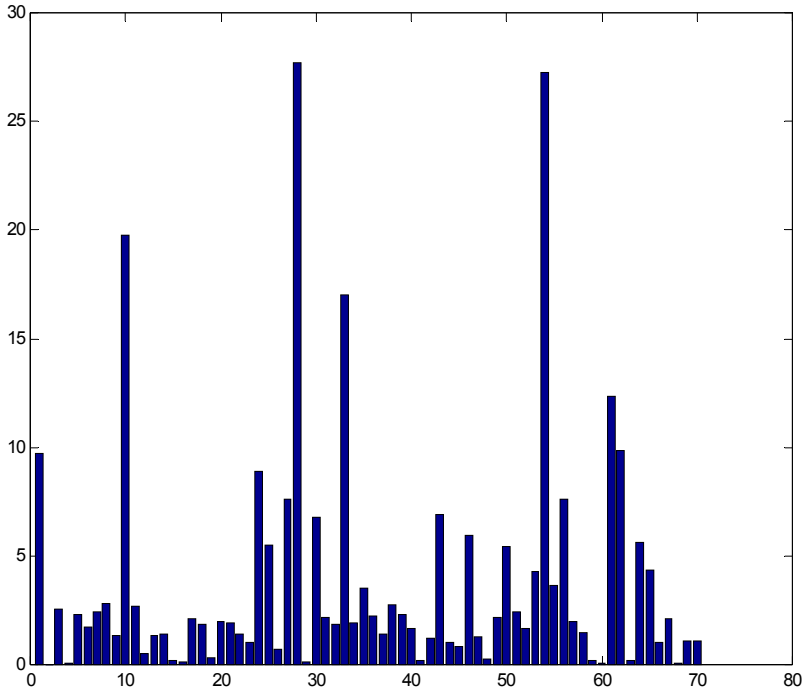
Example 2 – The “truth” is not known

- 70 genetic markers for 234 samples of a weed in Ethiopia
  - Factors:
    - Geographic regions
    - Climatic zones
    - Monthly temperature
    - Precipitation
    - Longitude and latitude
    - Altitude
- Model objective (on of many)
  - What genes changes with altitude?
- Variable selection to find significant (important) variables
  - Genetic algorithm (Riccardo Leardi’s implementation)
  - Estimate uncertainty by jack-knifing
  - Estimate uncertainty by unconditional bootstrapping

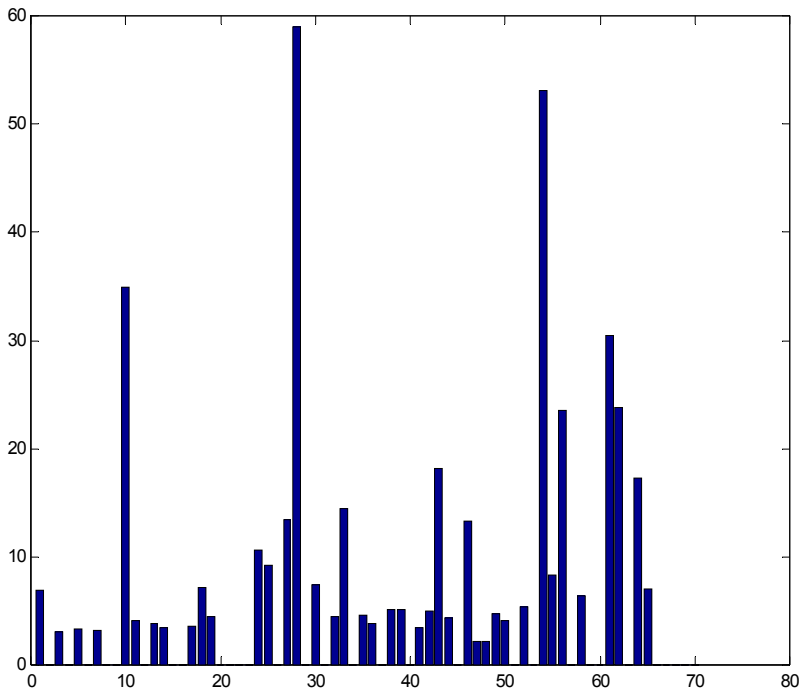
Results – Genetic data

Compare by plotting -log(p(b_k))
The 10 most significant variables are the same

Jack-knife (20 significant)

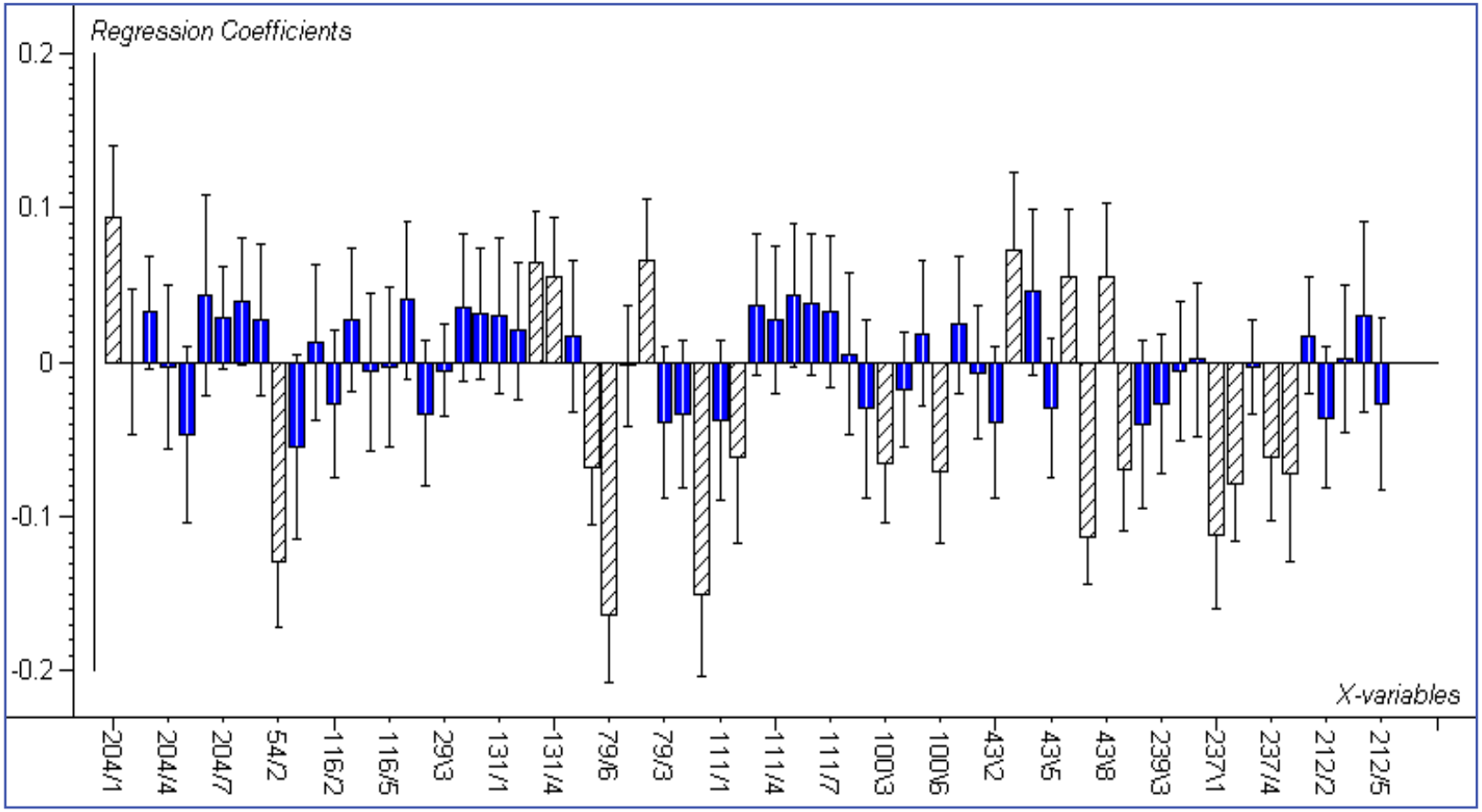


Bootstrap (38 significant)



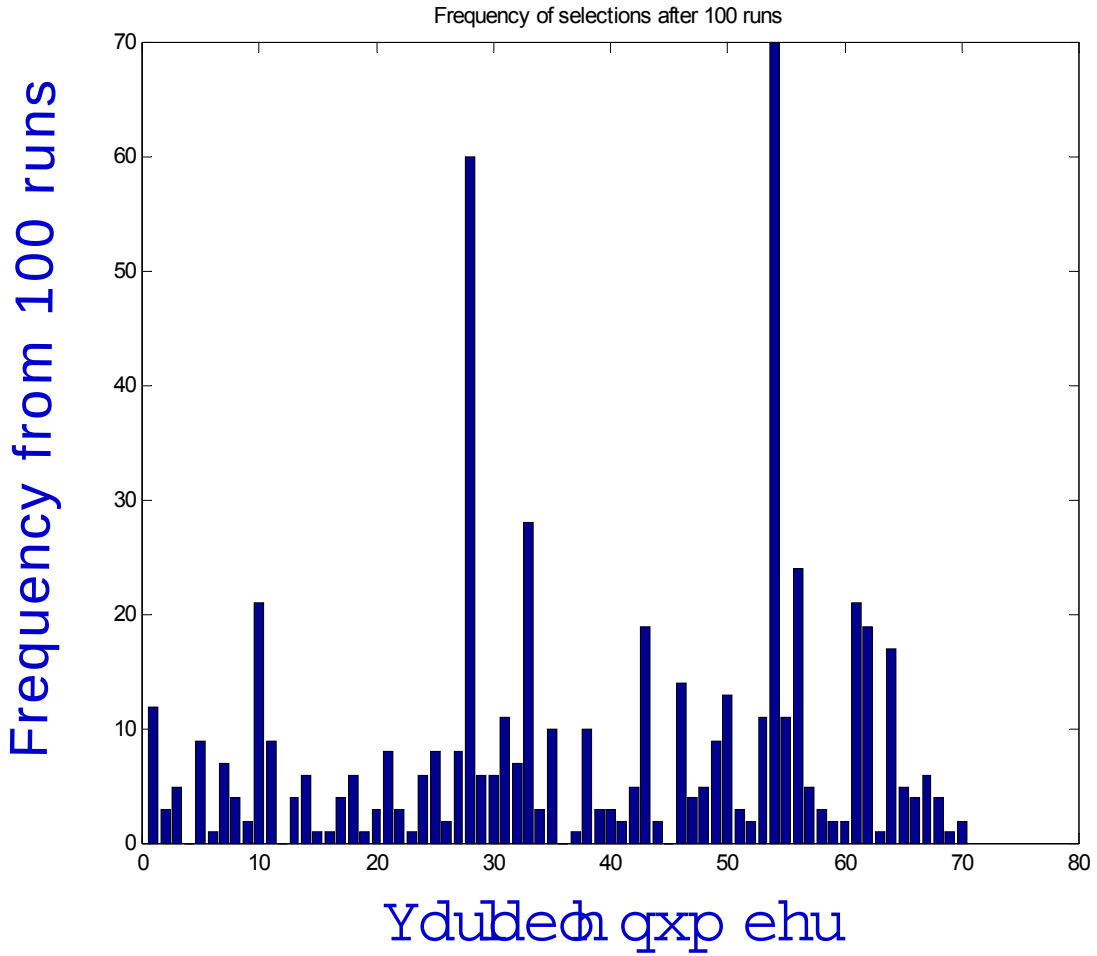
Regression coefficients, A = 3

Shown with “95% conf. interval”
20 significant genes marked (full CV)



imagination at work

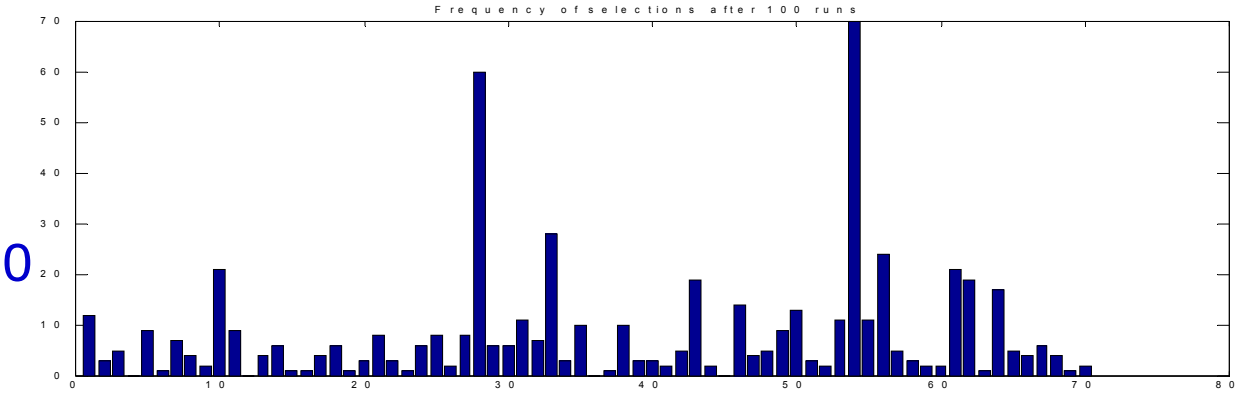
Selection frequency from GA



imagination at work

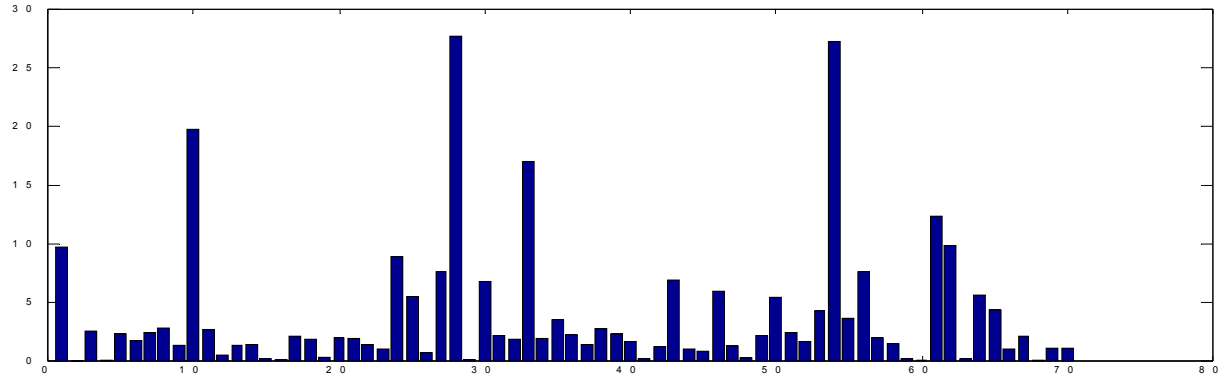
Comparison of JK PLSR og GA-PLSR

GA-PLSR:
Frequency from 100 runs



Variable number

JK PLSR:
-log(p(b))

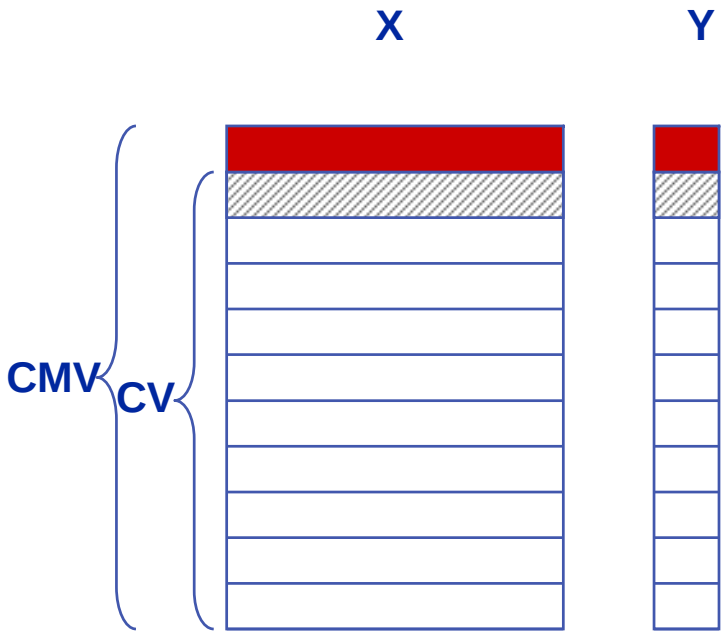


Variable number

imagination at work

Cross-model validation (CMV) with variable selection based on jack-knifing

0. Cross-validation on all objects
1. Take out e.g. 10% of the objects
2. Cross-validate the remaining
3. Find significant variables
4. Predict the objects that were kept out
5. Estimate RMSE (or explained variance)
6. Repeat 1 - 5 until all objects have been taken out
7. Show frequency of significance for all variables
8. Collect and predict an independent test-set!



Multiple testing - I

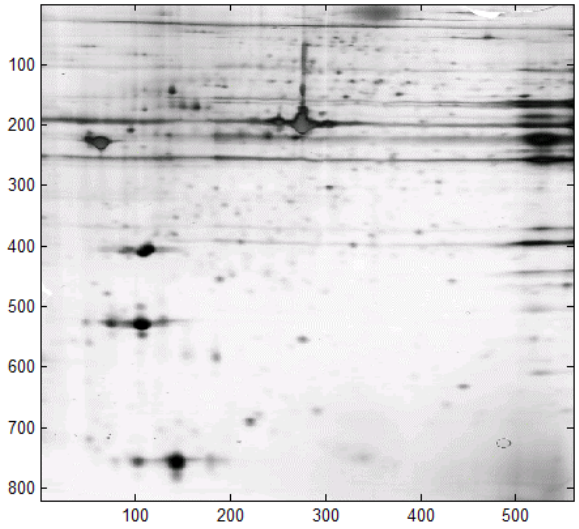
- Case 1: X = design, Y = multiple responses
- Case 2: X = genes, Y = binary (e.g. benign/malignant)
- Case 3: X = spectra, Y = genes
- Most work has been done in Case 1 (MANOVA) and Case 2 (individual t-tests)
- Issue: Type I *and* Type II error
- Many approaches to adjust for multiple tests:
  - Bonferroni (adjusted *p* = (number of variables)x*p*
  - False discovery rate (FDR) - Expected portion of Type I error among the rejected hypothesis
  - Family wise error rate (FWER) - Probability of at least one Type I error
  - Permutation/rotation tests

Multiple testing - II

- One strategy: Reduce the number of variables prior to the modeling
  - Employ some clustering on the variables (NB! Validation)
  - Use score vectors from some latent variable model
- How many tests are performed compared to testing individual variables?

