Path modelling by sequential PLS regression

Tormod Næs, Oliver Tomic,
Bjørn-Helge Mevik, Harald Martens

Nofima Mat

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

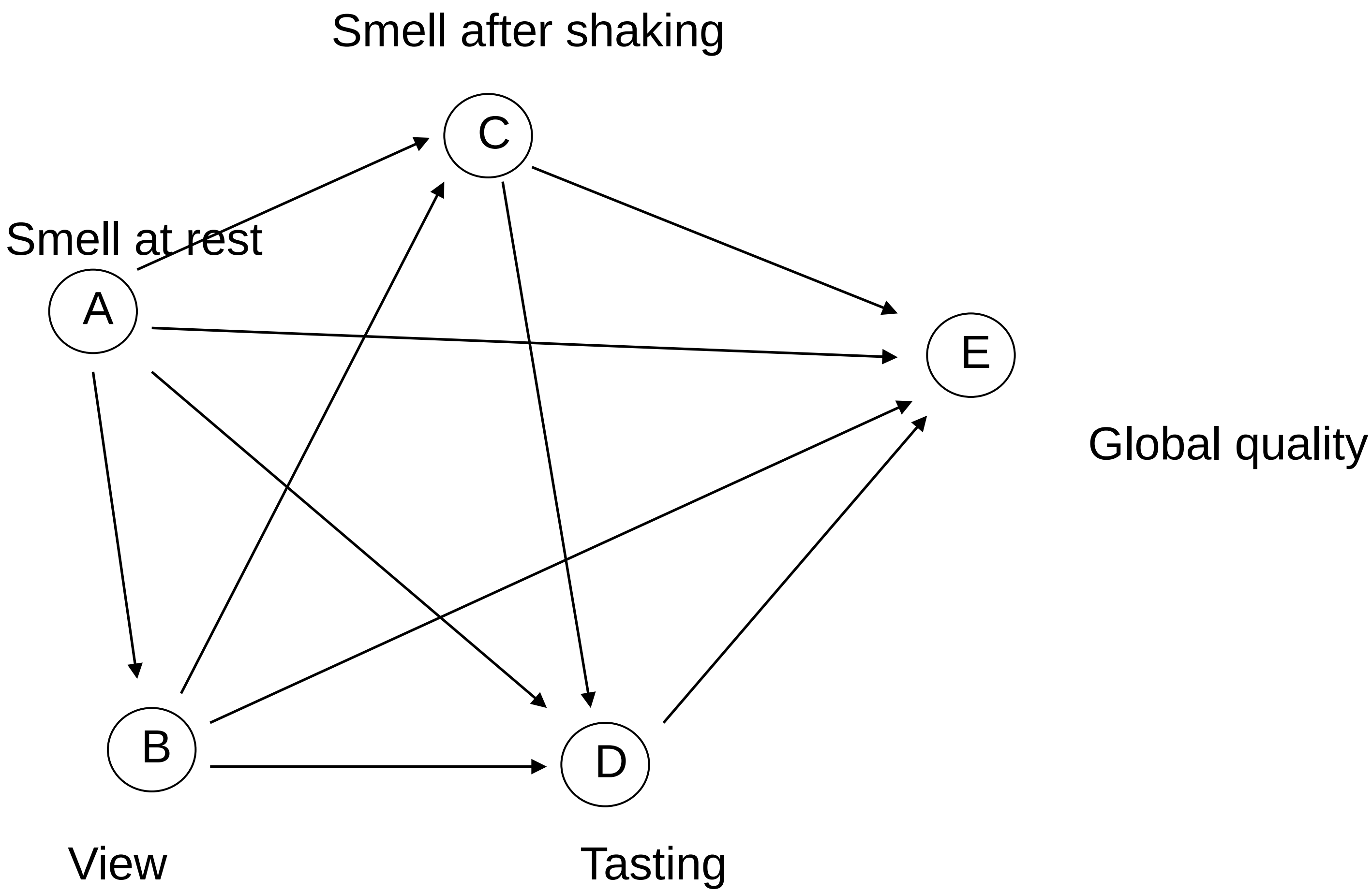
- **Structural equations modelling (SEM)**

  - 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

Sensory analysis of wine



Escofier, B. and Pages, J.(1988).
Tenenhaus, M. and Vinzi, V.E. (2005)
Martens, M., Tenenhaus, M, Vinzi, V.E. and Martens, H. (2007).

