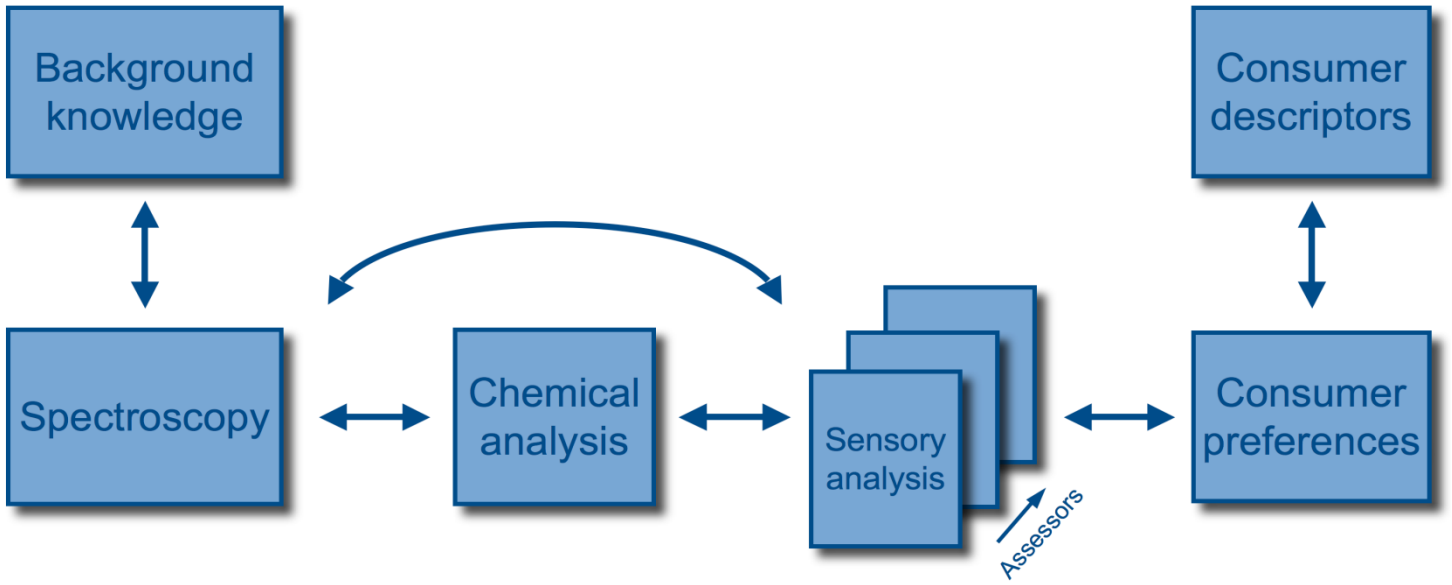
Multi-block regression based on sequential PLS regression

Tormod Næs

Nofima, Norway and Univ. Copenhagen, Denmark

Typical situation in food science

(Similar in other disciplines)

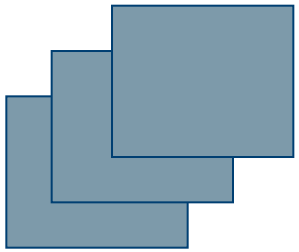


Understand and predict

Nofima

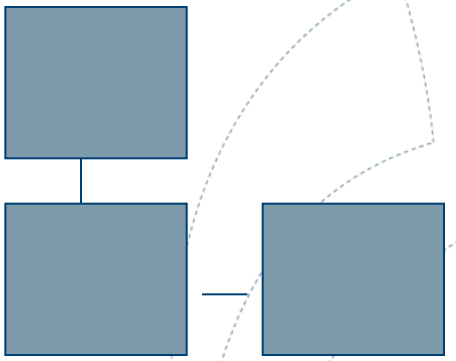
Important structures – main components

- Multi-way



- **Multi-block**
  - **Focus here**

- ”L-methods”



Nofima

Focus

- Discuss some new approaches based on PLS regression
    - PLS and orthogonalisation in sequence

  – Philosophy and results

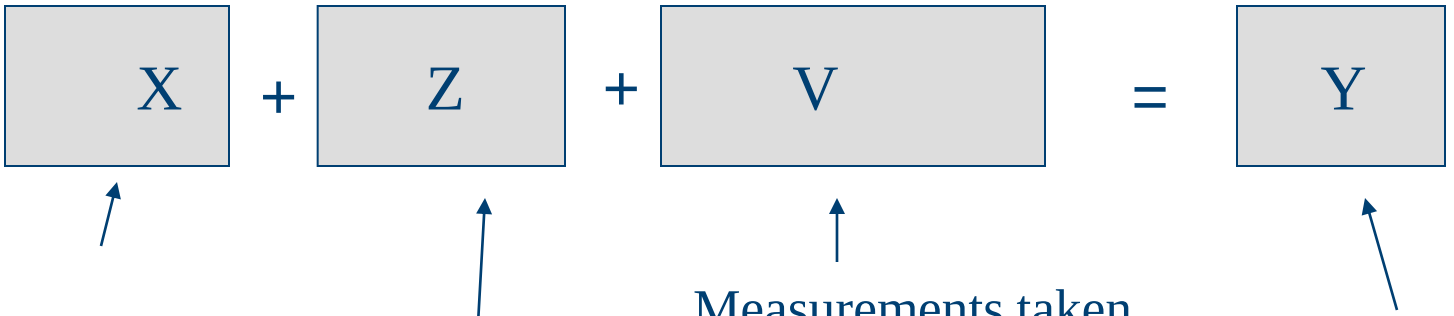
  – Closer to classical statistics than standard multi-block PLS regression (MB-PLS)

  – Invariance, explicit handling of different dimensionality of blocks

  – Interactions

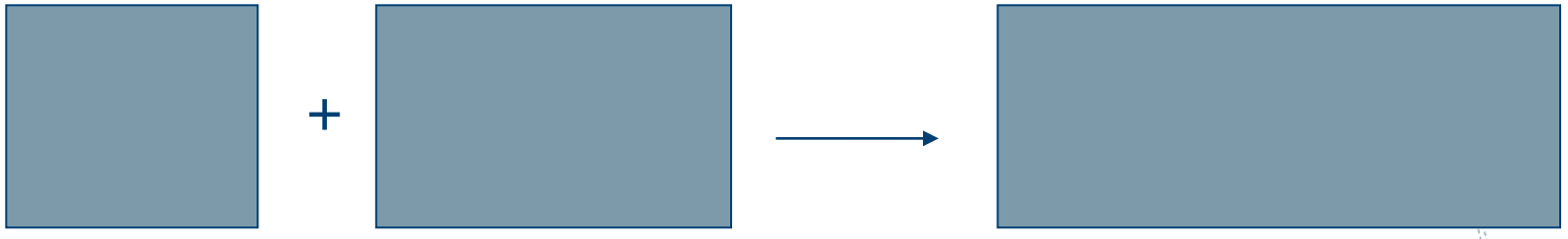
  – Relation to ANOVA

Typical example – process modelling



Jørgensen, K., Segtnan, V., Thyholt, K. and Næs, T. (2004). A comparison of methods for analysing regression models with both spectral and designed variables. J. Chemometrics, 18, 10, 451-464

Another typical example: sensory and consumer science



Prediction and interpretation

Måge, I., Menichelli, E. and Næs, T. (2011)- Preference mapping by PO-PLS: Separating common and unique information in several data blocks. Food Quality and Preference (in press).

