

Chemometrics from a statistical perspective.

Per Bruun Brockhoff
Dep. of Mathematics and Physics
KVL

- PCA (30 minutes)
 1. Classical textbook presentation
 2. The chemometric way
 3. PCA and univariate ANOVA
 4. PCA and multivariate ANOVA
- PLS (15 minutes)
 1. Why does MLR fail?
 2. Biased regression methods
 3. PLS and statistics.

PCA - classical textbooks

- Situation: Data matrix X (n observations of a p -dimensional vector)
- Finds the linear combinations (directions) that have maximal variability.
- Pearson (1901), Hotelling (1933).

Statistics/probability:

$$X_1, \dots, X_n \sim N_p(\mu, \Sigma), \quad (\text{Independent})$$

- Multivariate one-sample situation

PCA - math

Linear combination:

$$a^t X_i = a_1 X_{i1} + \dots + a_p X_{ip}$$

The variance to be optimized

$$\text{Var } a^t X_i = a^t \Sigma a, \quad a^t a = a_1^2 + \dots + a_p^2 = 1$$

Optimize using Lagrange Multiplier technique:

$$f(a) = a^t \Sigma a - \lambda(a^t a - 1)$$

$$\frac{\partial f(a)}{\partial a} = 2\Sigma a - 2a\lambda = 0 \Leftrightarrow$$

$$\Sigma a = \lambda a$$

Definition of eigenvalue/eigenvector!

In practice: Use $S = X^t X$ instead of Σ

PCA - classical

- The eigenvalue/eigenvector problem of $X^t X$.
- An empirical description of the variance/covariance structure.
- **Not** model based:
- No well-defined statistical model for, say, a 2-factor description

- Similar to Factor Analysis(FA):

$$X_i = \mu + \Gamma f_i + \varepsilon_i, \quad \text{Var}(\varepsilon_i) = \Psi$$

$f_i = (f_{1i}, \dots, f_{ki})$ common/latent factors/variables

- FA is a well defined (normal) statistical model for the variance/covariance structure:

$$\Sigma = \Gamma\Gamma^t + \Psi$$

- **Loadings of FA:** Parameters (Γ) of the model.
- **Scores of FA:** Realizations of the random common/latent factors/variables

PCA - chemometrics

- PCA is a bilinear model: (mean corrected data)

$$X_{ij} = \sum_{a=1}^k t_{ia} p_{aj} + \varepsilon_{ij}$$

$$X = TP^t + E$$

- Description of “mean”/”fixed” structure.
- Well defined (up to rotations) statistical model
- Both loadings and scores are estimated in the model.
- PCA is a consequence of the LS-minimization:

$$\|X - TP^t\|^2$$

- PCA is the Maximum likelihood estimates of the structure under the assumption of homogeneous and independent errors:

$$\text{Var}(\varepsilon) = \sigma^2 I_p$$

PCA and univariate ANOVA

Multiplicative interaction in a 2-way setup

- Usual ANOVA:

$$E Y_{apr} = \alpha_a + \nu_p + \gamma_{ap}$$

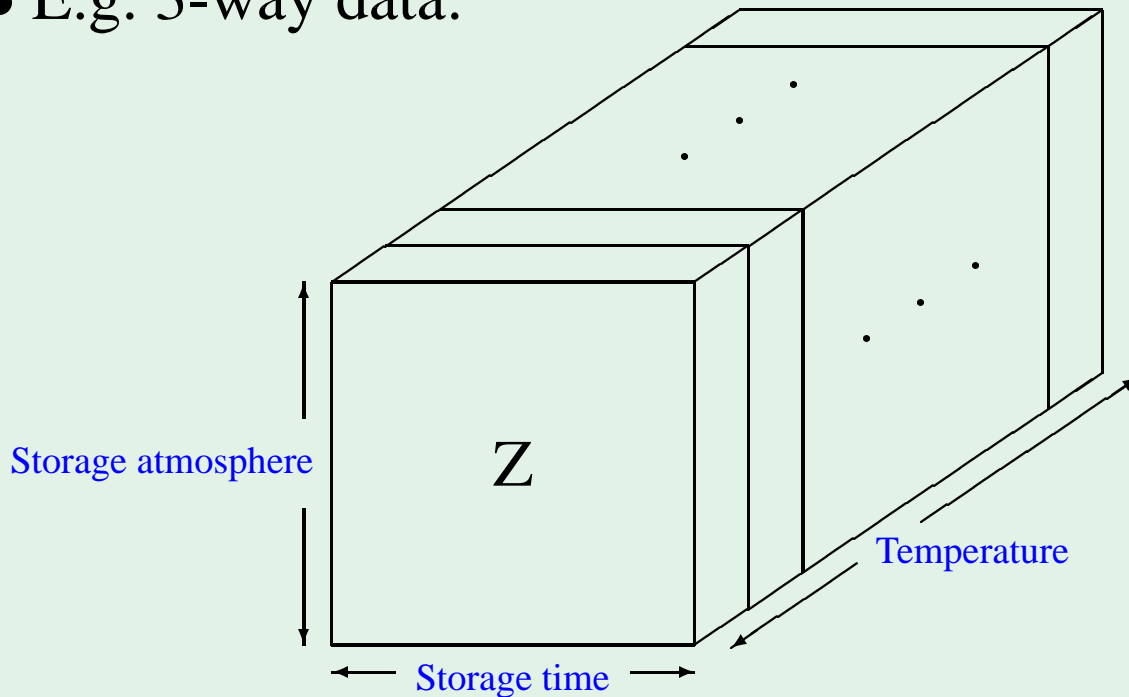
- Yates and Cochran (1938): Regression of \bar{y}_{ap} on row/column means.
- Mandel (1961, 1969, 1971):

$$E Y_{apr} = \alpha_a + \nu_p + \sum_{i=1}^m \lambda_i \delta_{ai} \kappa_{ip}$$

- AMMI-models for Genotype(variety)-by-Environment data in Plant Breeding:
 - Finley and Wilkinson (1963)
 - Kempton (1984)
 - Gauch (1988)

Multiplicative interactions in general

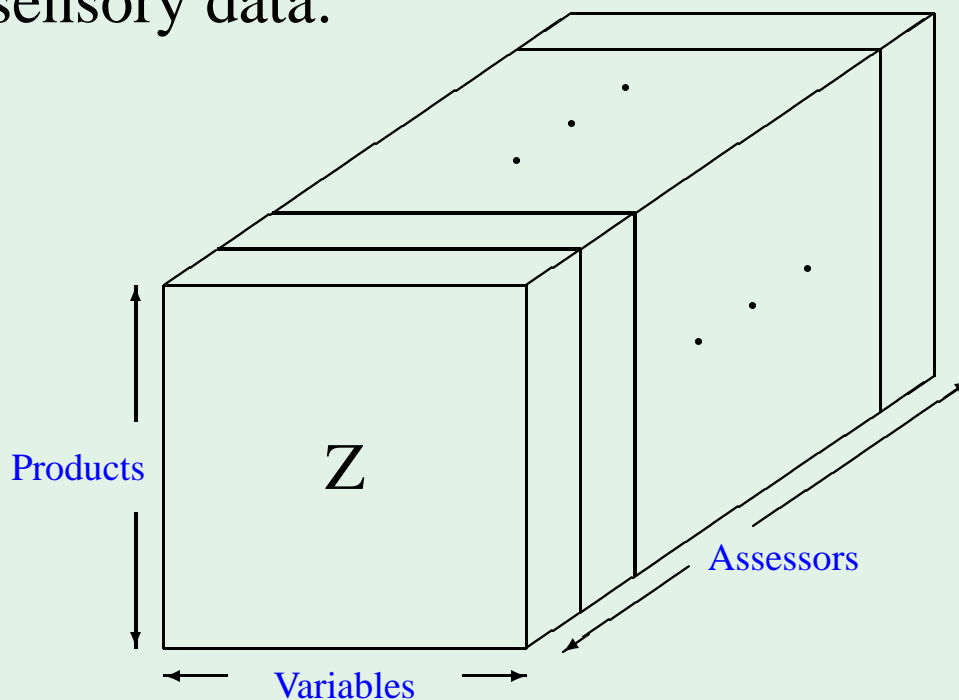
- E.g. 3-way data:



- Effects can alternatively (to classical ANOVA) be modelled as multiplicative.
- May offer easier interpretation of higher order interaction effects.
- GEMANOVA: GEneralized Multiplicative ANOVA
- Multiway PCA
- PARAFAC
- Tucker-1,2,3

Data with “replications:”

- E.g. sensory data:



- Multivariate ANOVA situation (MANOVA)
- Classical hypothesis:(say) H_0 : The objects mean structure lies in a 2-dimensional hyperspace
- So this is “PCA structure” for the object averages.
- The analysis/test of this hypothesis: Canonical Variates Analysis (CVA)
- If the error is assumed homogeneous and uncorrelated then: CVA=PCA.

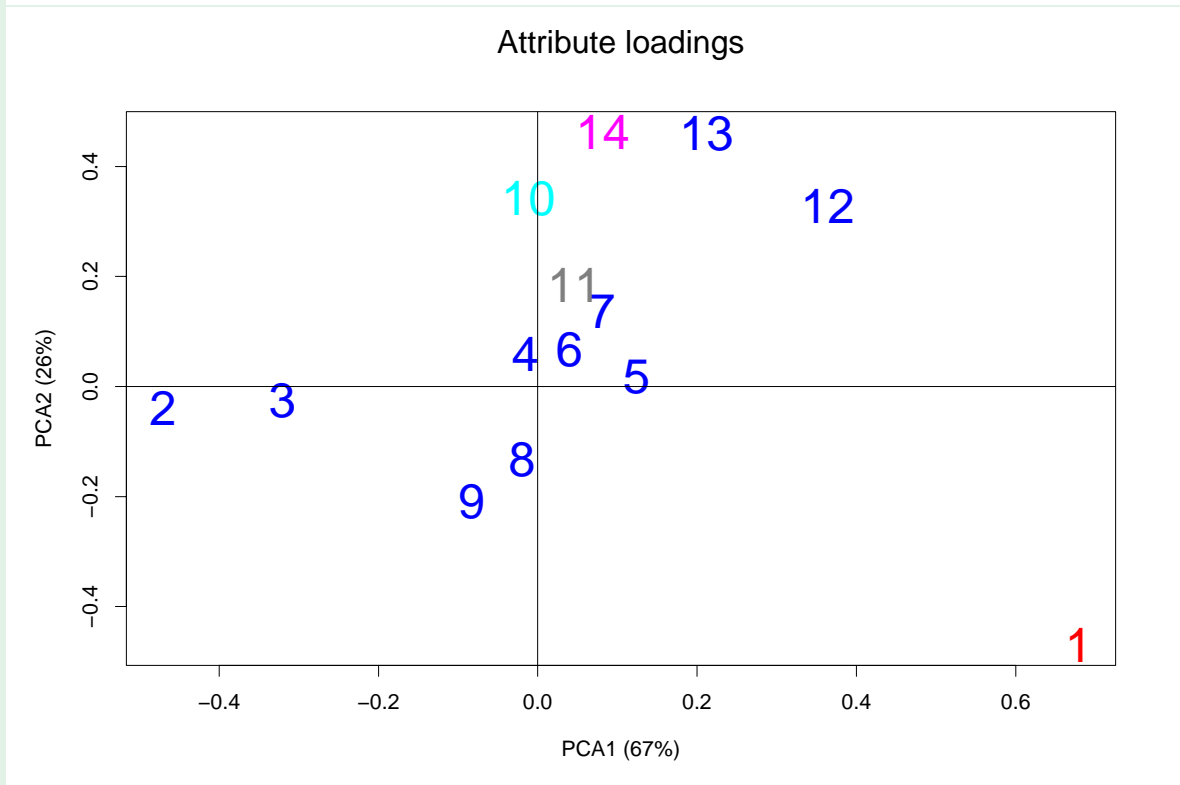
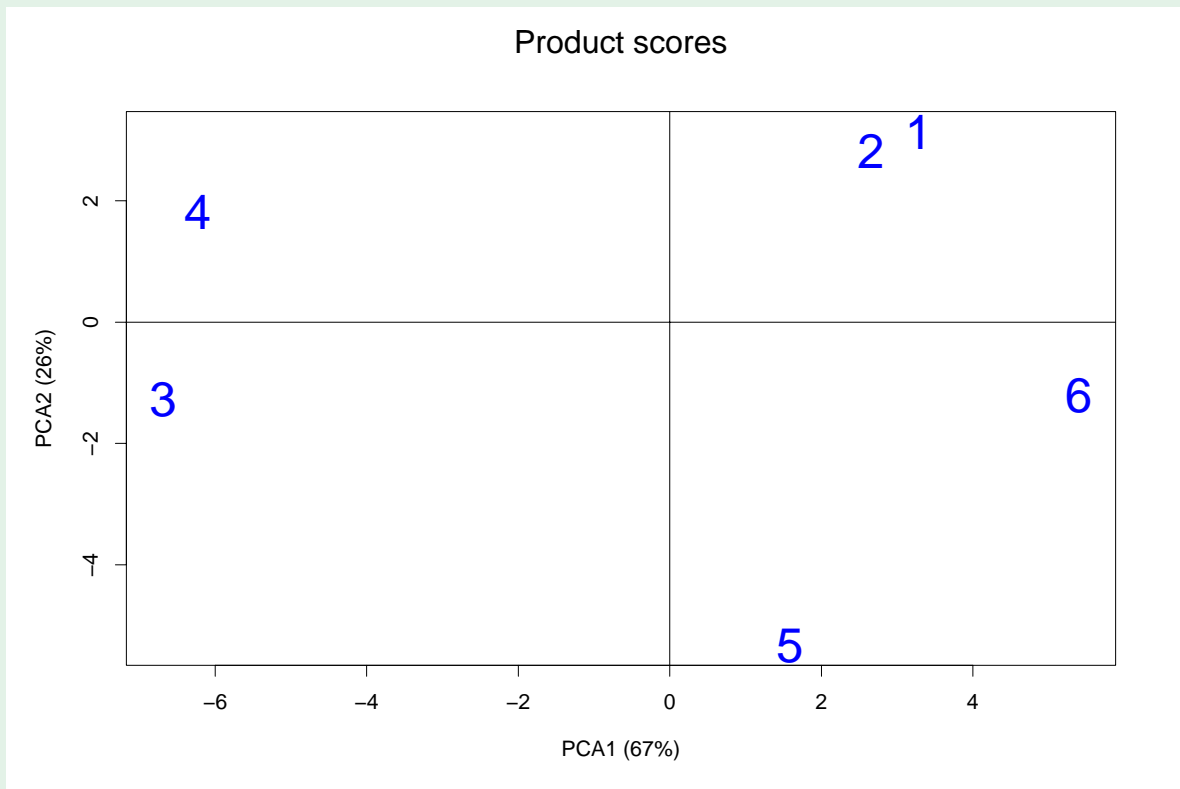
Example:

- 6 products
- 3 replications
- 9 assessors
- 14 attributes

Typical analysis:

- PCA of 6-by-14 matrix of product averages.