Two important traditions

- PLS
  - Algorithmic foundation, not so easy to understand why works
  - The criterion is somewhat complex (some new results and recent modifications exist)
  - Convergence generally works well in practice
  - Can handle collinearity and more variables than samples
  - Emphasis on both scores and structure (population and individual differences)

- ML - LISREL
  - Model and criterion based (statistical)
  - More samples than variables are required (at least for the classical solution)
  - Less emphasis on scores, mostly on structure
  - Sometimes identification and convergence problems

Possible problems

- One-dimensional blocks
  - PLS. Some attempts have been made to solve it
    - Deflation and PLS for the outer relations
  - ML. Can be done, but possibly quite complex (identification and convergence)
- Same information used for prediction and to be predicted in each block
  - No reason to expect that
  - Are SEM models appropriate?

New approach

- Instead fo repairing already exsisting methods
  - New approach from scratch

- Explorative, focus on interpretation, but only in validated models

- **Two elements** (estimation and interpretation)

1. SO-PLS for each endogenous block – separate models
   - Sequential and orthogonalised PLS (SO-PLS)
   - Cross-validation (global and incremental)

2. Principal components of prediction (PCP) for interpretation

SO-PLS, Regression method based on serial/sequential modelling (focus on incremental contributions)

$$\boxed{Y} = \boxed{X} + \boxed{Z} + \boxed{V}$$

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

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

SO-PLS
Sequential orthogonalisation and the use of PLS

- Fit first block Y to X with PLS (scores and loadings)
- Orthogonalise Z with respect to X
- Fit Y to the Z(orth) (scores, loadings)
- Orthogonalise V wrt X and Z
- Fit Y to V(orth) (scores and loadings)
- Fit Y to scores TX, TZ and TV (independent)

At each step: Fit Y to the part of a new block that is orthogonalised to previous blocks.

Properties of SO-PLS

- Scale invariant wrt blocks
- Different dimensionality in each block allowed
  - Can combine design variables and others
- Incremental contributions.
  - Type I strategy (ANOVA)
- Many more variables than samples allowed
- Good prediction and improved interpretation as compared to joint PLS.
  - Can interpret each block separately
- No convergence problems
- LS if all components are included
- **In this context: The problem of same information for prediction and to be predicted vanishes**
  - **extends the standard SEM assumptions**

PCP for interpretation

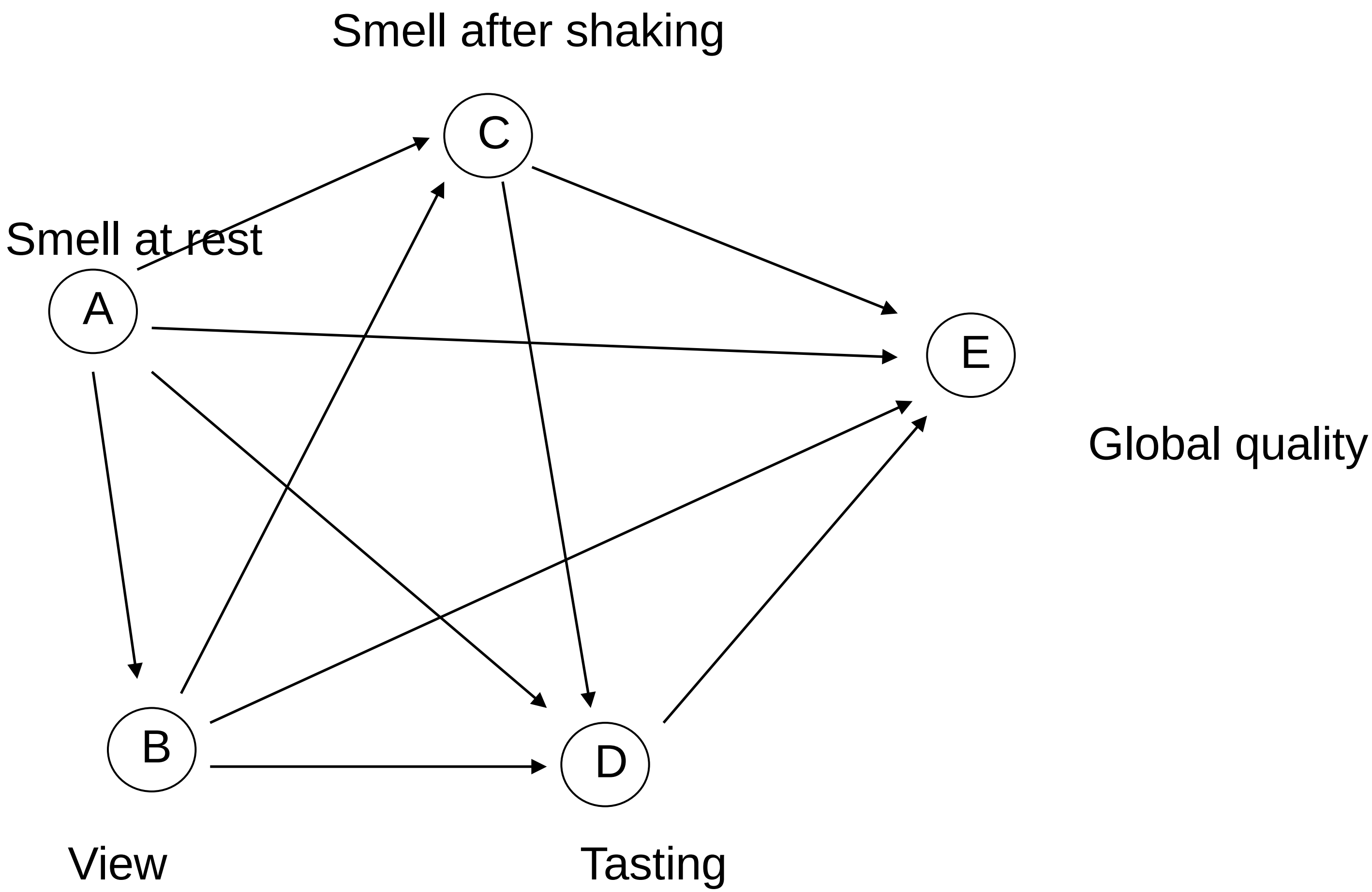
- The SO-PLS leads to many plots in this context.
  - We want one plot for each endogenous block
  - Use PCP – principal components of prediction