Multiple testing - III

- 2D gel-electrophoresis data (20 samples)
- Scenario 1:
  - 800 spots are quantified and aligned
  - 100 spots are “relevant”
- Scenario 2: Changing threshold for detecting spots
  - 300 spots are quantified and aligned
  - 100 spots are relevant
- In an ideal world: Regardless of scenario 1 or 2, the 100 relevant spots should be found
- How will the various methods handle this situation?

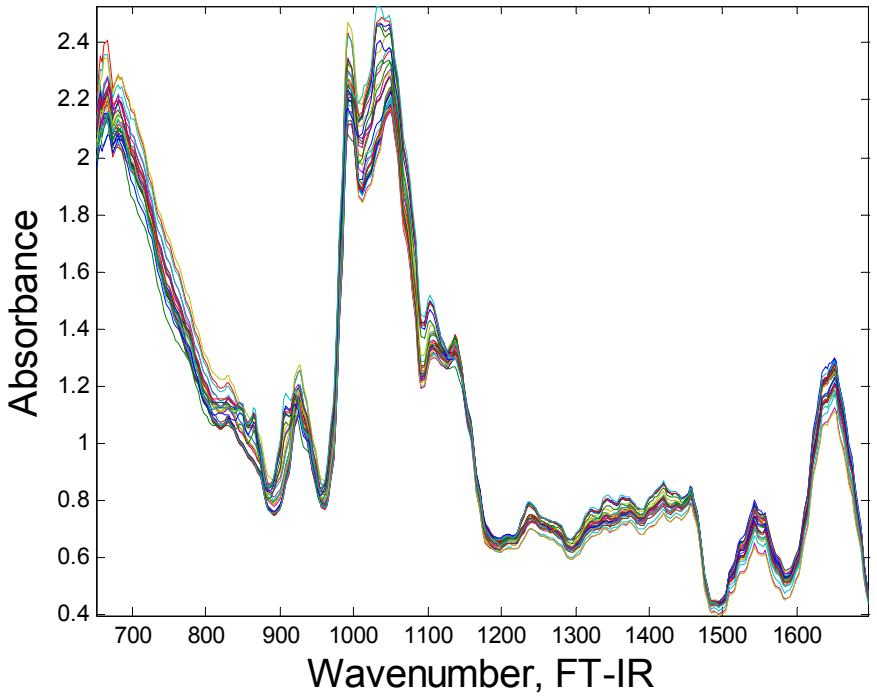


Cross model validation to assess significance in both X and Y –Example 1

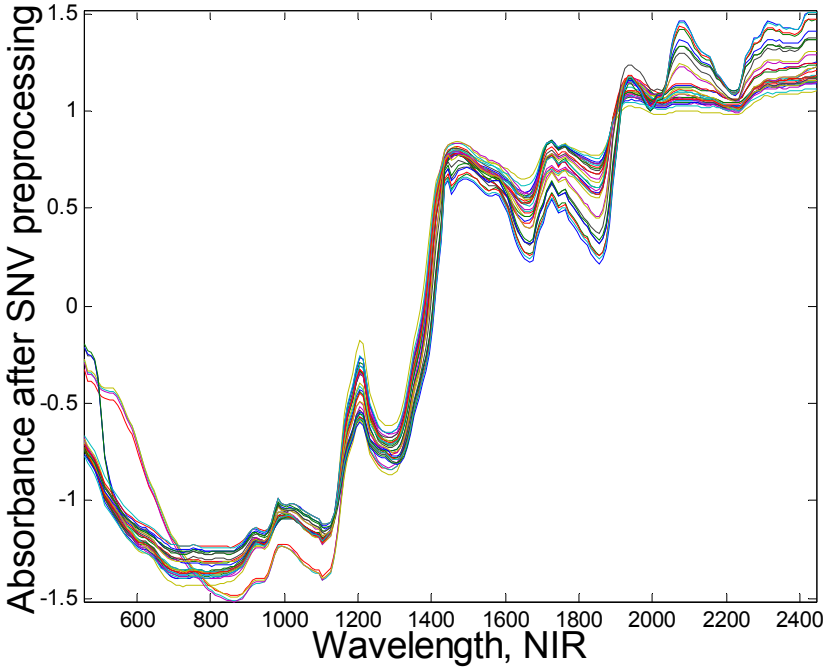
- 32 samples of marzipan
- VIS/NIR (400-2500 nm) and FT-IR (1700-600 cm^{-1}) spectroscopy
- Preprocessing the VIS/NIR spectra with signal normal variate (SNV; autoscaling objectwise)
- 16 random validation segments
- PLS regression with cross model validation
- Purpose:
  - Illustrate cross-model validation
  - Investigate how many variables that are regarded as significant in a permutation test