PCA results for example



Principal Component Analysis (PCA)

PCA looks at $S_{product}$

$S_{product}$ is the product averages variance-covariance matrix.

Loadings = The eigenvectors of $S_{product}$

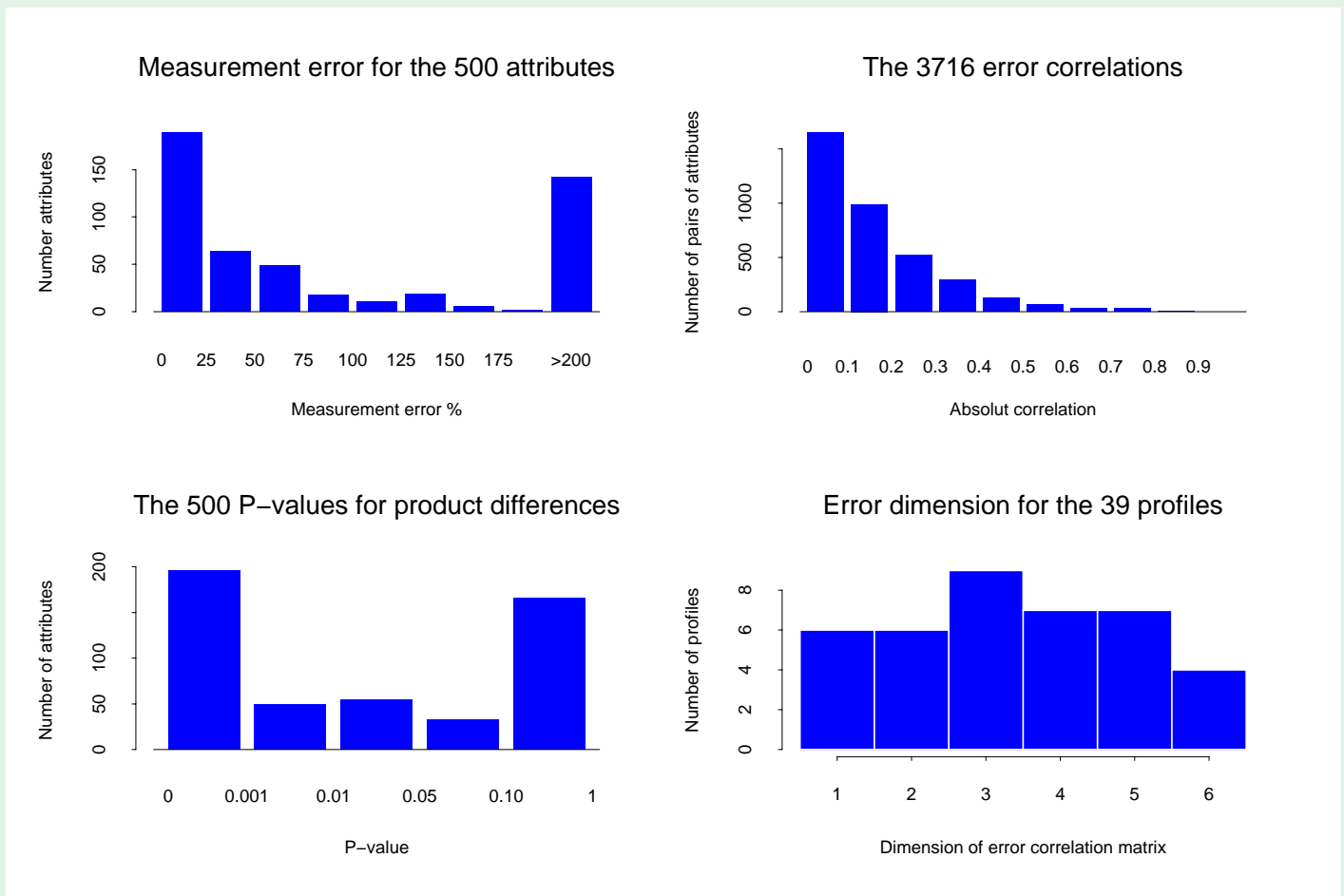
Scores = Product projections on the subspace

Problem of averaging

The error information is ignored

leading to inefficient and possibly biased (wrongful) analysis.