Situation and model considered

- Multiblock regression model
  - (with interactions, see later)

$$Y = X\beta + Z\gamma + e$$

- X and Z can be anything: design, highly collinear etc.
  - Y can be multivariate

- Concentrate on two input blocks, but methodology can be extended

Nofima

Some possible approaches to multi-block regression

- Full LS (ANCOVA). Often impossible due to collinearity – large number of variables
- Full (concatenated) PLS of **Y** vs. **X**, **Z**
  - useful, but possibly problems with relative weighting and different dimensionality of blocks
- MB-PLS regression
  - Concatenated PLS with additional tools - common components etc.
    - Better, but similar problems as for concatenated PLS regression
- LS regression of **Y** vs. principal components of **X** and **Z** (computed separately).

SO-PLS – sequential and orthogonalised PLS regression

1. Fit block Y to X using PLS regression (compute scores and loadings)
2. Orthogonalise Z with respect to X (or PLS components from X), Z^{orth}
3. Fit Y to the Z^{orth} (scores, loadings)
4. Fit Y to scores T_X, T_Z^{orth} (independent, orthogonal)

- For more than two input blocks, the same procedure is repeated

SO-PLS – some properties

$$\hat{\mathbf{Y}} = \mathbf{T}_X\mathbf{Q}_X^t + \mathbf{T}_Z^{orth}(\mathbf{Q}_Z^{orth})^t = \mathbf{X}\hat{\mathbf{V}}_X\,\hat{\mathbf{Q}}_X{}^t + \mathbf{Z}^{orth}\hat{\mathbf{V}}_Z^{orth}(\hat{\mathbf{Q}}_Z^{orth})^t$$

- Can be back-transformed to original units for X and Z
- Invariant wrt. relative weighting of blocks
- Different dimensionality - explicitly handled
  - For instance: Design variables and large multivariate blocks

Nofima

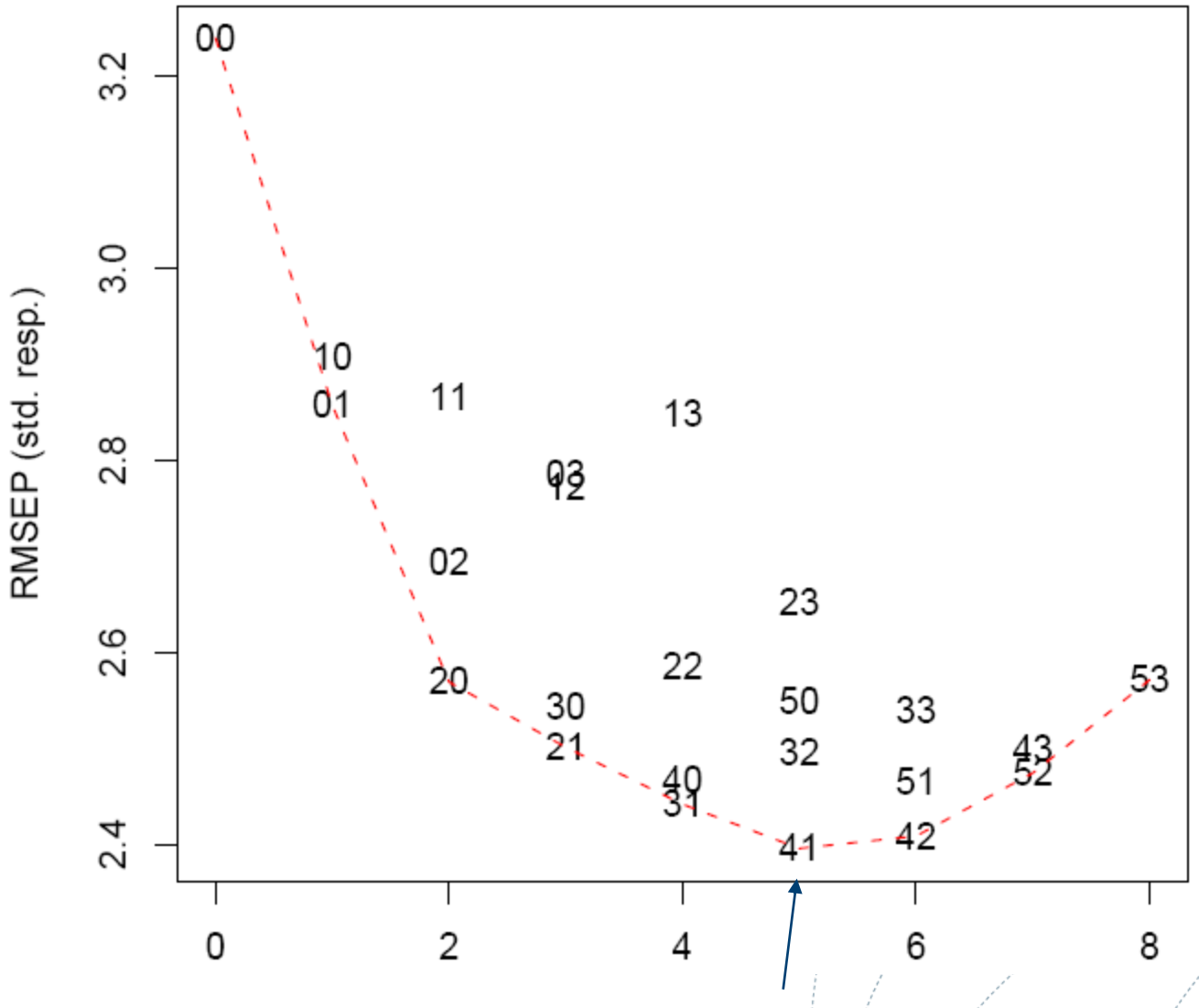
Order of the blocks?

- Sometimes obvious
  - Design + extra information (process and raw materials, ANCOVA)
- In other cases not obvious
  - Often similar prediction ability
  - Interpret both ways – additional information?
    - Problem is turned to an advantage

Validation

- CV as usual
  - Determine the number of components by Måge plot (see later)
- Two options: Universal vs. sequential optimisation
- Sequential fits better to the idea of the method
  - Universal – better predictions
- Special interest in incremental improvement

Måge plot for component selection



Interactions

- Add columnwise products of linear functions of X and Z, XV_1 and ZV_2.(denote by *)

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + (\mathbf{X}\mathbf{V}_1) * (\mathbf{Z}\mathbf{V}_2)\boldsymbol{\varphi} + \mathbf{e}$$

  - Includes direct mulplication
  - Includes direct multiplication after variable selection
  - Includes principal components of X and Z.