Spectral data

FT-IR spectra

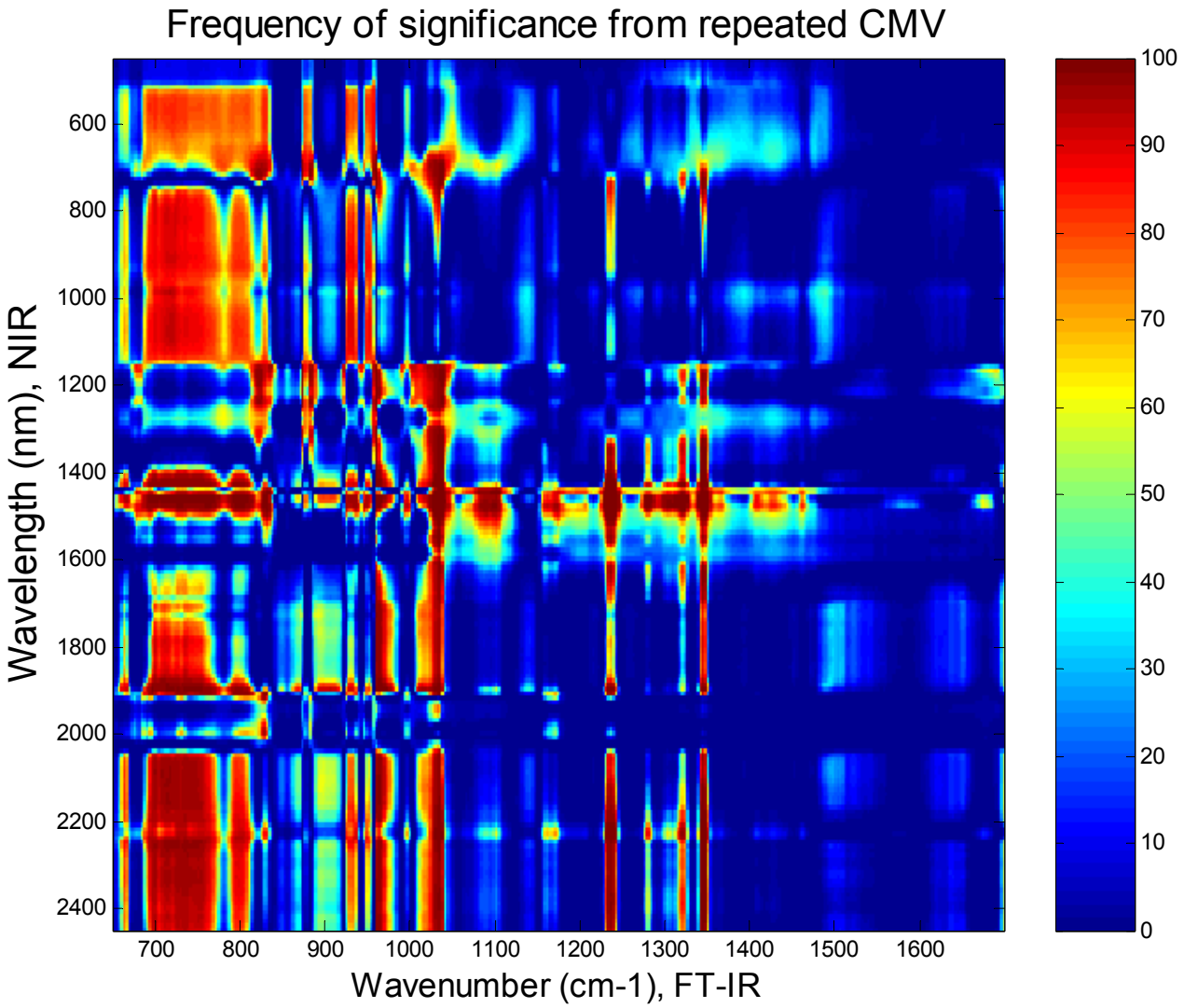


NIR spectra
SNV transformed



imagination at work

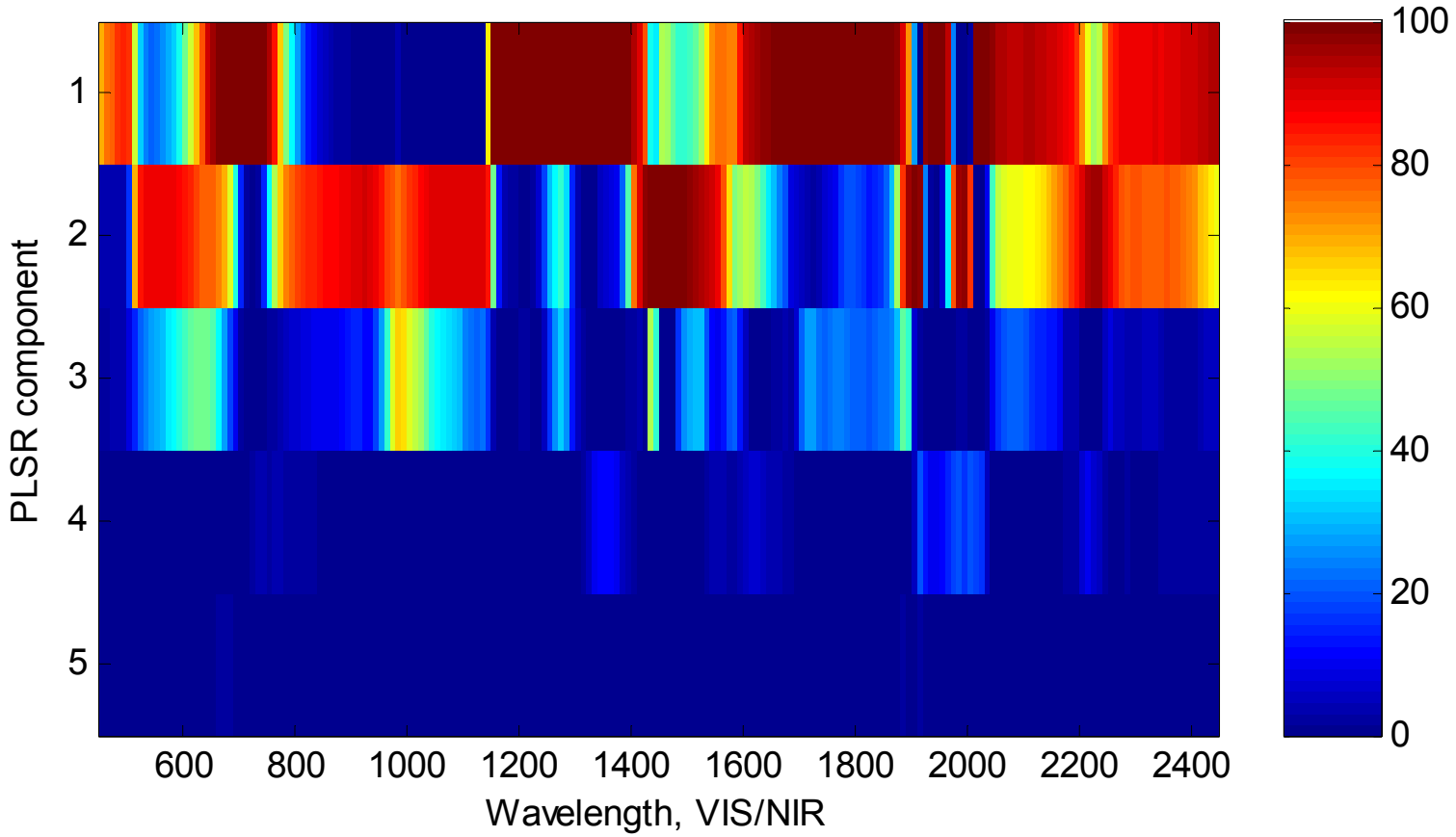
Map of significance, regression coeff. matrix



Map of significance, y-loadings

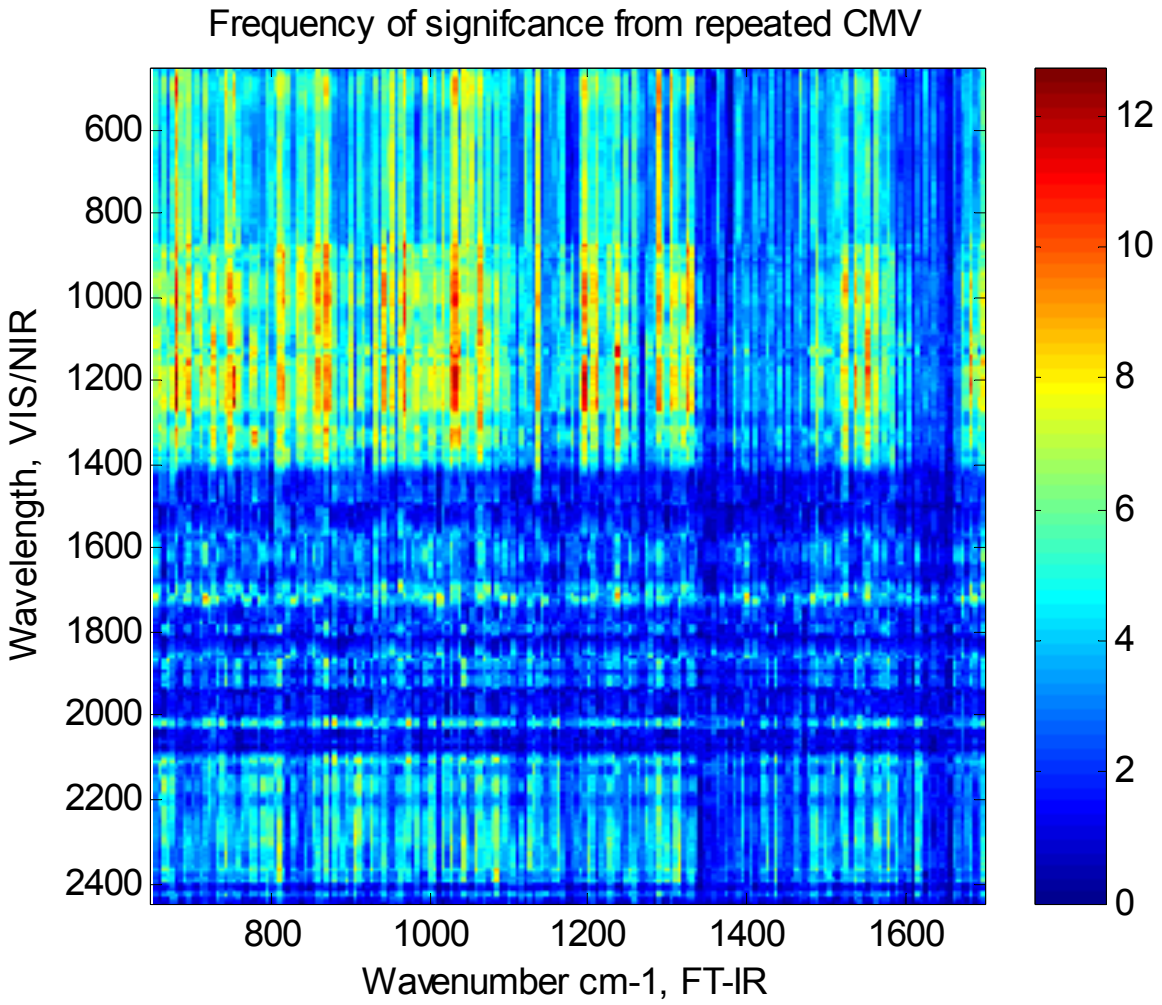
Significance for individual y-variables

Useful as an additional approach to decide on the model dimensionality



imagination at work

Map of significance, regression coeff. Matrix
Permutation test: Y-data randomly shuffled



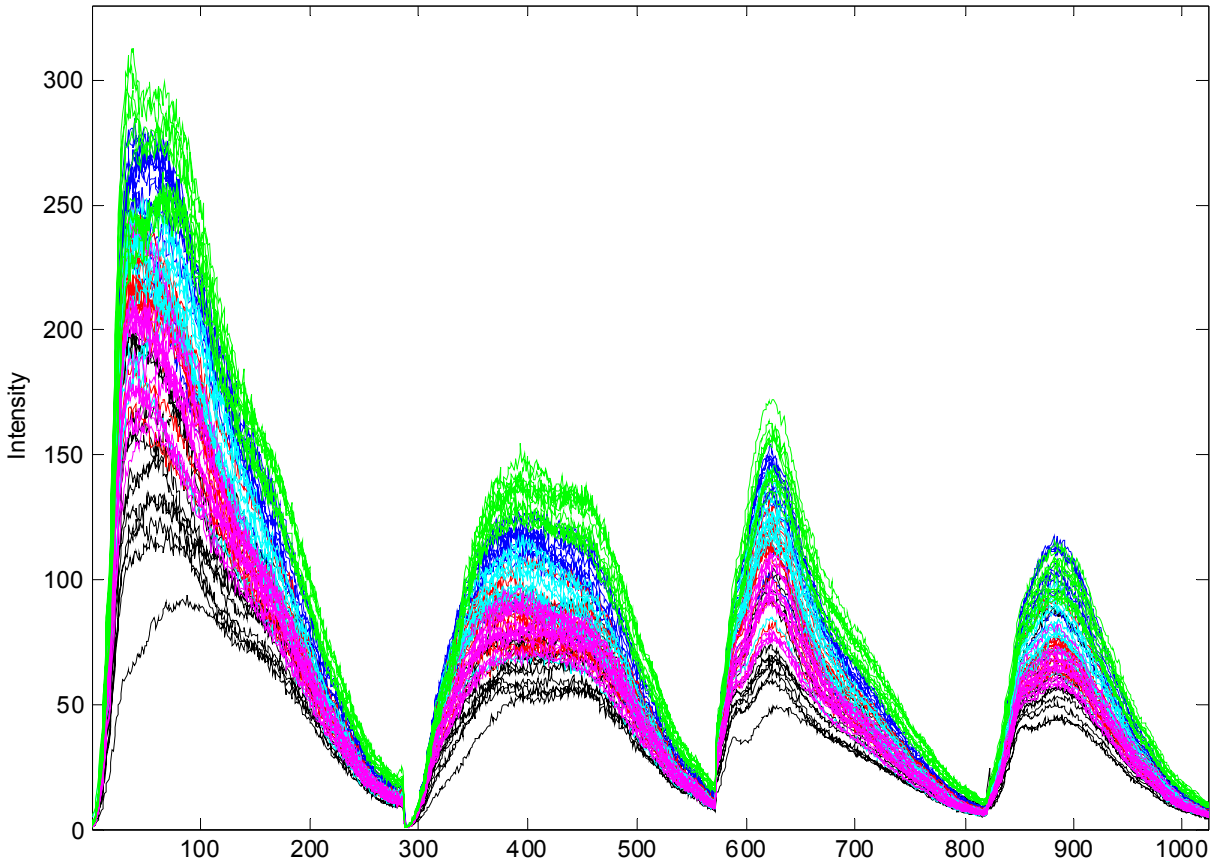
No variable > “12% significant” although performing 70000 individual t-tests

imagination at work

Example 2: Fluorescence spectra of sugar

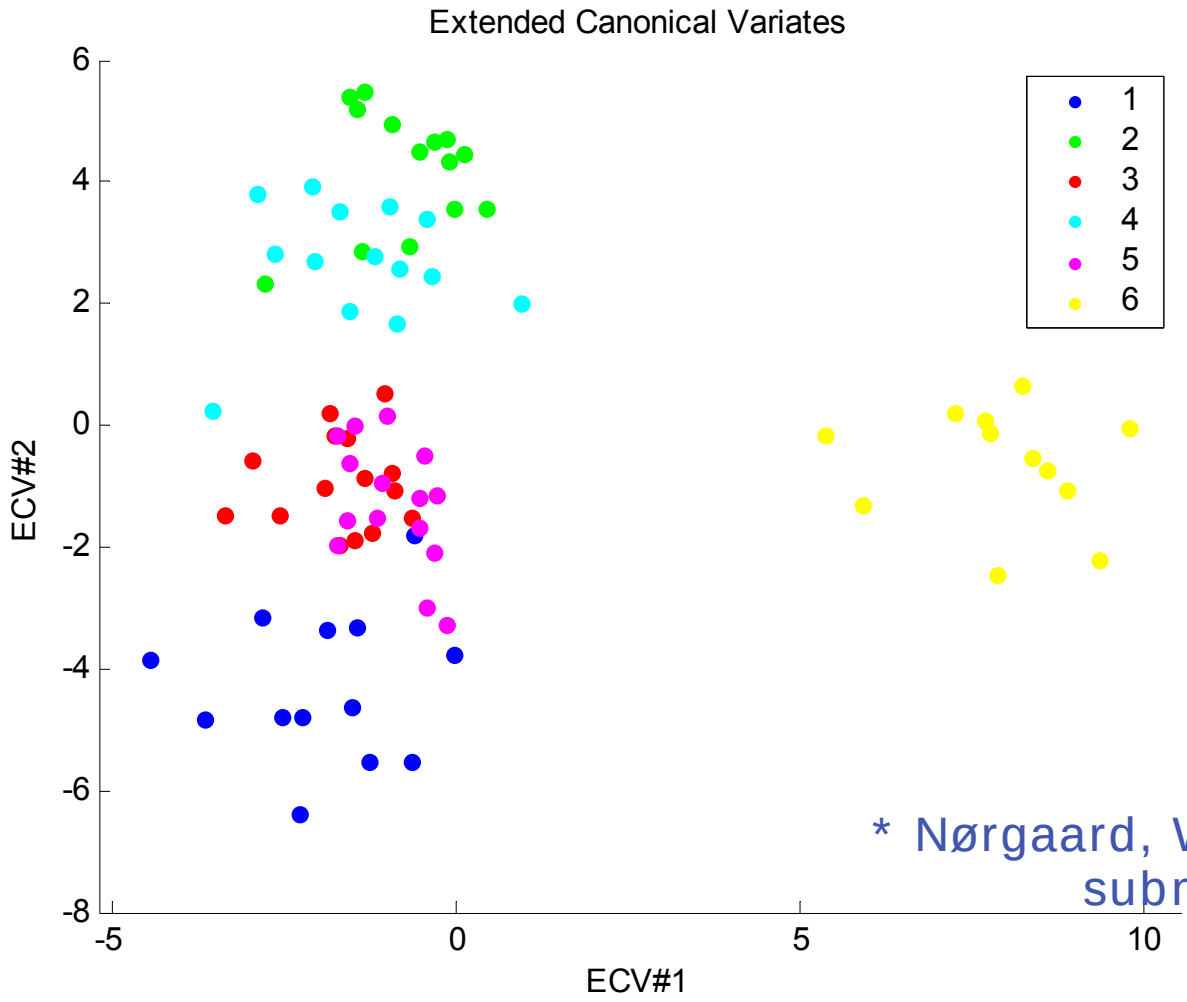
- Spectra from 6 different factories (83 samples)
- Four excitation wavelengths (230, 240, 290 and 340 nm)
- Emission (275-560 nm; 275-560 nm; 311-560 nm; 361-560 nm)
- The spectra are concatenated (1083 variables)
- Various classification methods are employed with factories as classes
- Full cross-validation

Fluorescence spectra of sugar
Just the data…



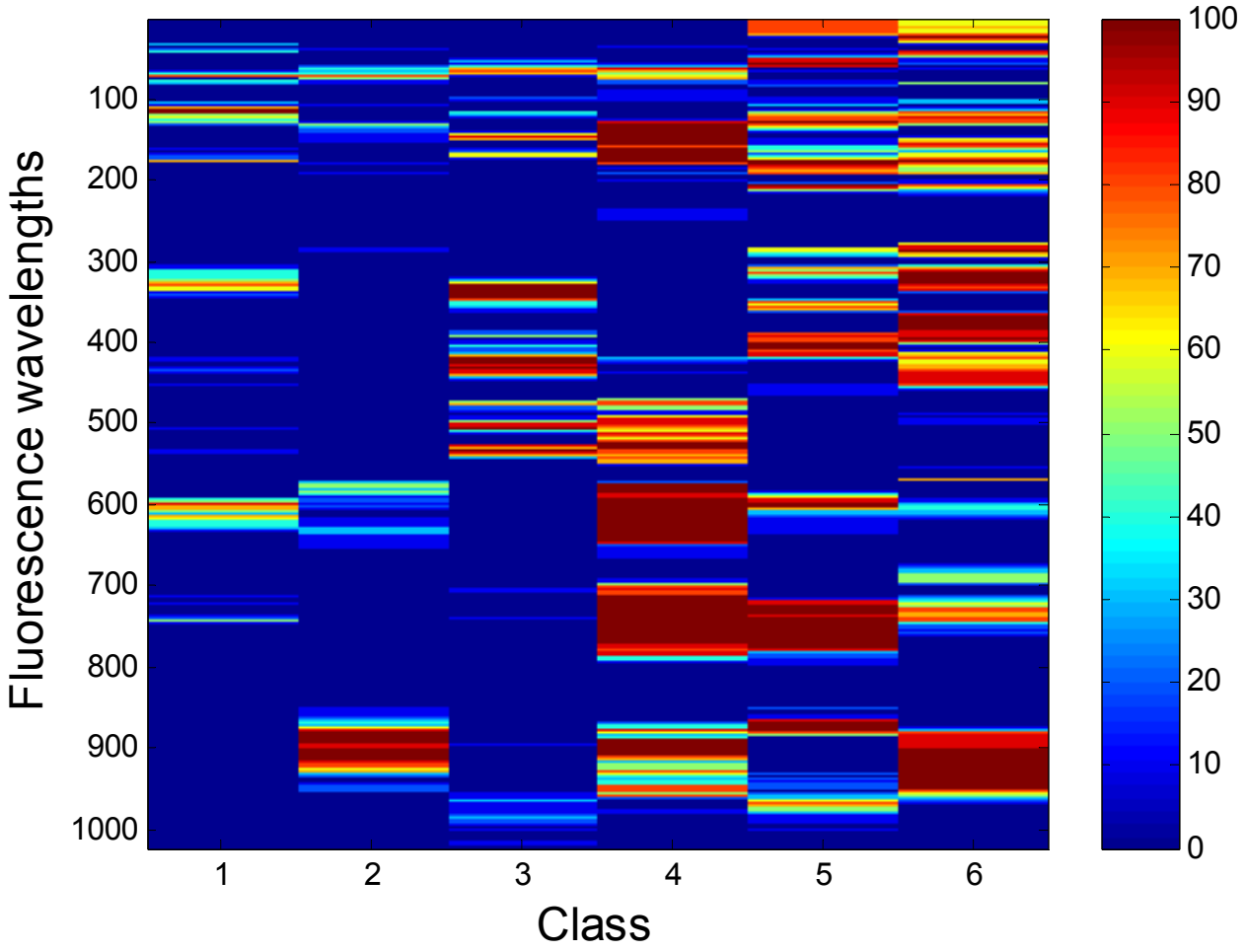
imagination at work

Fluorescence spectra of sugar
Extended Canonical Variates*



* Nørgaard, Westad, & Bro; submitted

Fluorescence spectra of sugar Significance from CMV-PLSR



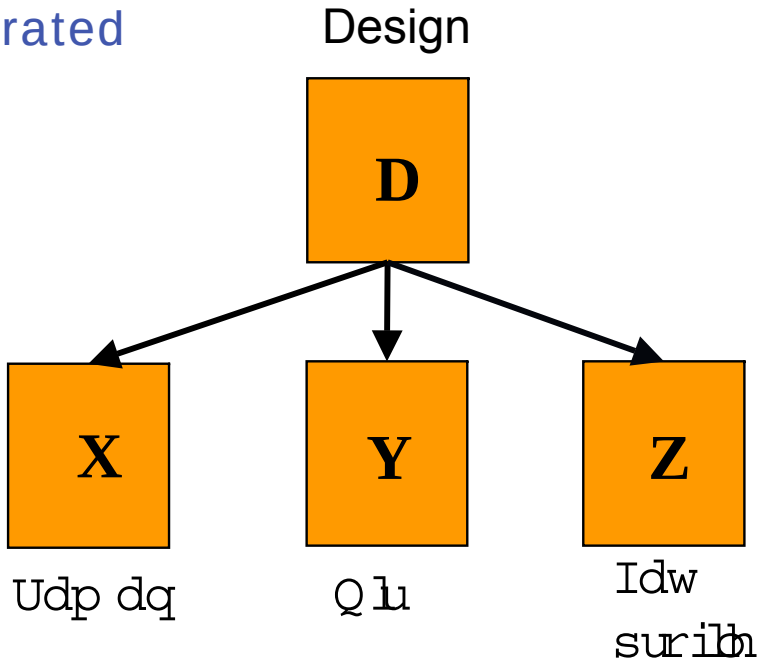
Passifying variables for a total interpretation

- Goal: Predict quality from spectroscopy
- Visualise the underlying design
- Downweigh variables that do not contribute to the prediction ability
  - Equivalent to computing correlation to the score vectors
  - Examples:
    - Design variables
    - Unimportant spectral regions
    - Observed process variables

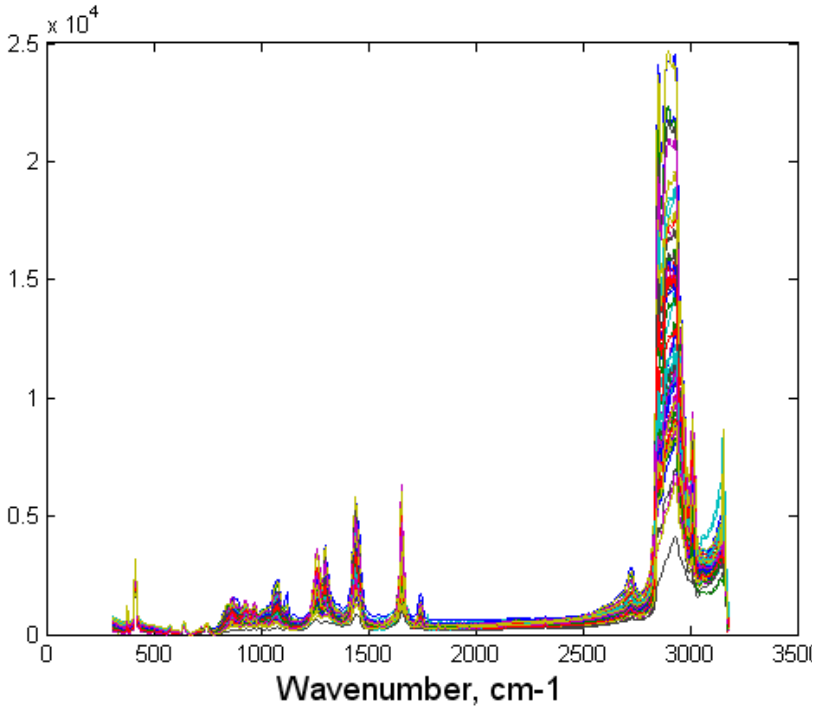
Example 3: Raman and NIR spectroscopy

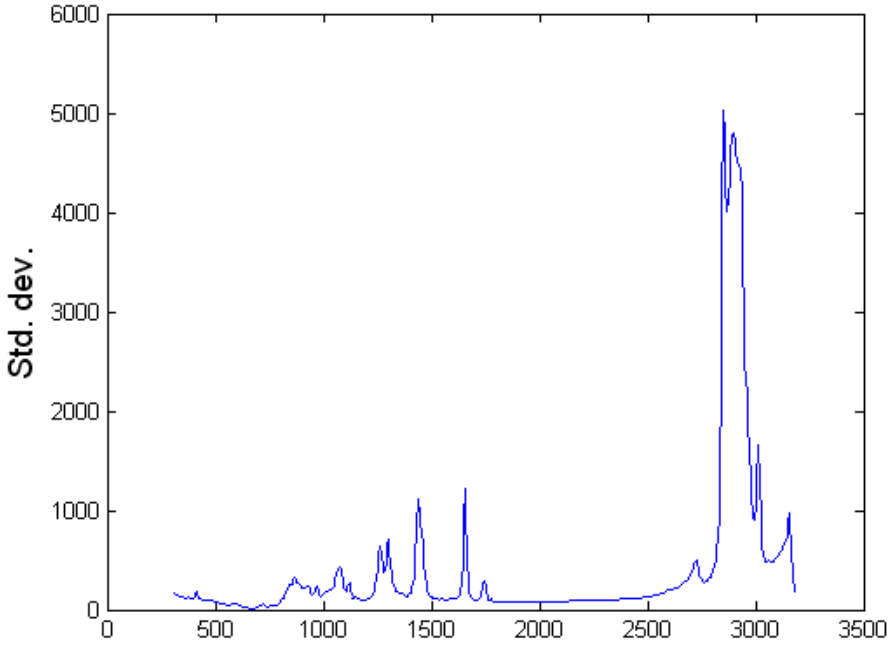
- Data:
  - Experimental design (double mixture) – 69 unique samples
    - Protein, fat and water
    - 5 sources of fat in a mixture (salmon, cod, soy, coca, olive)
  - Raman spectra, 310-3181 cm $^{-1}$; 575 variables
  - VIS/NIR spectra, 404 – 2494 nm; 210 variables
  - Fat profile (C12 – 22) + % saturated

- PLS regression with repeated cross-model validation; 9 segments
- Model 1: X =NIR , Y = Raman; Aopt = 4
- Estimating uncertainty for regression coefficients and y-loadings



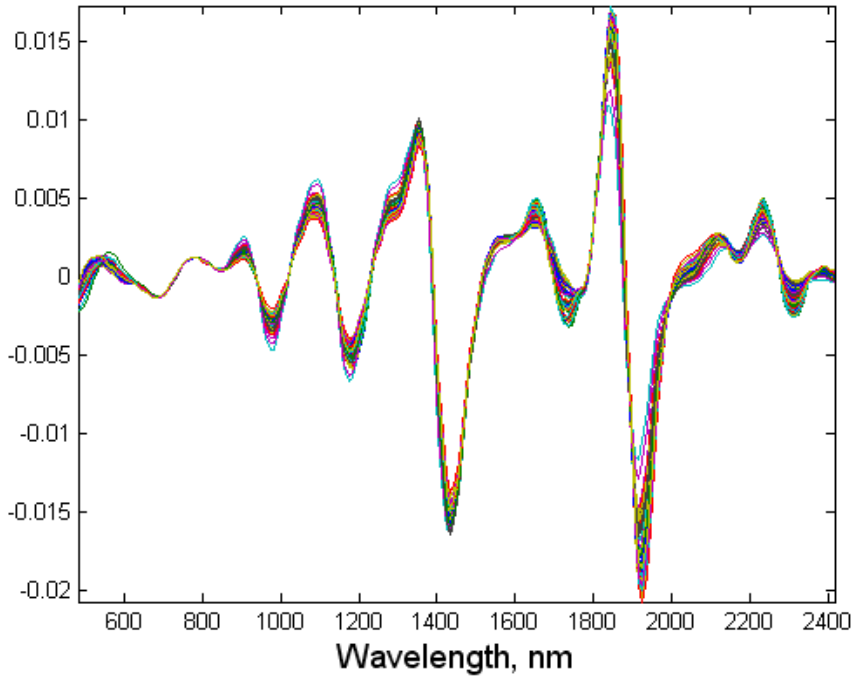
Raman spectra and standard deviation

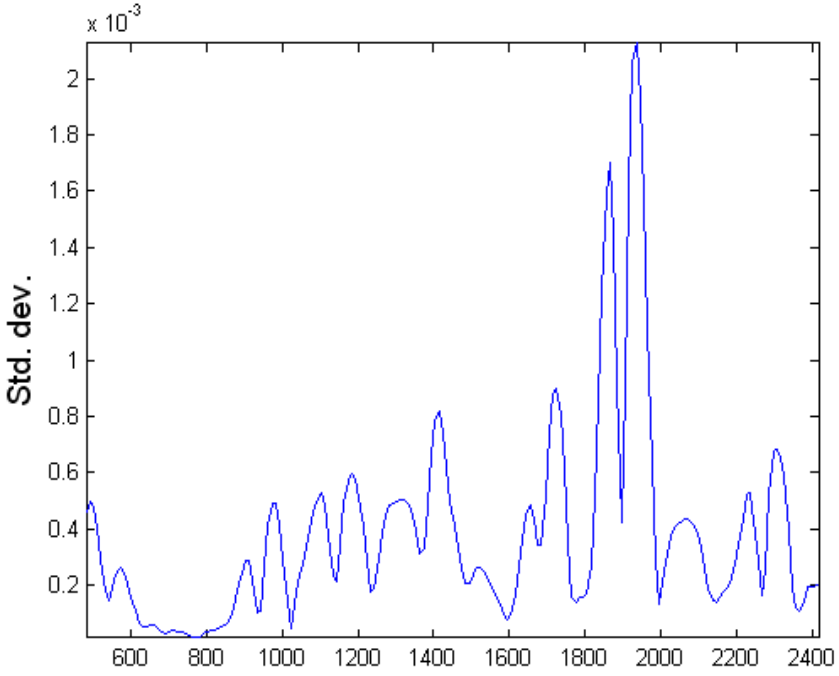




imagination at work

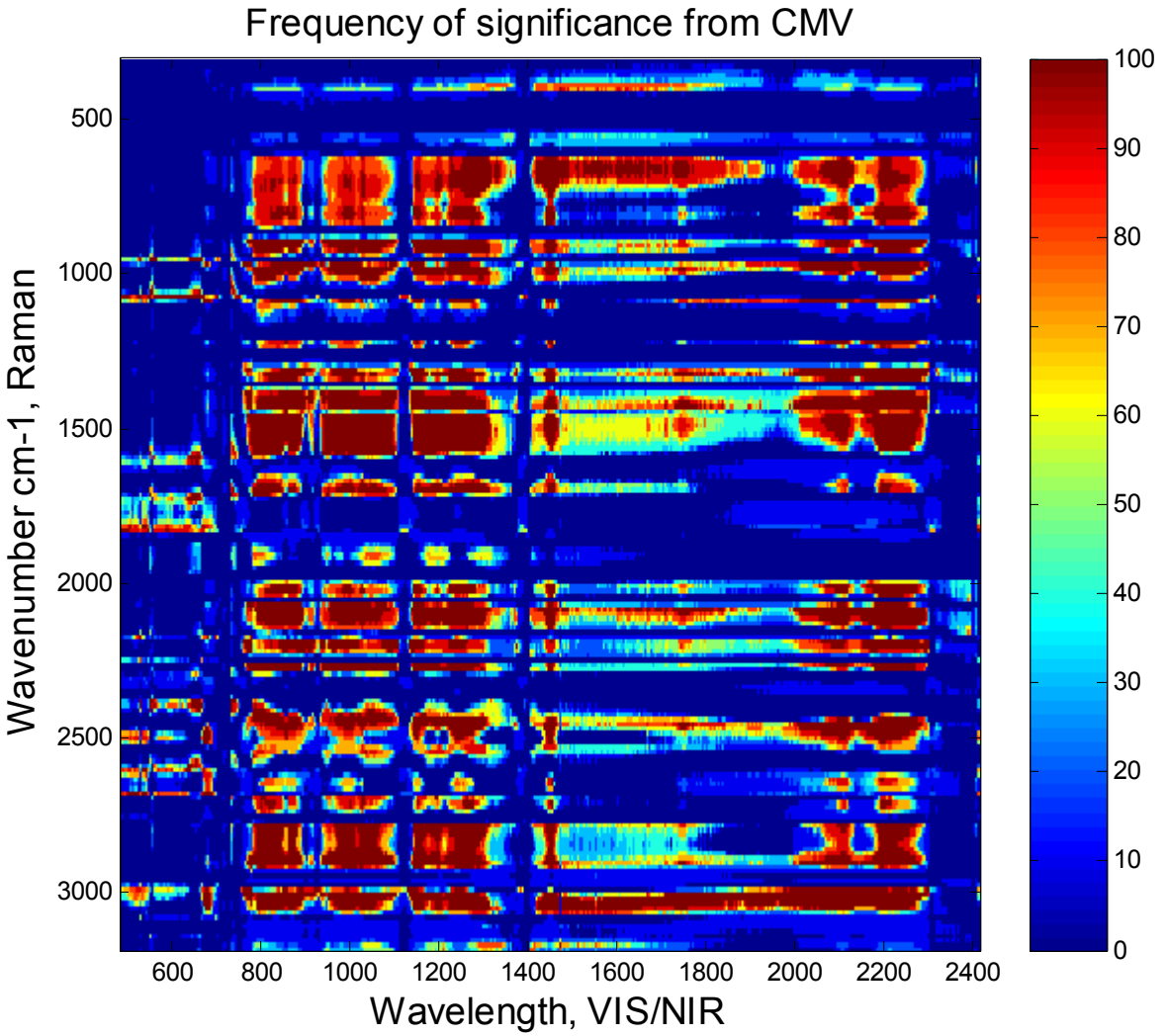
VIS/NIR spectra, 2nd derivative and standard deviation





imagination at work

Frequency of significance from CMV, B-coeff.

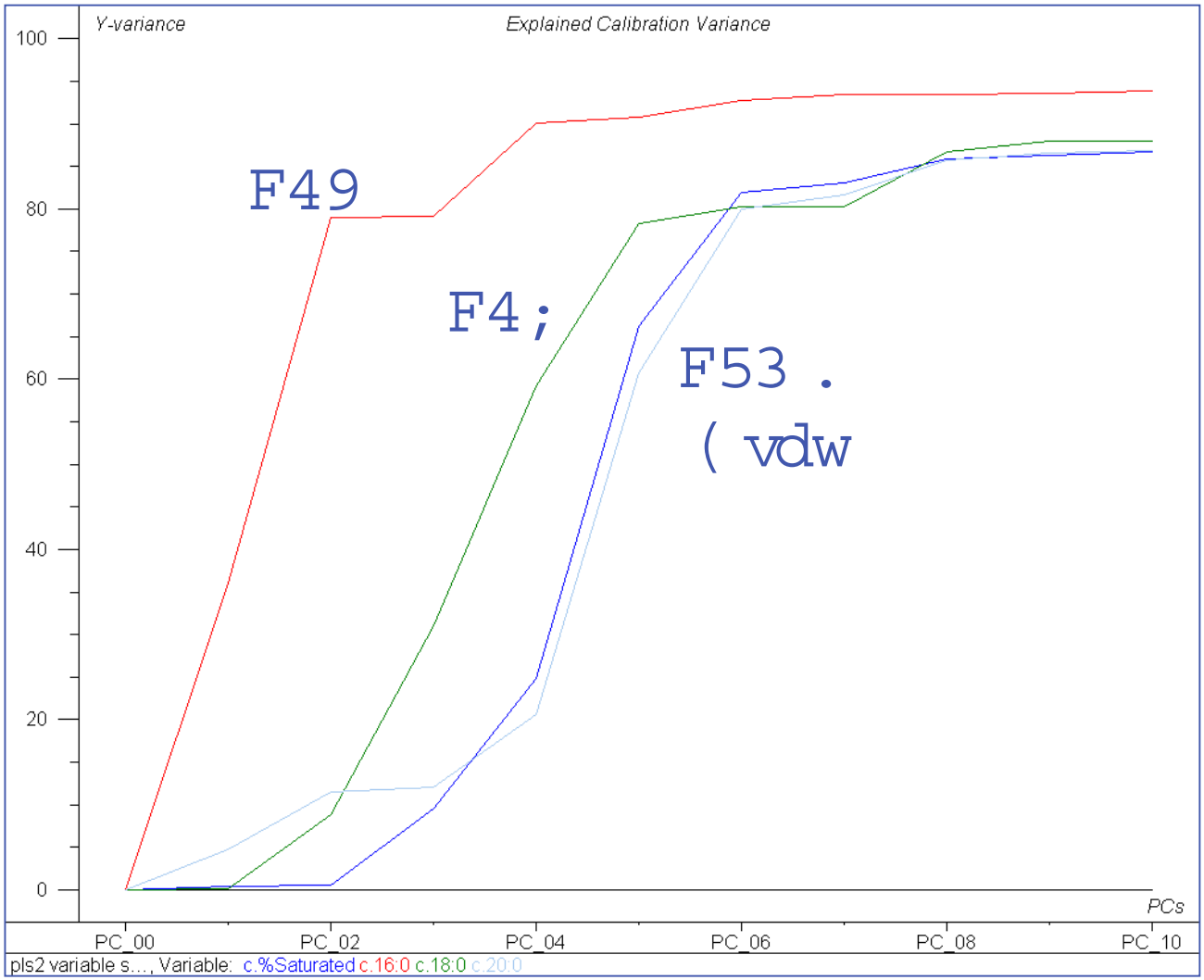


imagination at work

Visualising the oil composition

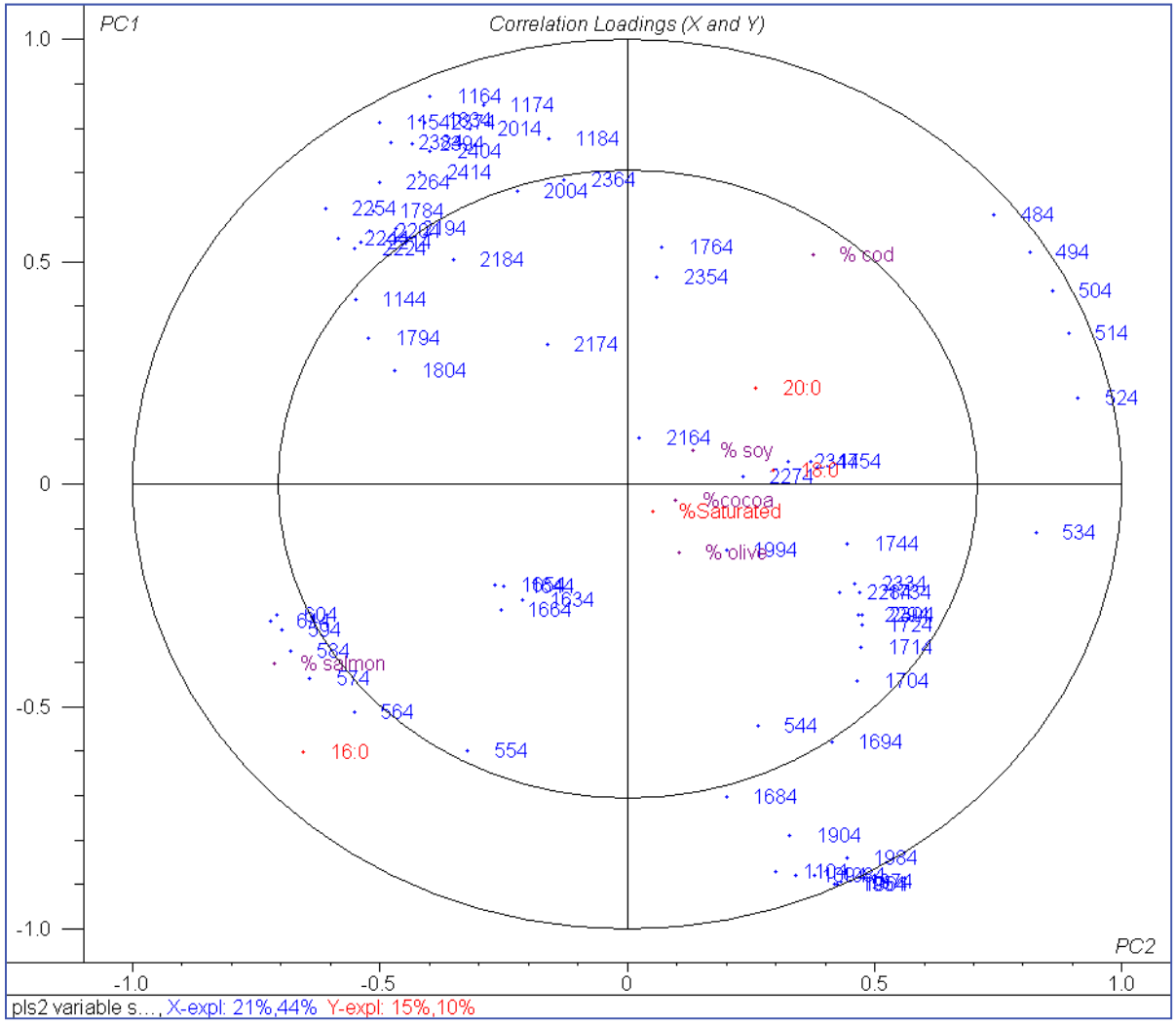
- Model 1: X =NIR , Y = %saturated fat, C16,C18,C20
- Amount of each oil-type as passified variables
- Use correlation loadings to interpret the underlying cause for the fat profile

Explained variance, Aopt = 6?



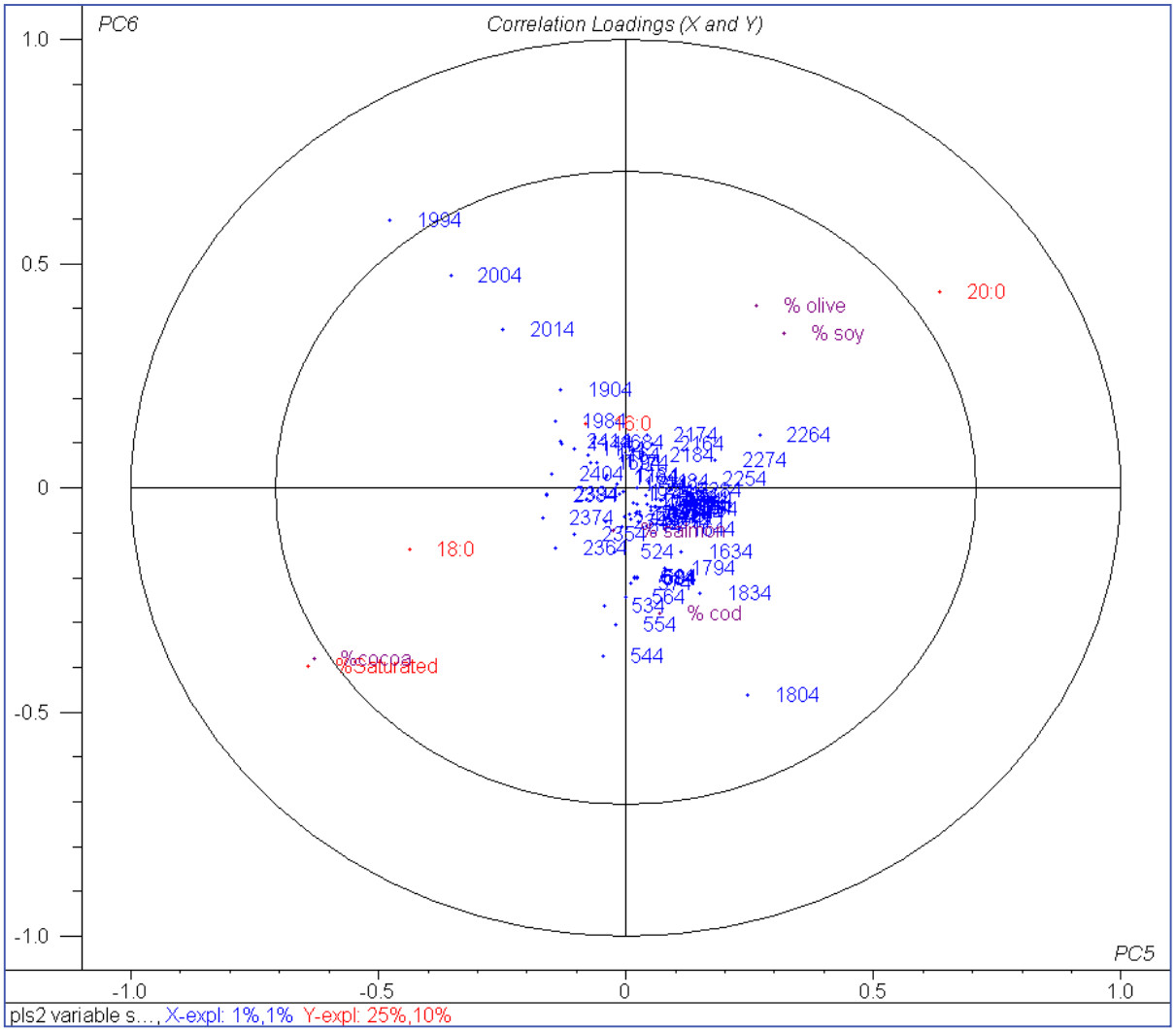
imagination at work

Correlation Loadings PC 1, 2



imagination at work

Correlation Loadings PC 5, 6



imagination at work

Summary

- The cross model validation acts as an efficient filter to reduce Type I error
- May also do repeated CMV if the structure of the objects allows
- Passifying variables is a way to visualise different types of background information
- Further studies are needed to compare CMV with rotation tests etc.

Acknowledgement

... Nils Kristian Afseth, Matforsk, Norway

Thanks for your attention!

... and may your hypotheses be with you