Measurement errors for 500 sensory attributes



Meta study of 39 sensory profiles

- A large proportion of sensory attributes has considerable measurement error.
- The measurement errors are of quite different size.
- For non of the 39 profiles the error correlation matrix was independent.

Error variability

$$E(S_{product}(p, p)) = \text{Var}_{\text{true}}(\text{product}) + \text{Var}(\text{error})$$

where

$$\begin{aligned} \text{Var}(\text{error}) = & \frac{1}{9} \text{Var}(\text{product} \star \text{assessor}) \\ & + \frac{1}{3} \text{Var}(\text{product} \star \text{replication}) + \frac{1}{27} \text{Var}(\text{residual}) \end{aligned}$$

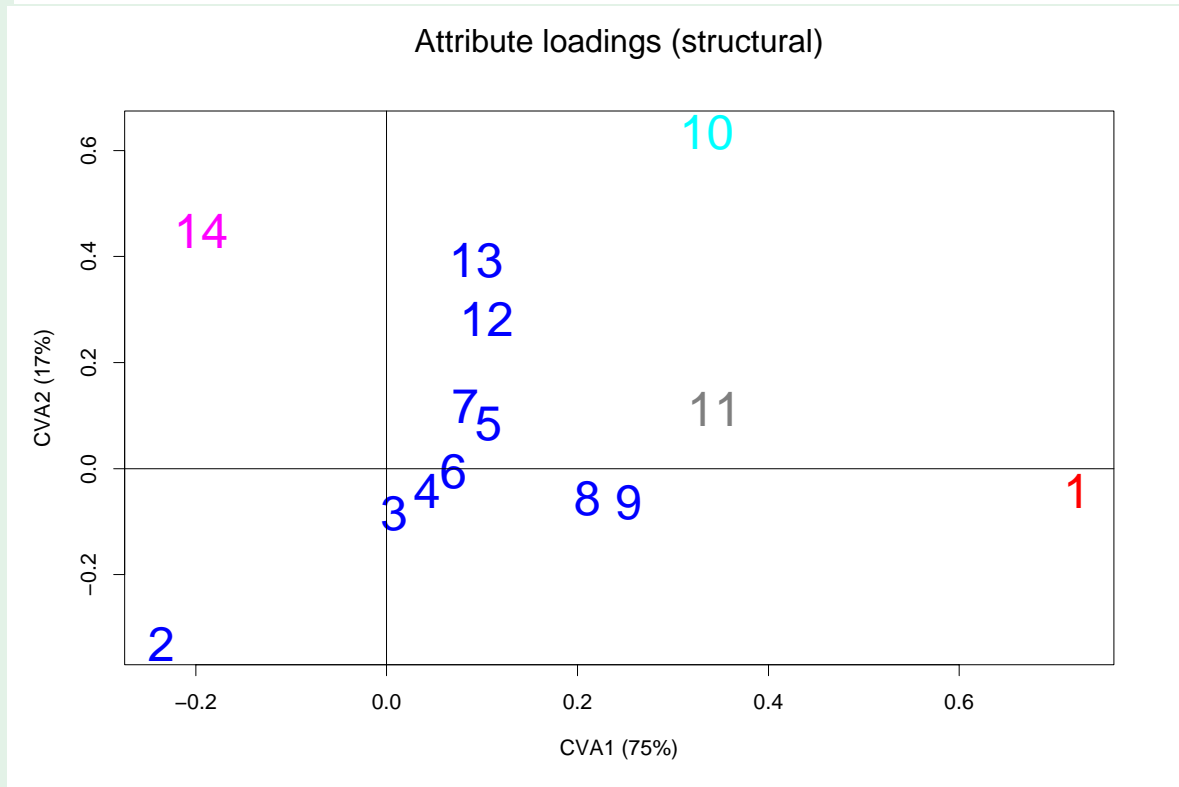
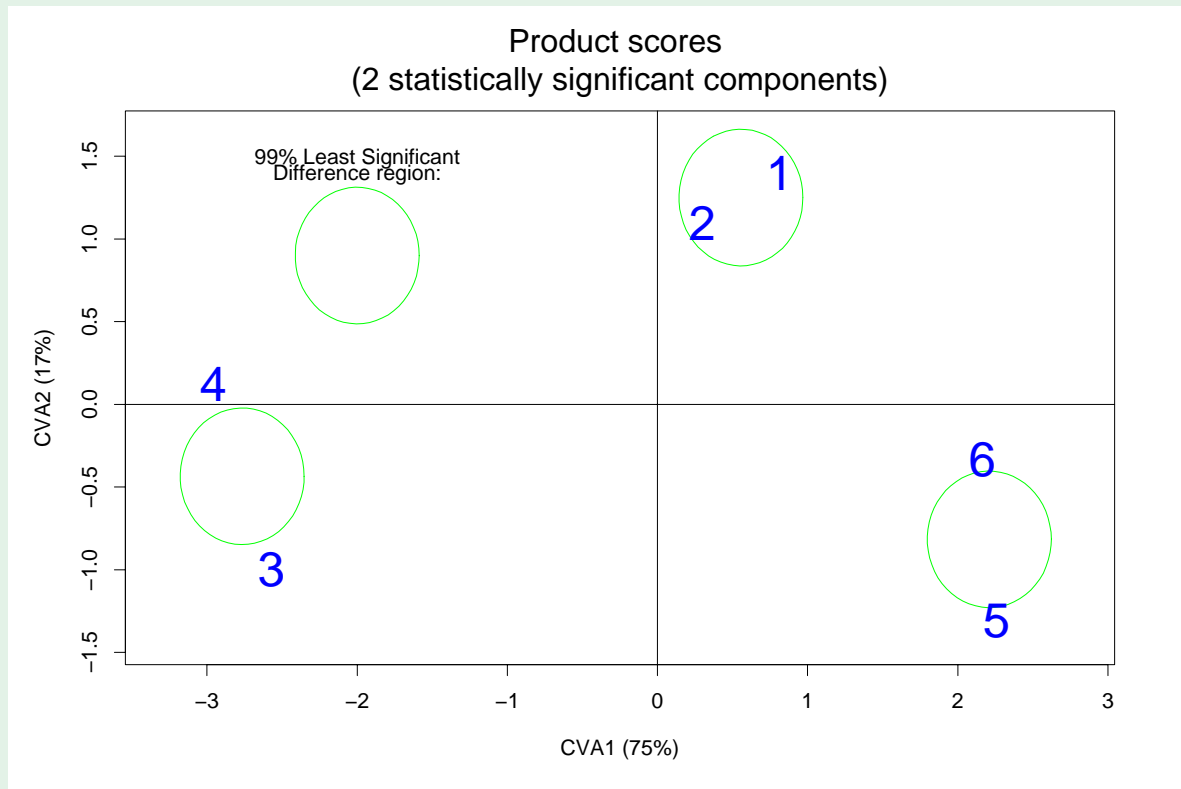
Less wellknown:

Exactly the same structure holds for the covariance (off diagonal) elements of $S_{product}$ also.

$$S_{product} \text{ is perturbed by } S_{error}$$

where S_{error} is the error variance-covariance matrix.

CVA results for example



Canonical variates analysis(CVA)

CVA looks at $S_{error}^{-1}S_{product}$

$L_{discrim}$ = The eigenvectors of $S_{error}^{-1}S_{product}$

Scores = Projections on $L_{discrim}$.