Næs, Måge, Segtnan (2011). Incorporating interactions in multi-block SO-PLS regression. J. Chem (in press)

Estimation procedue

- SO-PLS with three blocks: Fit X and Z before X*Z
- Preserves invariance (wrt. relative weighting of the blocks)
  - Orthogonalisation and column-wise multiplication
- Direct generalisation of ideas from polynomial regression and ANOVA
- Non-linearities handled similarly (new blocks or in the same)

Interaction - example

Salting of salmon

X - Design variables

salmon size (3 categories)

salting level (3 levels)

Z - NIR measurements of fillets

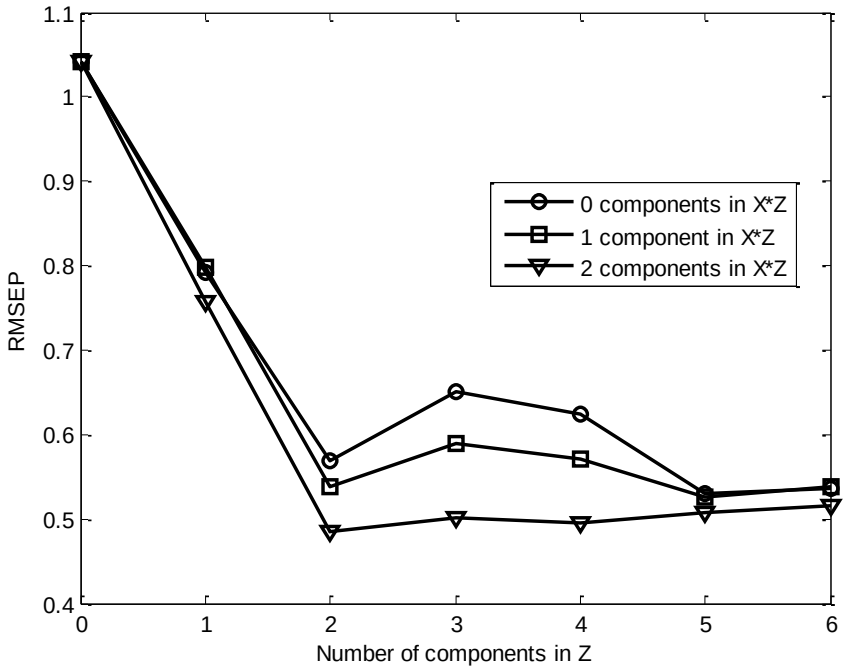
(6 highly collinear wavelengths, in the fat/water area)

Y – salt content after salting and storage

Nofima

Results

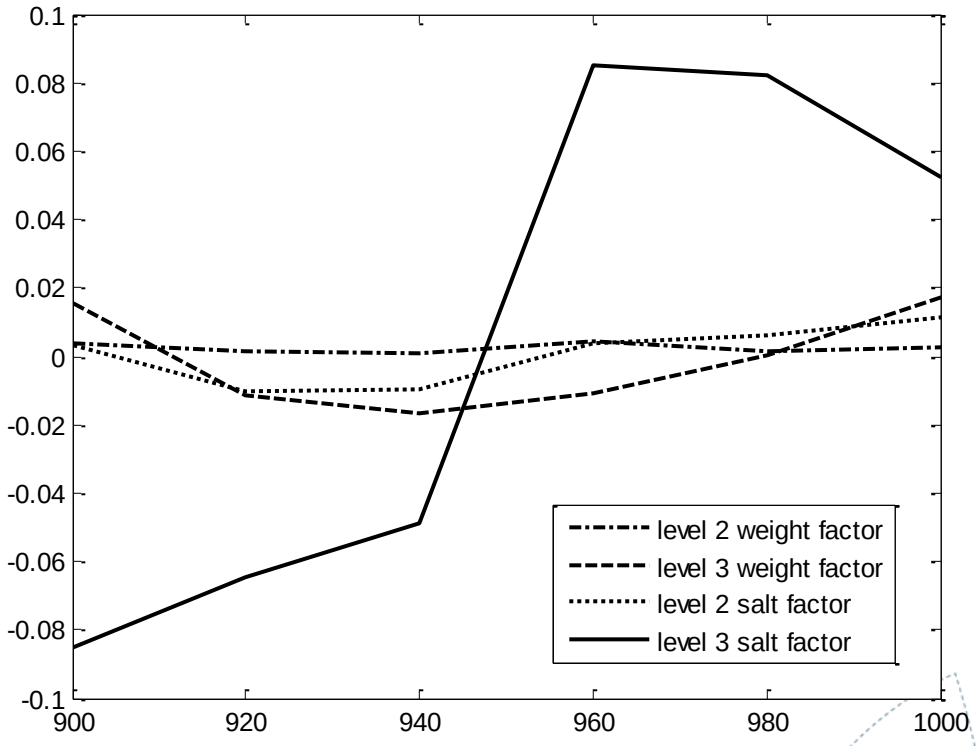
	Mean	Only X	X and Z	all blocks
RMSEP	1.21	1.04	0.57	0.49



The different lines correspond to different choice of components for **X*Z**. The horizontal axis represents the number of components in **Z**
X fitted by LS

Nofima

Regression coefficients for **X*Z** based on PLS.



The one with the largest deviation from 0 (_____) is the one corresponding to level 3 for factor 2.

Relations to standard Type I ANOVA

- Sequential fitting
  - Linear effects before interactions – the same idea as underlying Type I ANOVA.

- PLS is equal to LS for the maximum number of components
  - Direct generalisation of ANCOVA for data that can not be analysed by LS

- Information about incremental contributions (improvements)
  - Can decompose SS into a sum of contributions from each block and residuals (orthogonal)

$$SS_{Tot} = SSX + SSZ^{orth} + SS(XZ) + SS_{E}$$

- Testing is more problematic since DF's are not known for PLS
  - Can use CV-ANOVA (possibly also the bootstrap)

Indahl, U.G. and Næs, T. (1998) Evaluation of alternative spectral feature extraction methods of textural images for multivariate modelling. Journal of Chemometrics, 12,4, 261-278

”ANOVA” table based on CV-ANOVA

Source	RMSEP	“MS” (SS/N)	p-values (2-way CV-ANOVA)
First Matrix	288	2978228	0.017
Second Matrix	284	67671	0.725
RES		2095970	
Total	467	5668190	

First matrix - design, second matrix - spectral data

Interpretation

- The model gives independent PLS models, regression coefficients (original or orthogonalised units) and prediction after each block
- Three plots based on these aspects
  - Direct interpretation of PLS models
    - Useful for outlier detection and also for interpretation
  - PCP
    - Method-independent based on regression coefficients
      - After back-transformation to original units
  - Projection onto PC's of predicted Y
    - Projections based on X and based on X,Z (compare)

PCP for interpretation

- Basic idea. PLS components are introduced for prediction and do not necessarily reflect the natural dimension of the problem.
  - Also sometimes difficult to interpret if many

- PCA of predicted Y
  - Scores and Y-loadings
  - The scores are linear functions of the independent variables (X-loadings)
    - Linear functions of linear functions
  - The latter gives X-loadings (coefficients)
  - Plot the usual way (as for PLS)

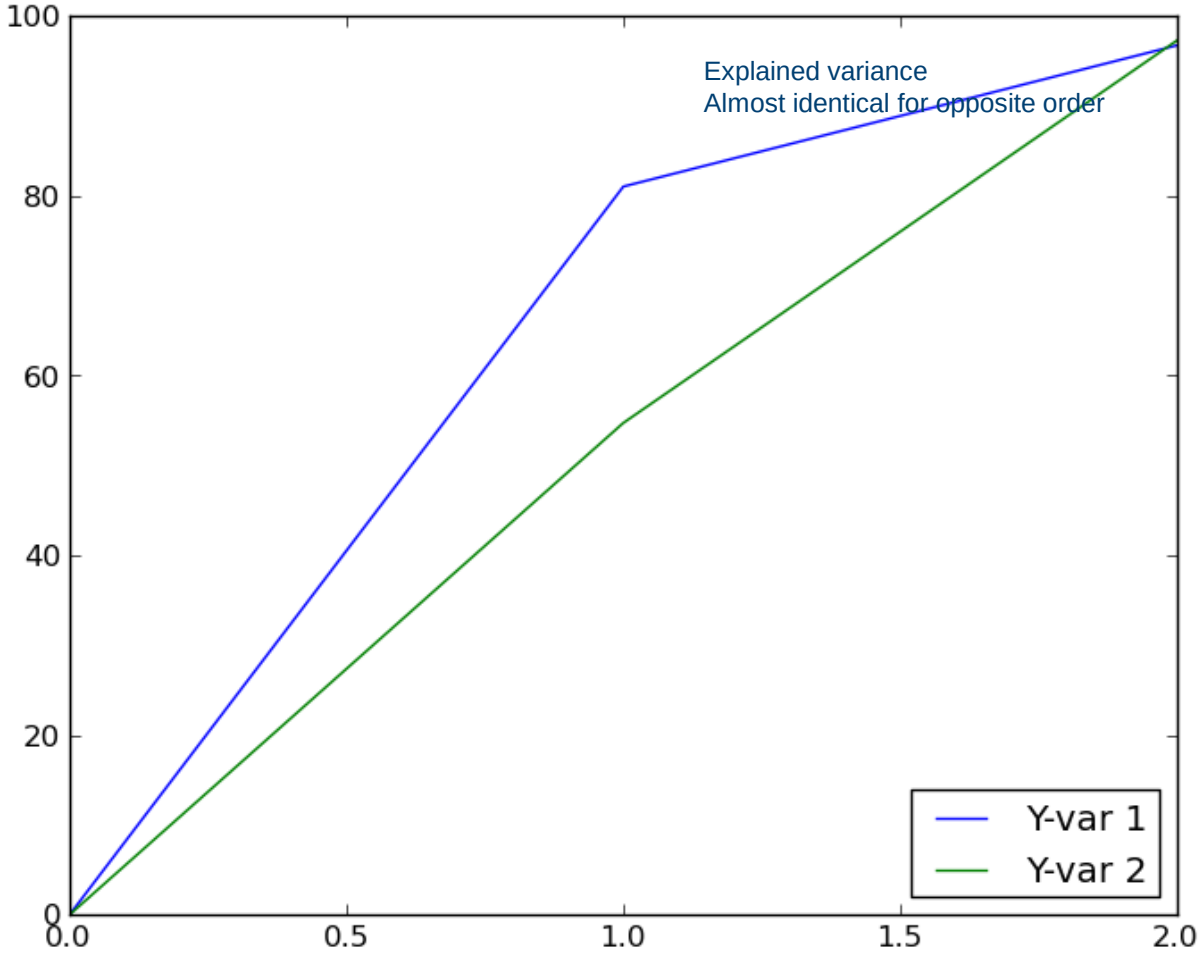
- If only one Y - corresponds to regression coefficients

Langsrud, Ø., Næs, T. (2003). Optimised score plot by principal components of prediction. Chemolab. 68, 61-74.

Example: Y- two-dimensional, X: NIR – Z: Raman

7 comp - NIR

5 comp - Raman



Y1 = PUFA%emul

unsaturated fat as % of sample

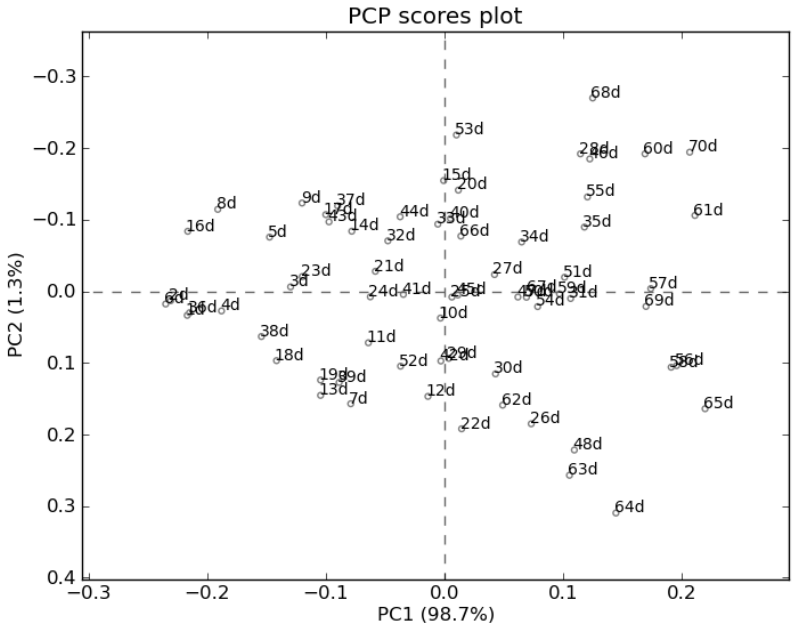
Y2 = %Pufa

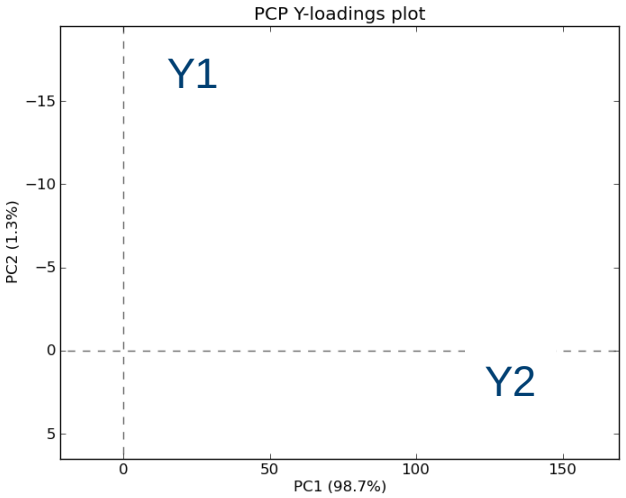
unsaturated fat as % of fat

Nofima

Block 1: NIR – Block 2: Raman

PCA of predicted Y

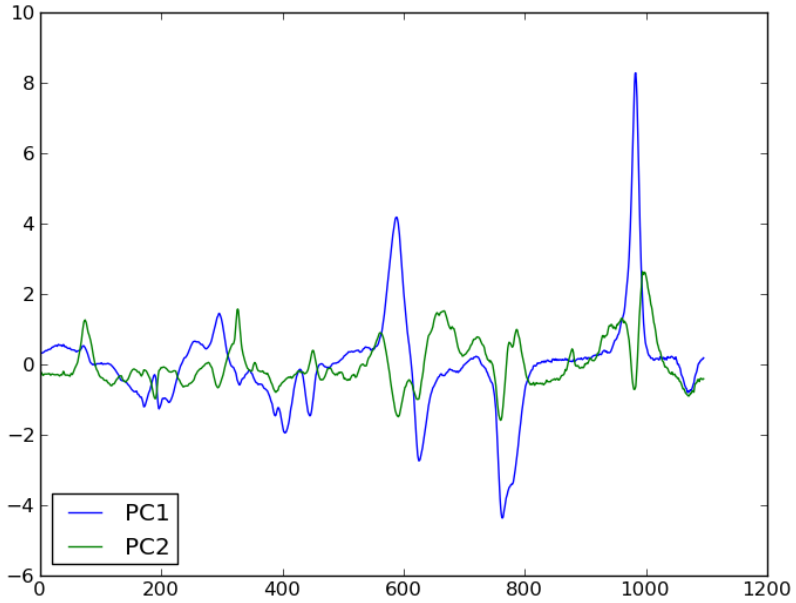




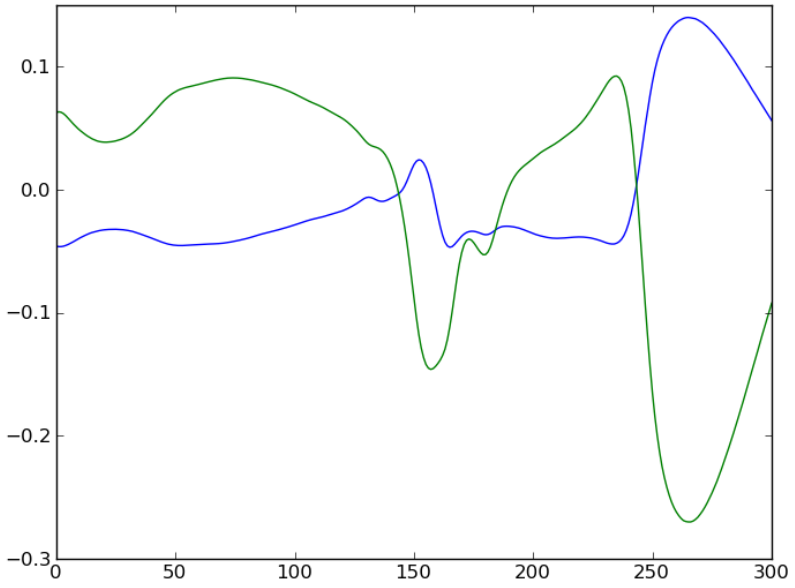
Y-loadings

Raman and NIR X-loadings from PCP

Raman



NIR

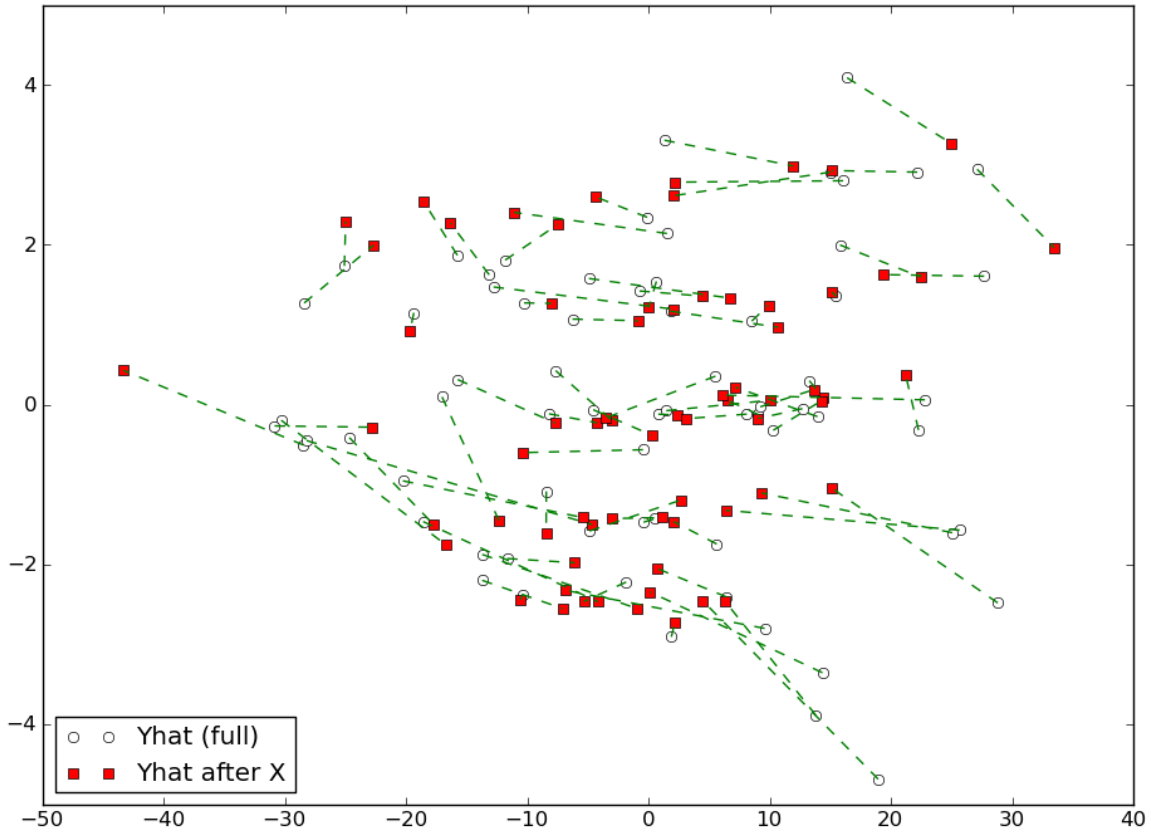


Blue related to %PUFA = Y2
Green related to PUFA%emul = Y1

Illustration of projection approach

Plot of predicted values after X and after (X, Z) projected onto the PCA space for predicted Y

Raman improves prediction along first component, %Pufa related



SO-PLS for path modelling

- Methodology for linking several data blocks (manifest variables) according to a given relation between the blocks (path diagram-arrow diagram)
  - causal or other

- **Structural equations modelling (SEM)**

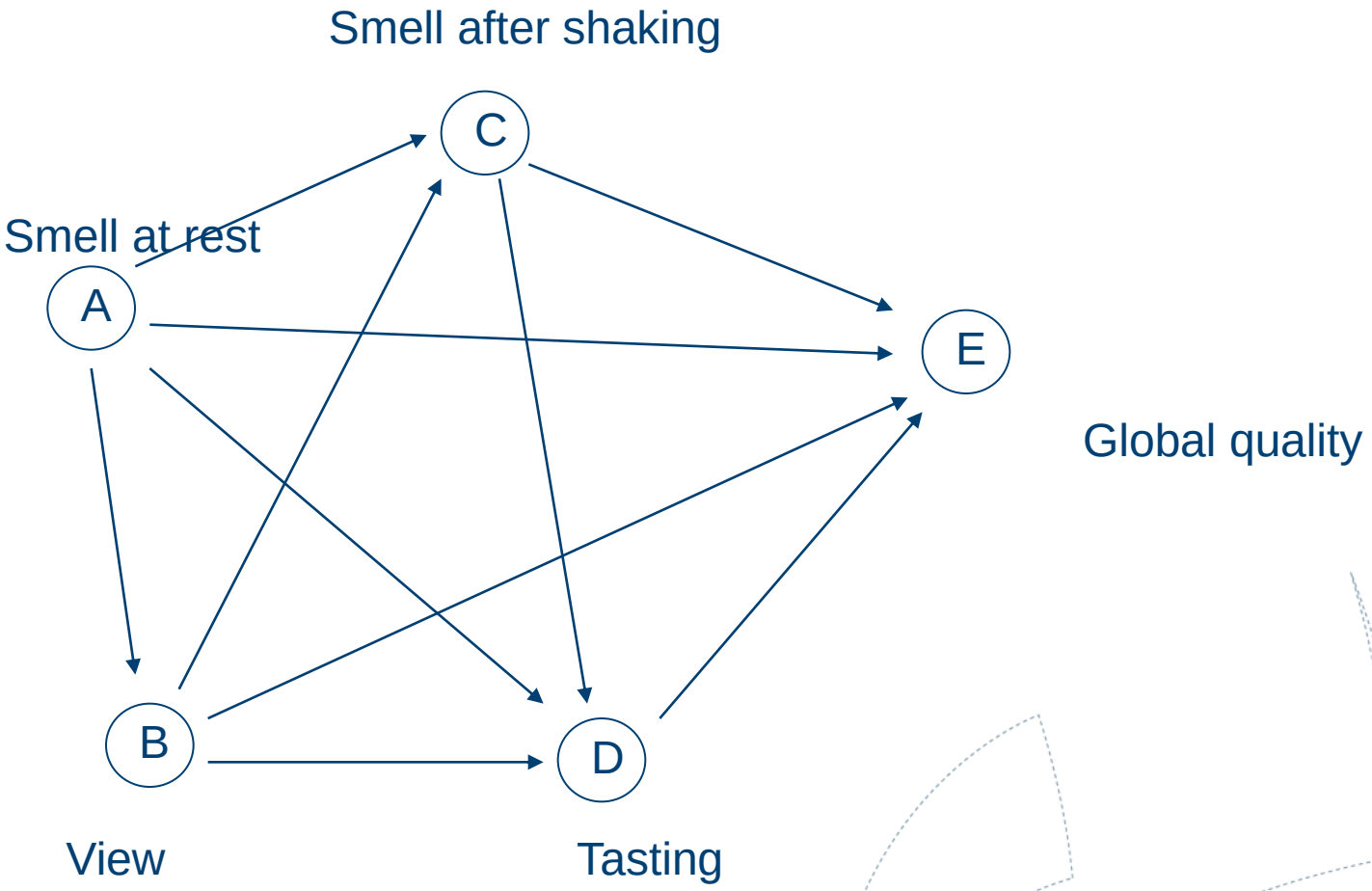
  - Models based on two elements/parts

    - Measurement model for each manifest block, outer relations (Factor analysis model)

    - Path model in the latent variables (inner relations)
      - Joint set of regression models

Nofima

SO-PLS has also been used for path modelling



Næs, T. Tomic, T., Mevik, B-H. and Martens, H. (2011). Path modelling by sequential PLS regression. Journal of Chemometrics, 28-40

New approach

Two elements (estimation and interpretation)

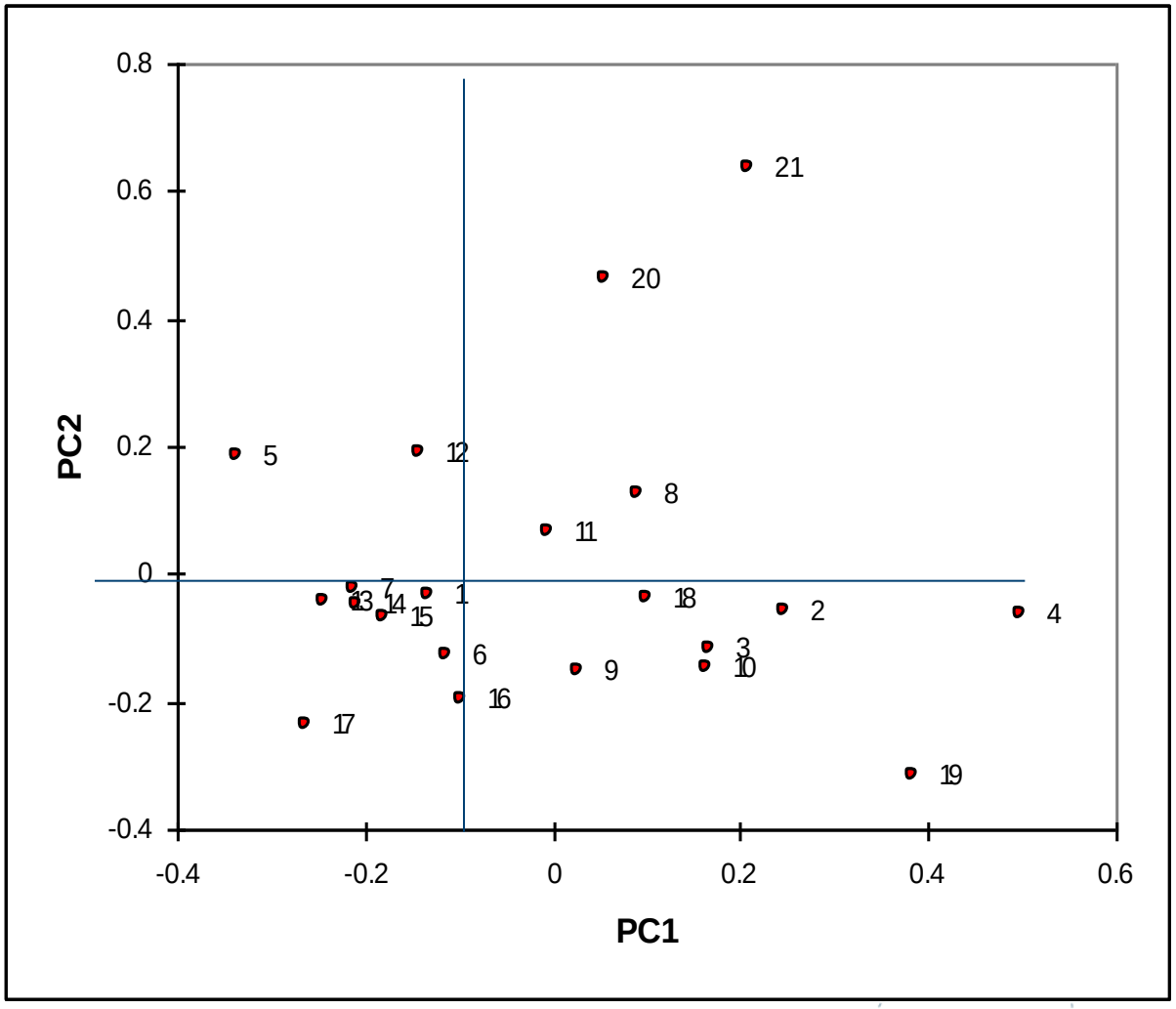
1. SO-PLS for each endogenous block – separate models
2. Principal components of prediction (PCP) for interpretation

Allows for

Different dimension in each block

Different information used for prediction and to be predicted in each block

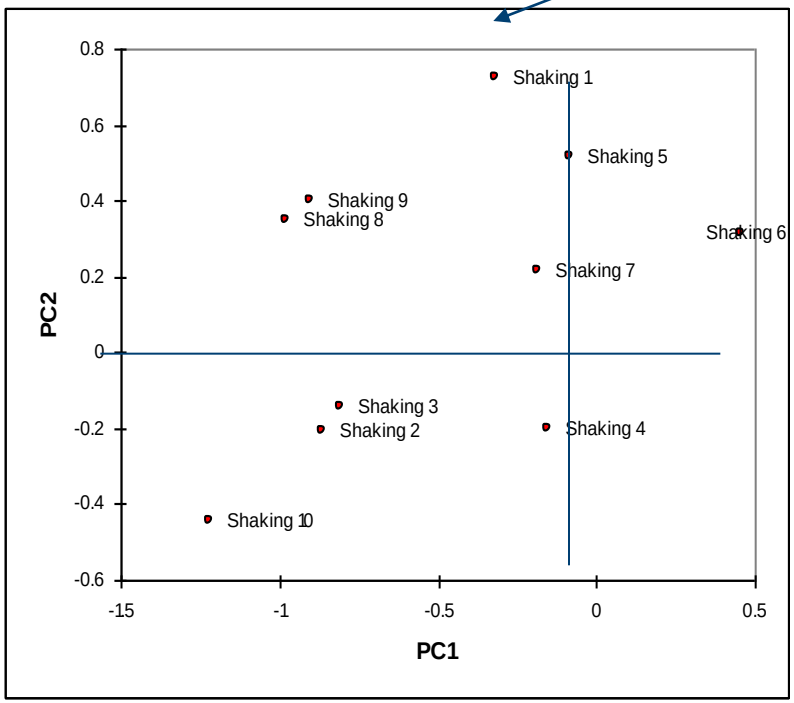
20%



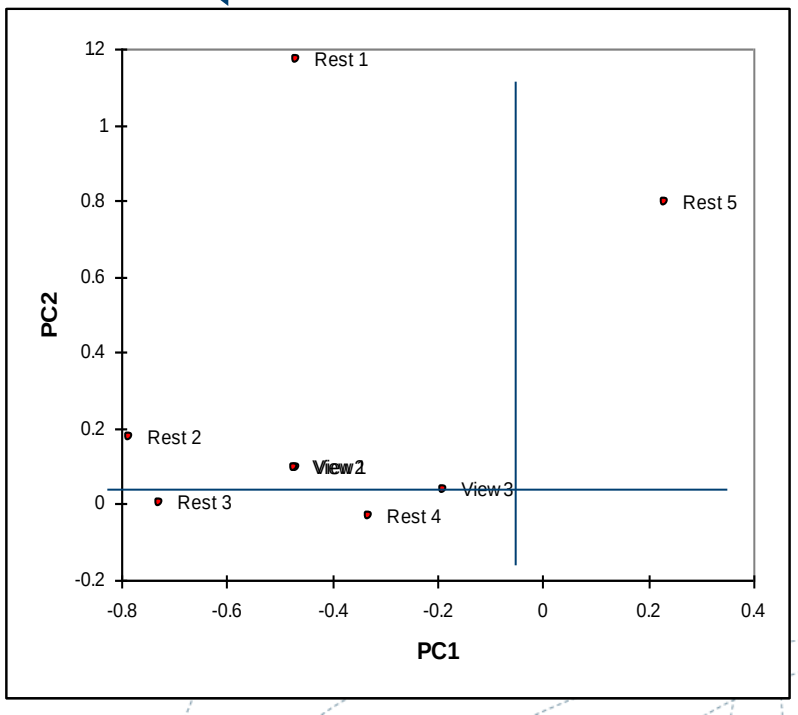
61%

Scores plot for model 2, C predicted from A and B

Loadings –plots for Y and X, PCP



C



A, B

PO-PLS

- Similar to SO-PLS, but focus on common varibility, not on additional variability

- First define the common variability space as the space spanned by linear combinations with correlation close to 1 (canonical correlation).
  - Subspace shared by the two blocks
  - Reduce dimensionality first – stabilize or use regularised canonical correlation

- Then orthogonalise X and Z wrt. this space

- PLS of Y onto the orthogonalised versions of X and Z.
  - Then LS of all three scores matrices

- Can be combined with SO-PLS. First XZ^{common}, then X^{orth} and finally Z^{orth}.
  - All blocks orthogonal

- Also inviariant wrt. different scale of the blocks – and allows for different dimensionality of the blocks

Måge,I. Mevik, B-H. and Næs, T. (2008). Regression models with process variables and parallel blocks of raw material measurements. J. Chemometrics, 22, 443-456.

Closely related to confounding (or collinearity),
Confounding among blocks- not among variables

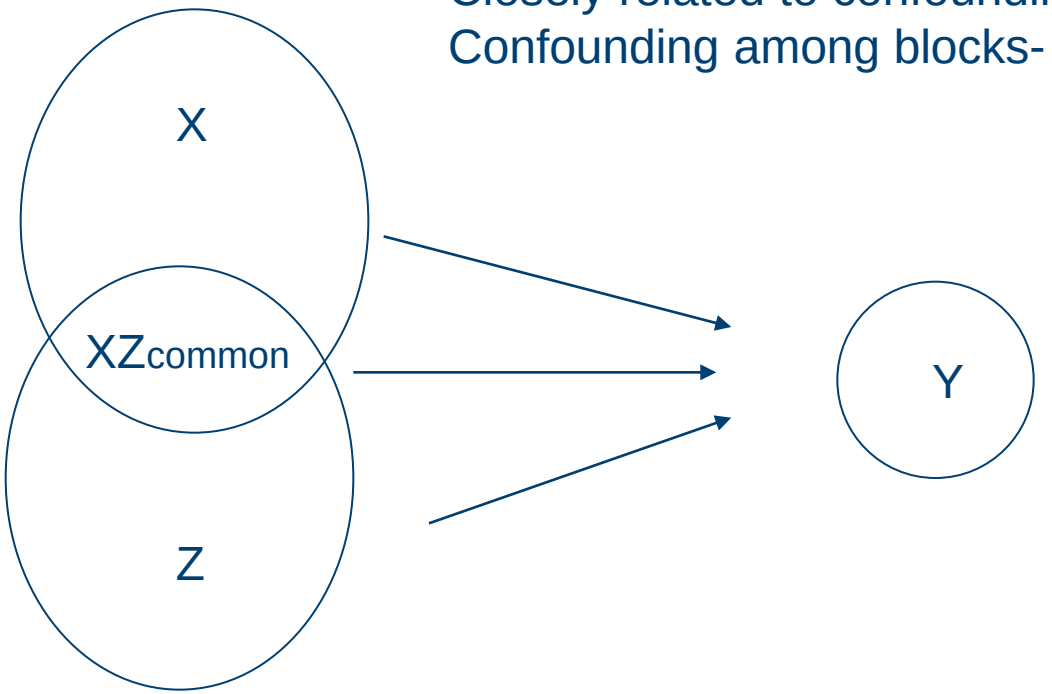
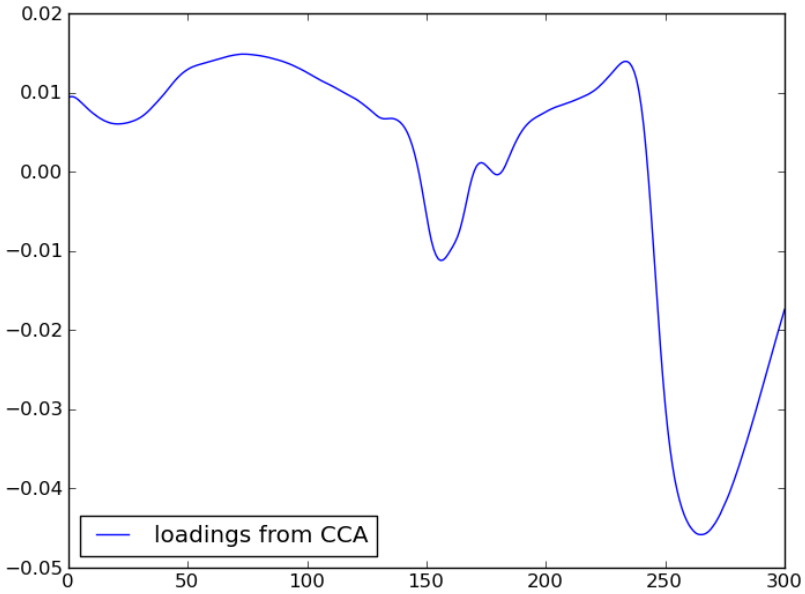
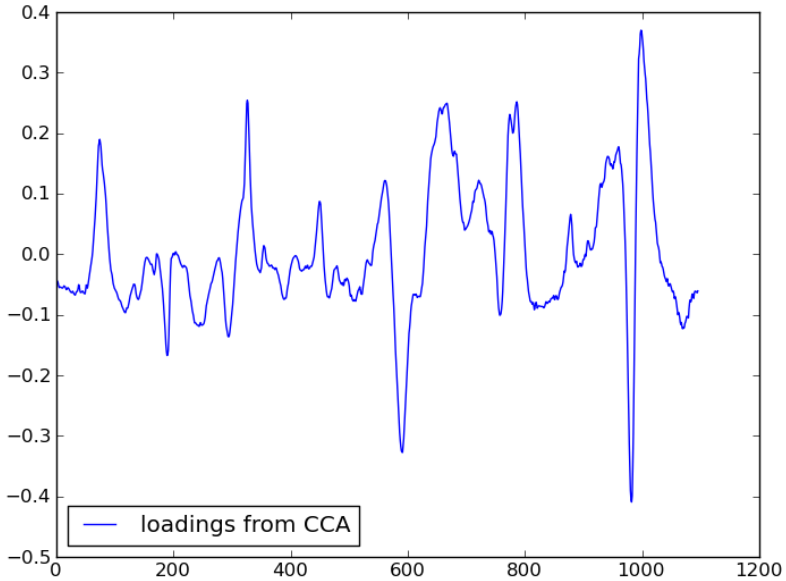


Illustration of common variability

NIR and Raman loadings from PO-PLS

Canonical correlatioin equal to 0.9, PCA on each first





Most closely related to Pufa%emul
Can be seen from plots and correlations

Summary

- SO-PLS and PO-PLS: New methods for multi-block regression
  - Explicit focus on additional and joint information
- Flexible in interpretation,
  - PLS models, joint interpretation after back-transformation (PCP), additional information - projections, outliers
- Invariant and different dimensionality
- Interactions can be allowed in SO-PLS
- Natural extension of Type I ANOVA
  - Close relation to statistics,
  - Fits to PLS philosophy of extracting information and using residuals
- Challenge: better testing (bootstrap?)

Nofima

References

Jørgensen, K., Segtnan, V., Thyholt, K. and Næs, T. (2004). A comparison of methods for analysing regression models with both spectral and designed variables. J. Chemometrics, 18, 10, 451-464.

Måge, I. and Næs, T. (2005). Split-plot regression models with both design and spectroscopic variables. Journal of Chemometrics. 19, 521-531.

Jørgensen, K. Mevik, B-H. and Næs, T. (2007). Combining designed experiments with several blocks of spectroscopic data. Chemolab. 88, 2, 143-212.

Måge,I. Mevik, B-H. and Næs, T. (2008). Regression models with process variables and parallel blocks of raw material measurements. J. Chemometrics, 22, 443-456.

Jørgensen, K. and Næs, T. (2008). The use of LS-PLS for improved understanding, monitoring and prediction of cheese. Chemolab, 93, 11-19.

Næs, T. Tomic, T., Mevik, B-H. and Martens, H. (2011). Path modelling by sequential PLS regression. Journal of Chemometrics, 28-40

Næs, T, Måge, I and Segtnan, V.H. (2011). Incorporating interactions in multi-block SO-PLS regression. Journal of chemometrics, 25, 601-609.

Måge, I., Menichelli, E. and Næs, T. (2011)- Preference mapping by PO-PLS: Separating common and unique information in several data blocks. Food Quality and Preference (in press).

Acknowledgements

- Oliver Tomic
- Ingrid Måge
- Vegard Segtnan
- Nils Kristian Afseth