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

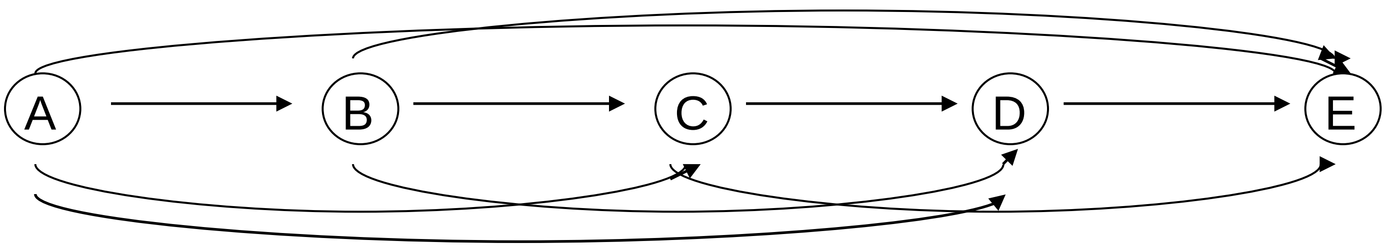
- PCA of predicted Y (scores and Y-loadings)
  - Scores and Y-loadings
  - The scores are linear functions of the independent variables (X-loadings)
  - The latter gives X-loadings

- Can also look at more details in the SO-PLS model

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

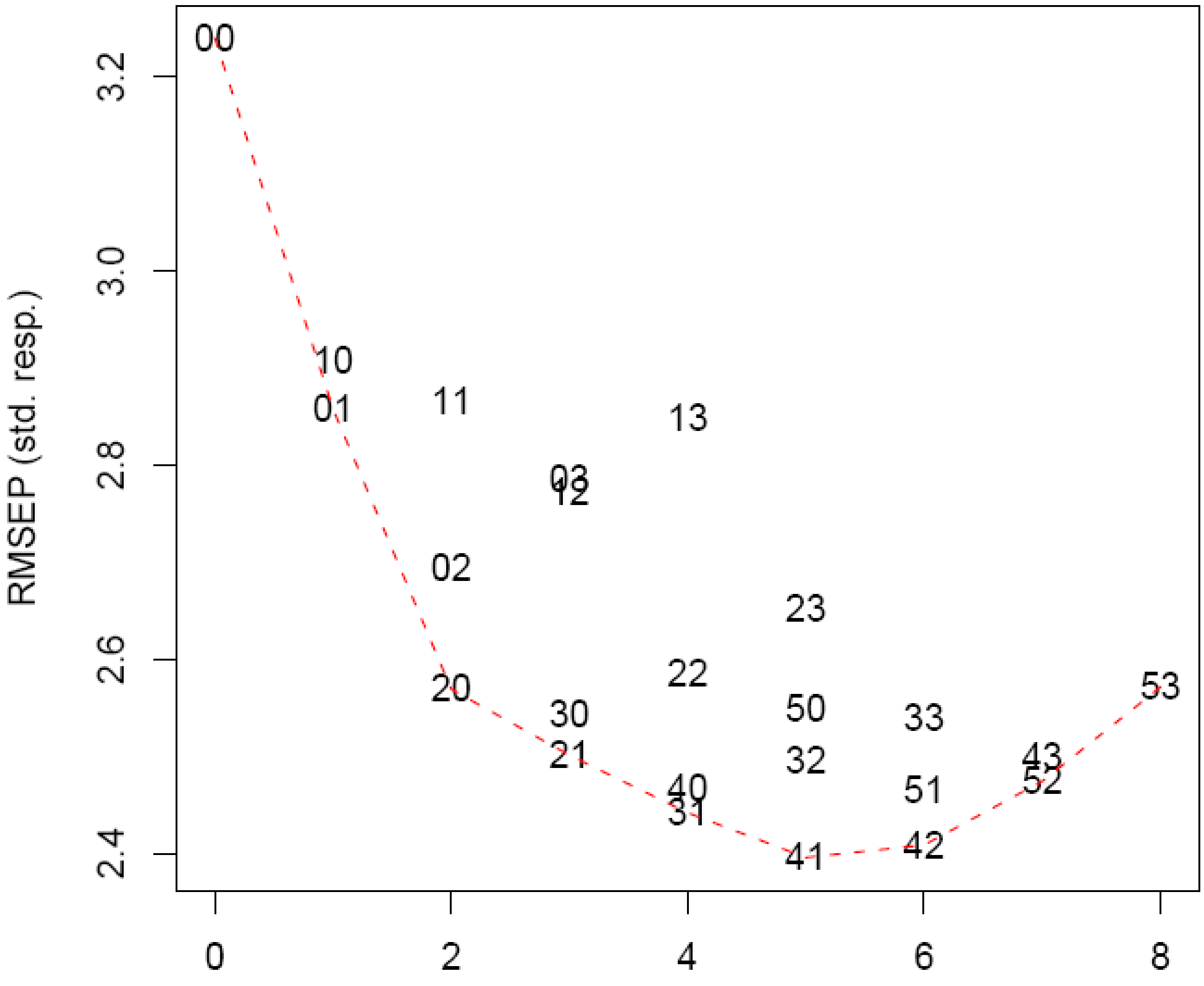


Number of manifest variables: 5, 3, 10, 9, 1
Number of samples = 21



Dependence diagram, usually quite obvious
(Sometimes a choice has to be made)

For each endogenous block, the arrows indicate the input



Måge plot for model 2, prediction of C from A and B

Explained variances (cross-validation) for the different input matrices in all the 4 models.

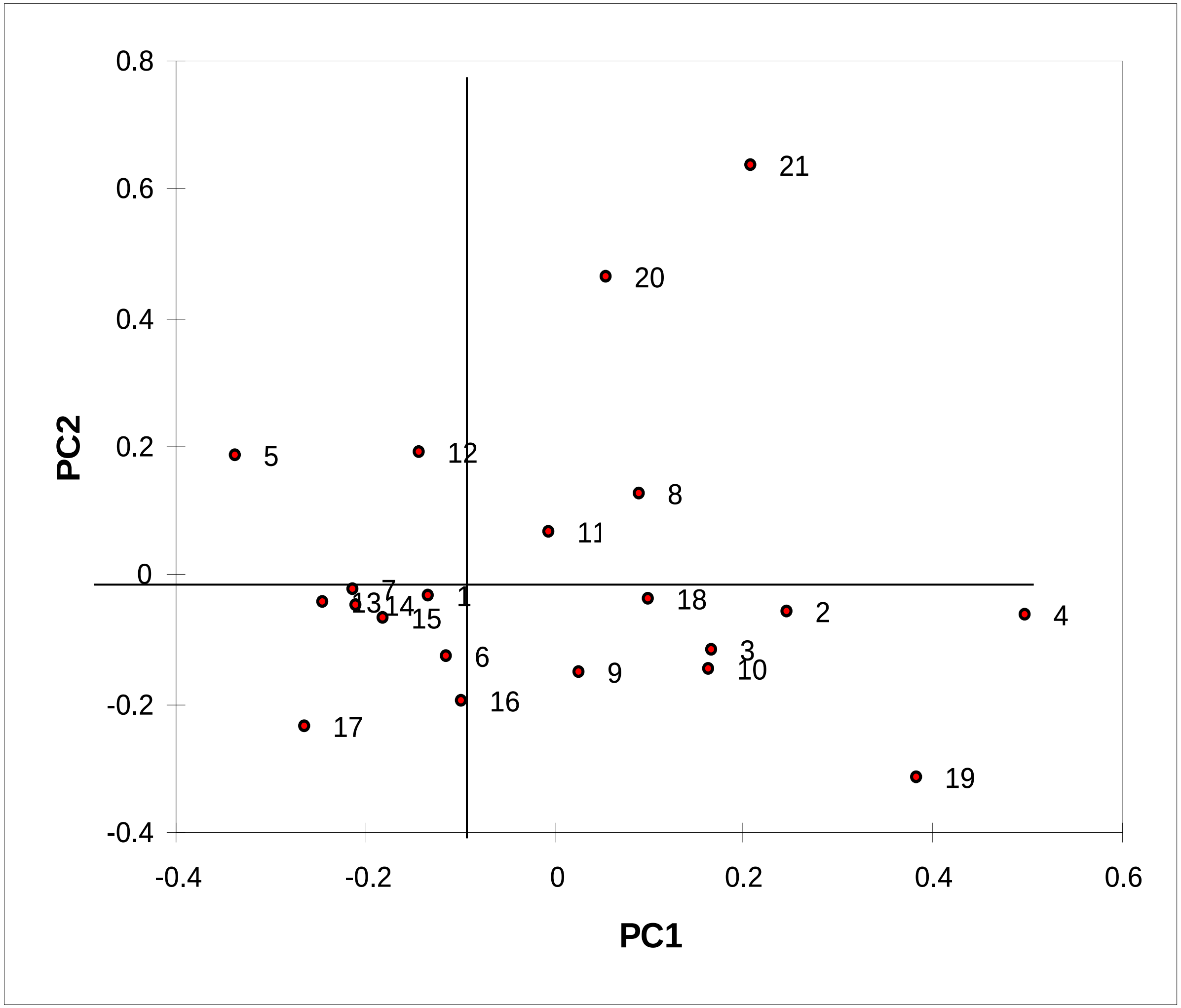
	Model 1	Model 2	Model 3	Model 4
Block A	37 (1)	42,5 (4)	0,0 (0)	0,0 (0)
Block B		45,3 (1)	41,1 (2)	0,0 (0)
Block C			50,9 (2)	78,4 (2)
Block D				96,5 (3)

Explained variances (in %) of the predicted Y (CV)

	Model 1	Model 2	Model 3	Model 4
1. component	100	61	85	100
2. component		81	96	
3. component		92	97	

Model 2 is clearly 2-dimensional

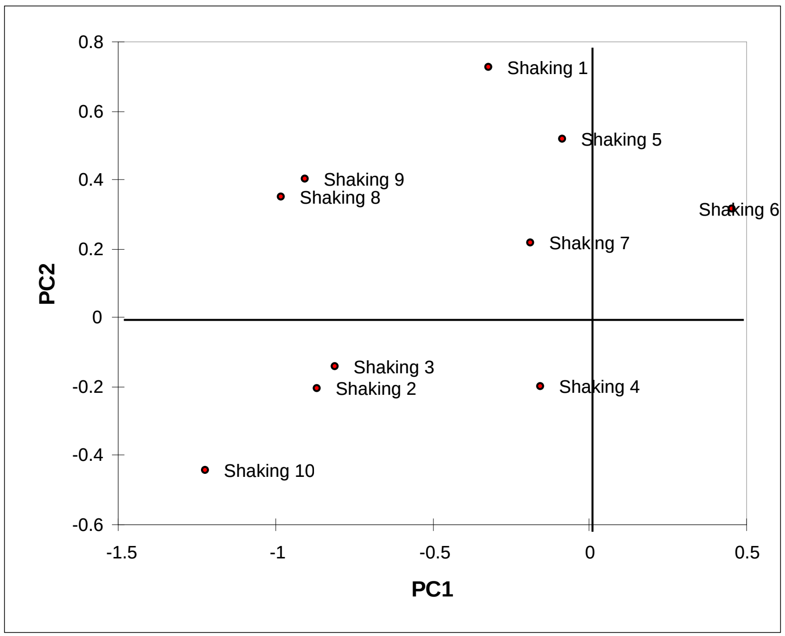
20%



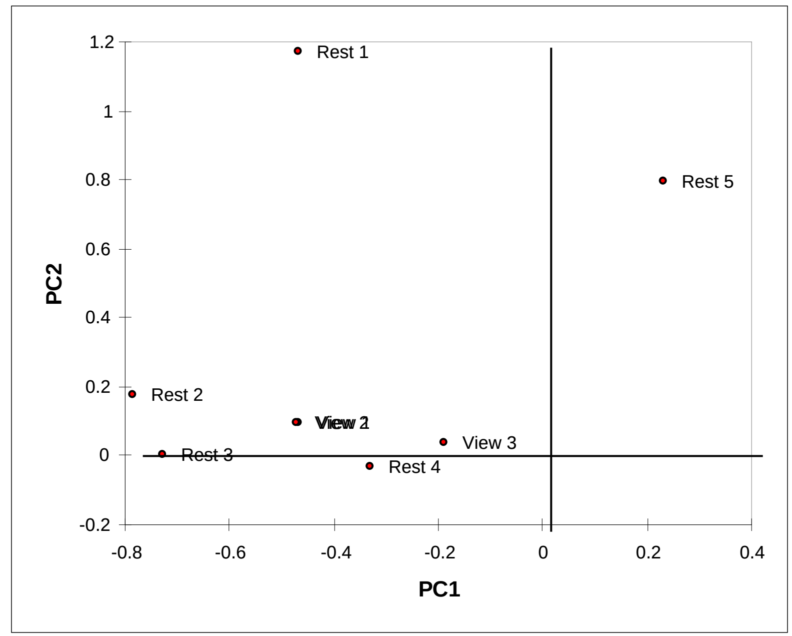
61%

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

Loadings –plots for Y and X, PCP

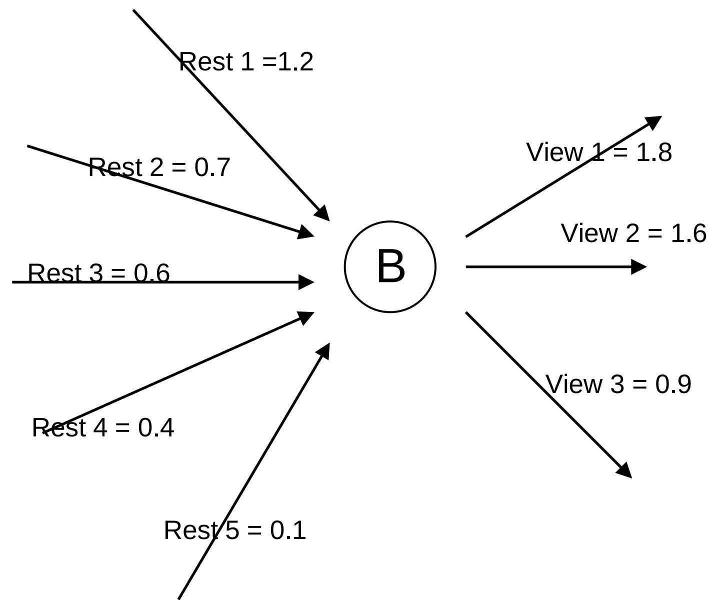


C



A, B

For one-dimensional blocks



Common variability for prediction and to be predicted?

- Block B contributes in addition to A for predicting C, but this contribution has no relation to the predicted values of B from block A.

- This shows that the part of block B that can be predicted from A has no overlap with the part of B that adds to predicting C.

  – There is more in block B that is useful than the part that can be predicted

  – SEM paradigm in this case?

Possible extensions

- Interactions and non-linearities
  - ”Simple” within this framework
    - Add extra matrices of products (like in standard PLS)
    - Or add extra marices based on principal components
    - Type I philosophy (or Type III)

- Variable selection
    - Jack-knife – technically not problem
    - Influence on validation?