Structural loadings = $S_{error}L_{discrim}$

PCA/CVA/chemometrics

- CVA is an analogue of PCA that takes uncertainty and error correlations into account.
- Is in analogy with:
 - Generalized Least Squares (GLS) versions of PCA recently promoted by Martens *al.* (2001).
 - Maximum likelihood PCA
- Often in chemometrics: “structural interferences” are more important than “random noise” (at least the measurement error)

PLS and statistics

Situation: one y (or more), several x 's:

$$y, x_1, \dots, x_p$$

Multiple Linear Regression(MLR):

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

$$y = X\beta + \varepsilon$$

Estimation:

$$\hat{\beta} = (X^t X)^{-1} X^t y$$

Error in estimation:

$$\text{Cov}(\hat{\beta}) = \sigma^2 (X^t X)^{-1} = \sigma^2 \sum_{i=1}^p (l_i l_i^t) / \lambda_i$$

λ_i eigenvalues for $X^t X$

l_i eigenvectors for $X^t X$

Problem of MLR

- Requires $n > p$.
- x-correlations, near multicollinearity:
 - Some very small eigenvalues
 - Error in estimation “explodes”.

For any predictor \hat{Y} :

Mean Prediction Error = Estimation Error

+ Natural Variability + Squared Bias

$$MSE(\hat{Y}) = \text{Var}(\hat{Y}) + \text{Var}(Y_{new}) + \text{Bias}^2$$

- MLR-predictor is “unbiased” ($\text{Bias}^2 = 0$) BUT $\text{Var}(\hat{Y})$ may be extremely large.
- Solution: Chose a predictor with some bias but much smaller estimation error.
- Variance/bias trade-off
- Minimize MSE by cross validation.
- Same principle used in e.g. non-linear smoothing techniques/ non-parametric regression.

Biased regressions

PLS is one choice of such a biased regression

Other possible choices

- Variable Selection techniques
 - Ridge Regression
 - * Use $(X^t X + kI)^{-1}$ instead of $(X^t X)^{-1}$.
 - * Has a Bayes interpretation
 - Use only a subset of the x-variables
- Principal Component Regression (PCR)
 - Use some of the x -PCA-components to predict y .
- Continuum Regression:
 - Includes MLR, PCR and PLS as special cases (3 points on a continuum of possible choices)
- Generalized Ridge Regression
 -
 -

PLS1 and Statistics ?

(Helland (2001). *Chem. Int. Lab. Syst.*)

- What does the PLS1 algorithm really calculate?
- Or similarly: What is the “projection space”?
 - Solved, the Krylow sequence:
$$s, Ss, \dots, S^{k-1}s, S = X^t X, s = X^t y$$
 - A theoretical basis for the practical fact: PLS1 and PCR performs almost equal, but PLS1 with fewer components.
- What is the statistical model/parameter setting?
 - Solved (at least one version has been given)
- Is the sequence of PLS1 population models theoretically meaningful?
 - Yes!

PLS1 and Statistics ?

(Helland (2001). *Chem. Int. Lab. Syst.*)

- Theoretical comparison between the various bi-ased regression methods
 - Something is done:
 - All methods are **shrinks** the OLS-solution
 - This shrinkage structure has been investigated
 - PLS1 is the only one that is not a "true" shrink-age method, i.e. it is NOT optimal.
- Comparisons of prediction ability
 - Frank and Friedman (1993): Simulation gives that Ridge Regression is (slightly) superior.
 - Critized by Svante Wold.
- Does the PLS1-algorithm give the optimal solu-tion/estimate for the PLS1 population model?
 - No!!
- The formal handling of **model reduction** is often ignored by applied statisticians.

A traditional comparison:

Statistics	vs. Chemometrics
Hard modeling	Soft modeling
Hypothesis testing (Deductive)	Explorative (Inductive)

My point of view:

