

# **PLS regression for analyzing two or more data tables: An overview of different approaches.**

**Frank Westad**

**Norwegian Food Research Institute, Norway**

**Sensometrics 2004**

**July 29, 2004**

# Outline

- ◆ Introduction (PLSR reminder)
- ◆ Preference mapping - external/internal (reminder)
- ◆ Down- and up-weighting of variables
- ◆ L-PLSR
- ◆ Orthogonal PLS regression
- ◆ Summary
  
- ◆ **What is not in here:**
  - Multiblock/Hierarchical PLSR (Wold, McGregor)
  - Multi-way PLSR (Bro)
  - Non-linear PLSR (Høskuldsson, Wold)
  - Logistic PLSR (Tenenhaus, Quannari)
  - Path PLS (Tenenhaus)

# PLS regression - reminder

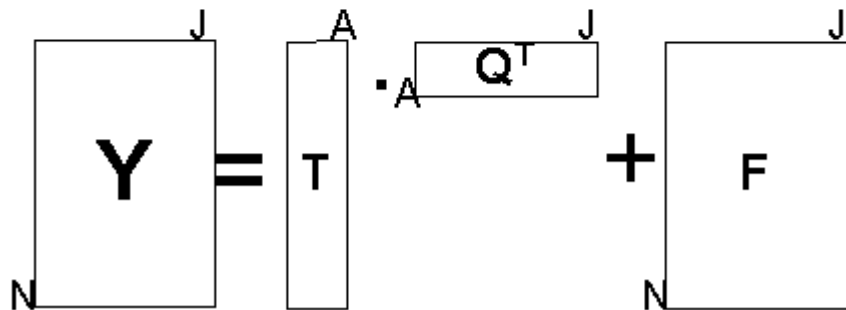
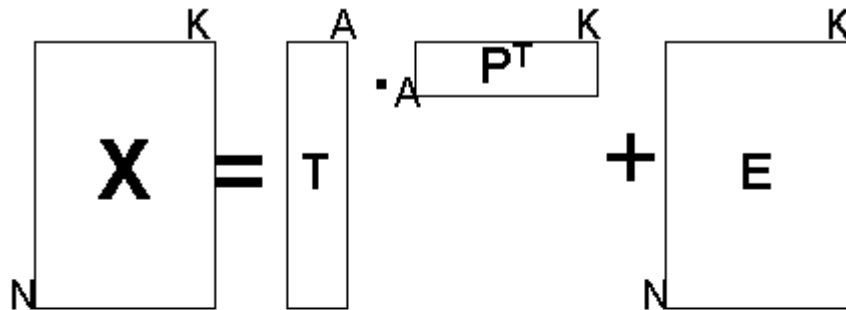
- ◆ There is nothing “magical” about PLS regression
- ◆ With one dependent variable, there should ideally be only one PLS-component
- ◆ The loading weight vector,  $w_1$ , for component #1 reflects the covariance between individual x-variables and y
- ◆ We’re still missing a straight forward formula for the degrees of freedom “eaten” for each component
  - Depends on the structure of X
  - E. g. when all eigenvalues in X are identical PLSR “eats” everything i the first PC (e.g. X = central composite design)
- ◆ When all variance in X is related to Y, PCR and PLSR are identical
- ◆ PCR and PLSR are identical to MLR (OLS) for a full rank model (but remember centering & scaling when interpreting)

# 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(X,Y))

As a linear regression model:

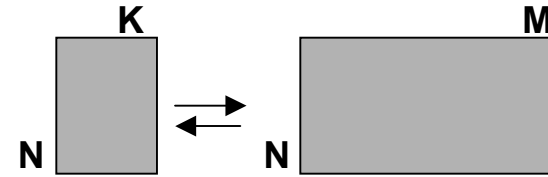
$$Y = XB + F$$

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

# Data structures in preference mapping

- ◆ Descriptive sensory data for a number of samples

- N\*K matrix of data (N samples and K attributes)
- Intensities, scores on a hedonic scale (typically 1-9)



- ◆ Consumer preferences for the same samples

- N\*M matrix of data (N samples, M consumers)
- Preferences (degree of liking, preference, purchase intent etc.)

- ◆ PCR was earlier the most used method:

- **MDPREF**, PCA of consumer data. The sensory attributes are regressed on the principal components. **Internal** preference mapping
- **PREFMAP**, PCA of sensory data. The consumers are regressed onto the principal components. **External** preference mapping

- ◆ With PLSR the two approaches are more alike (McFie, Sensometrics 2004)

# Detour: Scaling of consumer and sensory data

- ◆ **Should one scale the data to unit variance or not?**
  - Assume variables on a scale from 1 - 9
  - If not scaled, the variables with large variance will dominate
  - If scaled, then small numerical differences might (erroneously) influence on the result (accidental correlations when many consumers & few products)
- ◆ **The correlation loadings are the correlations between the variables and the PC's**
- ◆ **Invariant of the choice of scaling (but component's directions might change)**

$$r_{ka} = p_{ka} \sqrt{\mathbf{t}_a^T \mathbf{t}_a} / \sqrt{\mathbf{e}_{0,k}^T \mathbf{e}_{0,k}}$$

PCA model:  $\mathbf{X} = \mathbf{TP}^T + \mathbf{E}$

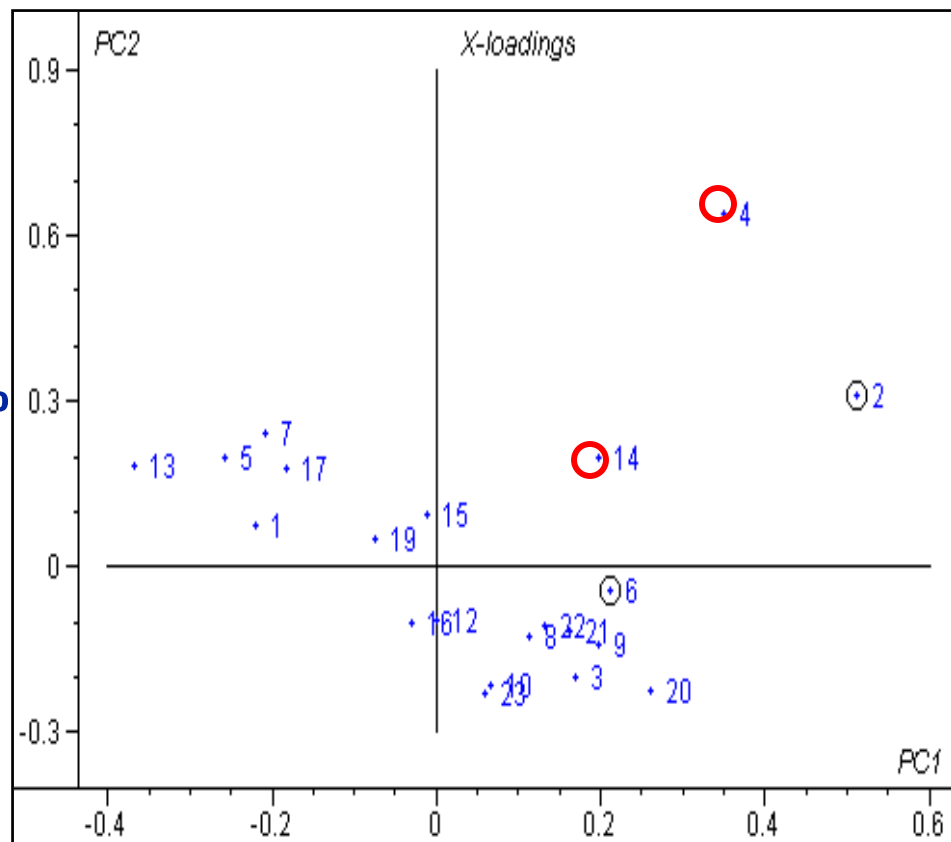
How much is explained in PC a?

Variance before modeling starts

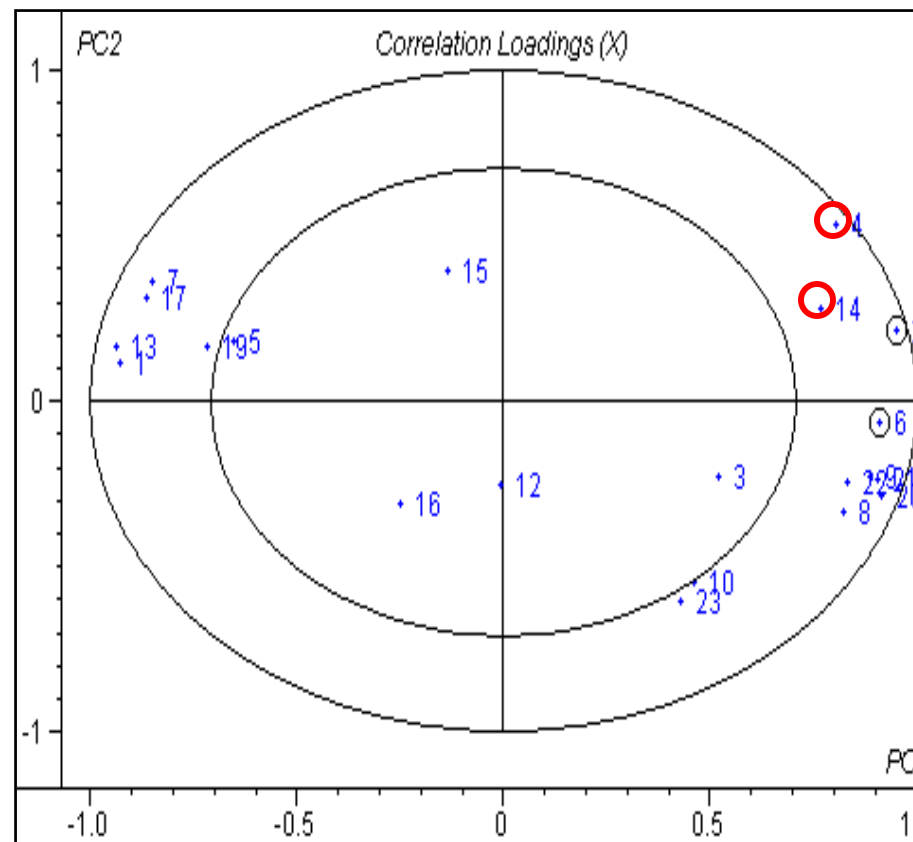
# Example - PCA on sensory descriptive data

- ◆ **Product: Ciabatta**
- ◆ **88 samples generated from an experimental design with**
  - **11 different meals**
  - **Yeast (2 levels)**
  - **Proofing time (90; 135 minutes)**
  - **Each experiment was replicated**
- ◆ **23 sensory attributes**

# Ciabatta: PCA on unscaled data

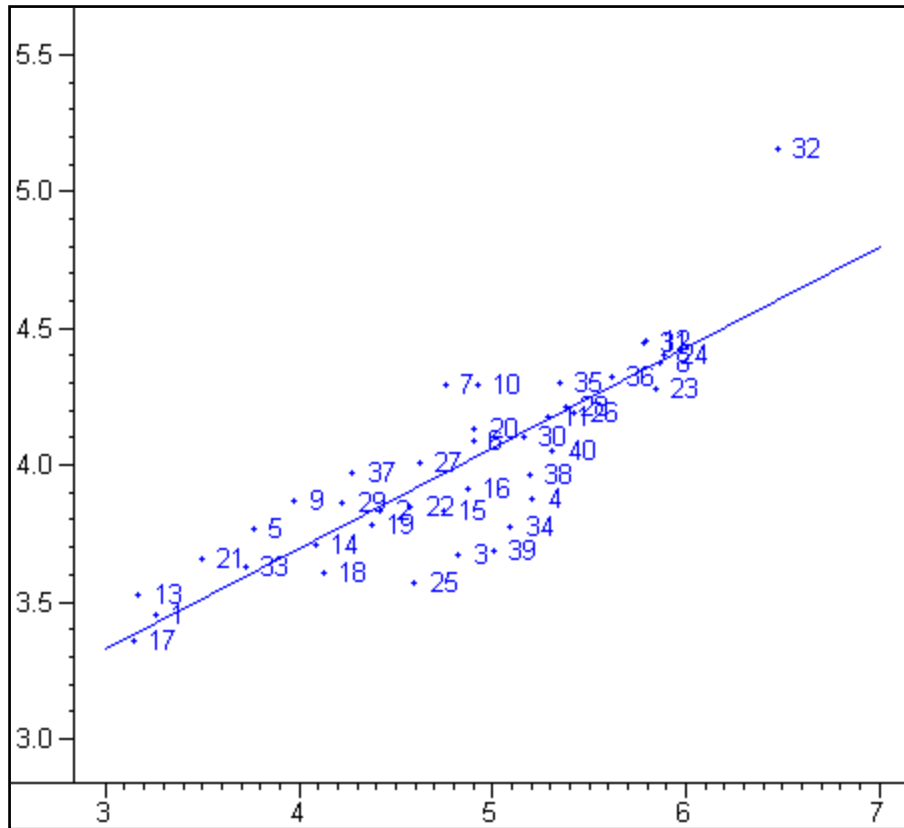


53 %



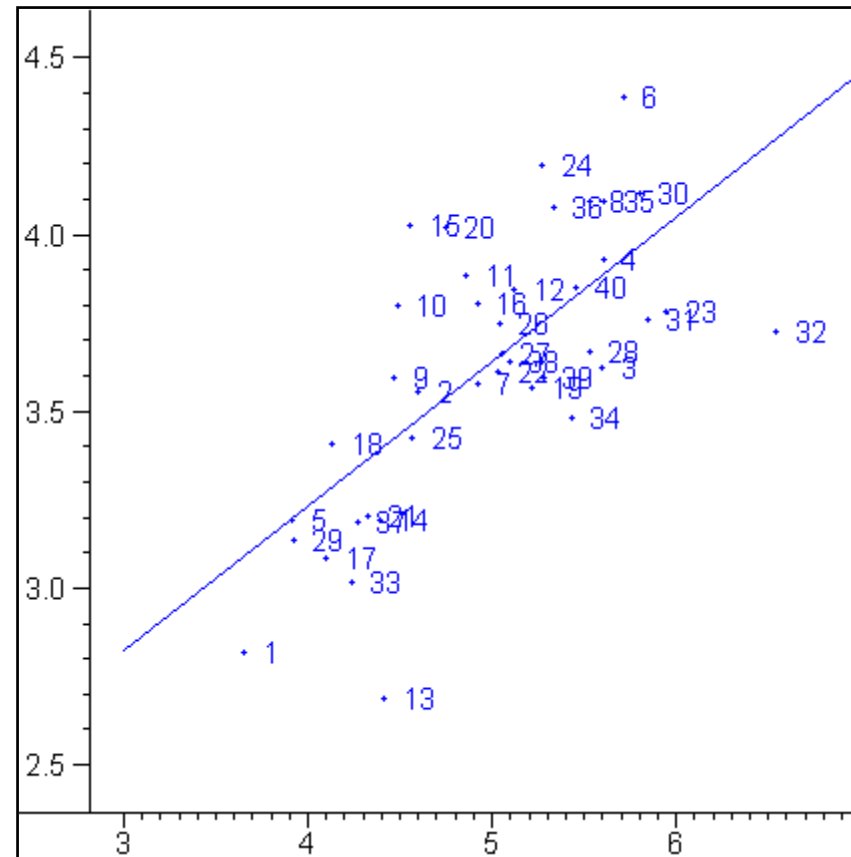
# Ciabatta: Raw data

Variable 0



Variable 2

Variable 14



Variable 4

# Validation

- ◆ **Validation is essential in all scientific work**
  - **Avoid overfitting and wrong interpretations**
- ◆ **External: Hypothesis driven**
  - **Repeat study over time, different countries etc.**
- ◆ **Internal: Data driven. Two types are often applied:**
  - **Cross-validation**
  - **Test set validation**
- ◆ **Bootstrapping/jack-knifing are general tools for estimating uncertainty for model parameters (ref. J.-F. Meullenet, Session 5)**
  - **Can also be applied for PCA**  
⇒ **t-test on individual variables on each PC**

# Uncertainty estimates

The variance of the model parameters may be estimated by jack-knifing  
Example: Regression coefficients,  $b$

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

Main model      Sub-model

$M$  = the number of segments

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

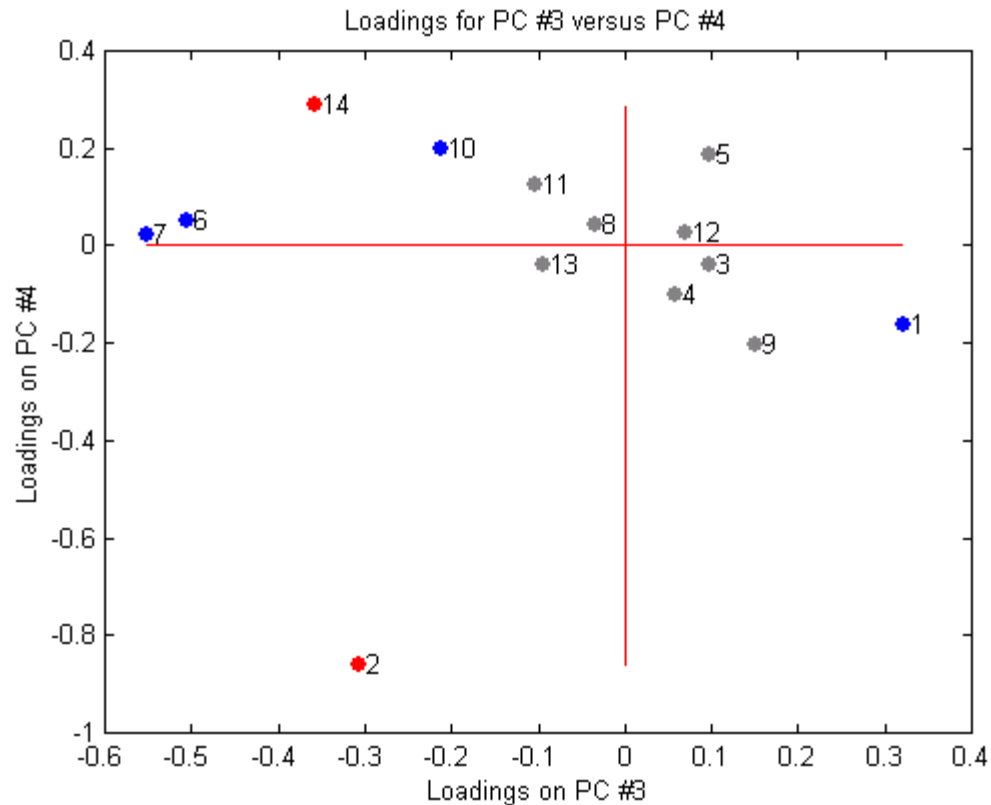
*Many tests - Adjusted Bonferroni is an option*

# Reflection of models is needed

- ◆ When the structured noise components (e.g.  $P_{Y_{osc}}$ ) are extracted during cross-validation, their orientation may be mirrored. Thus, in order to sum up the variances in a meaningful way, they must be flipped (and sorted)
- ◆ Some approaches:
  - Procrustes rotation
  - Flip and order vectors based on correlation between main and sub-models

# Loading plot with significance

Consumer data - eating habits (103 x 14) - PCs 3,4



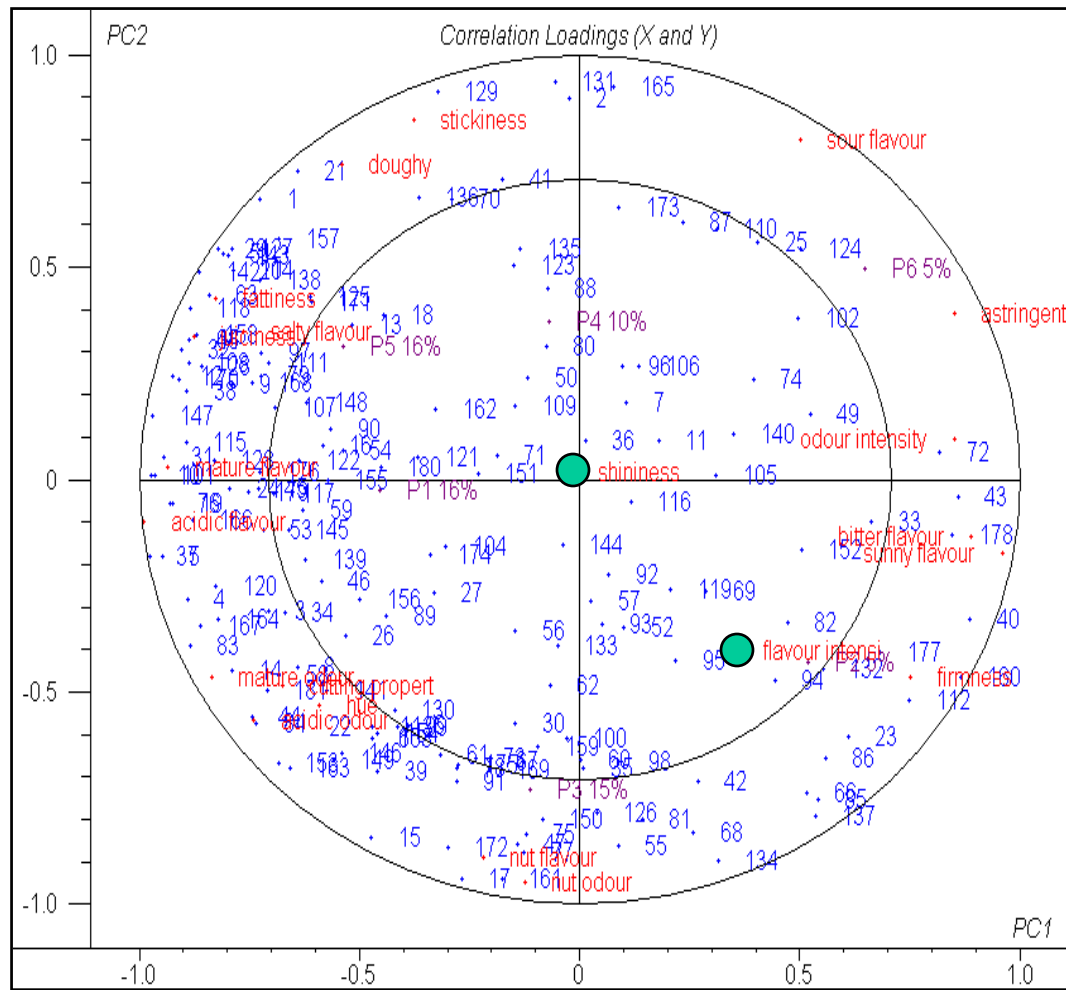
**..end of detour...back to**

# Example - External vs. Internal preference mapping

- ◆ 6 commercial semi-hard cheese products
- ◆ Preference for 181 consumers
- ◆ 26 sensory attributes



# Internal preference mapping



**28 % , 42 %**

**37 % , 48%**

# PLS regression with binary data

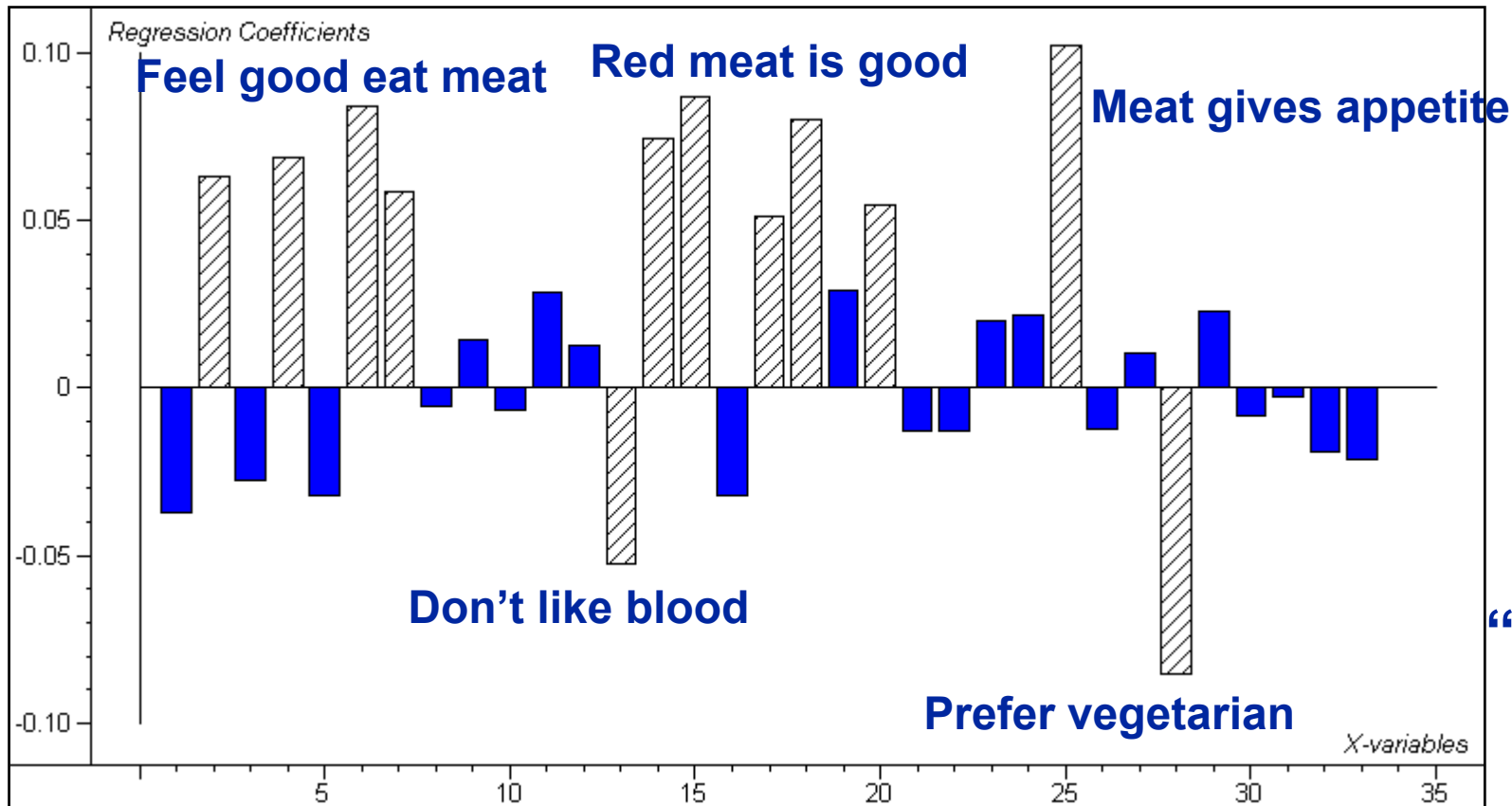
- ◆ ANOVA PLSR
- ◆ Discriminant PLSR - classes as 0/1 variables
- ◆ Objects as dummy variables as an alternative to biplot
- ◆ Remove specific effects in an “inverse” model
  - Remove unwanted variation by generating dummy Y-variables
    - ◆ remove effect of replicates in sensory data in the ciabatta example
    - ◆ keep the residual X-variance for further modeling

# Discriminant PLS regression

- ◆ **Case: Young people's attitudes towards meat**
- ◆ **Which attitudes are related to gender?**
- ◆ **180 consumers age 16-29**
- ◆ **Gender as 0/1 response variable (Y)**
- ◆ **Attitudes (34) as predictors (X)**

# PLS regression - gender

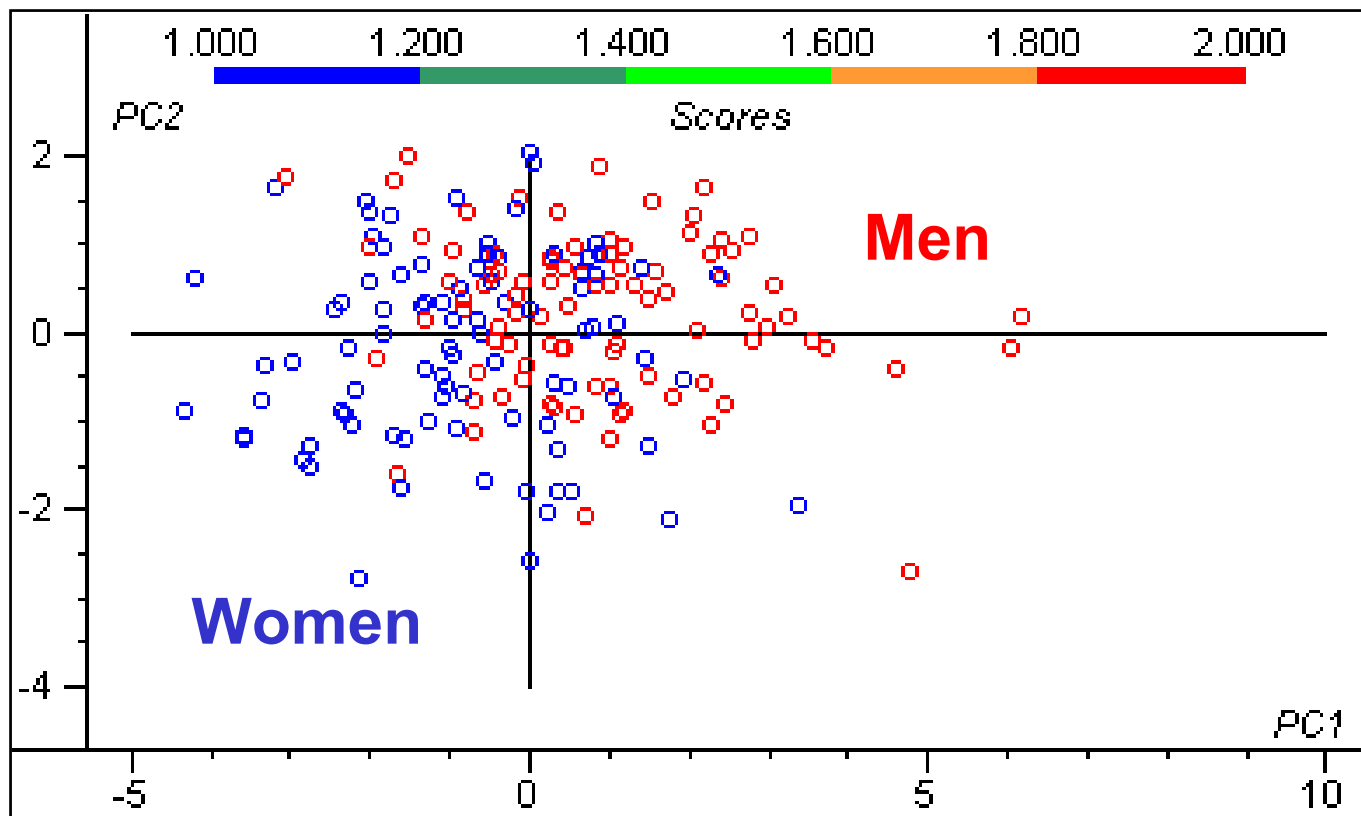
## Significant variables from jack-knife estimates



“Maleish”

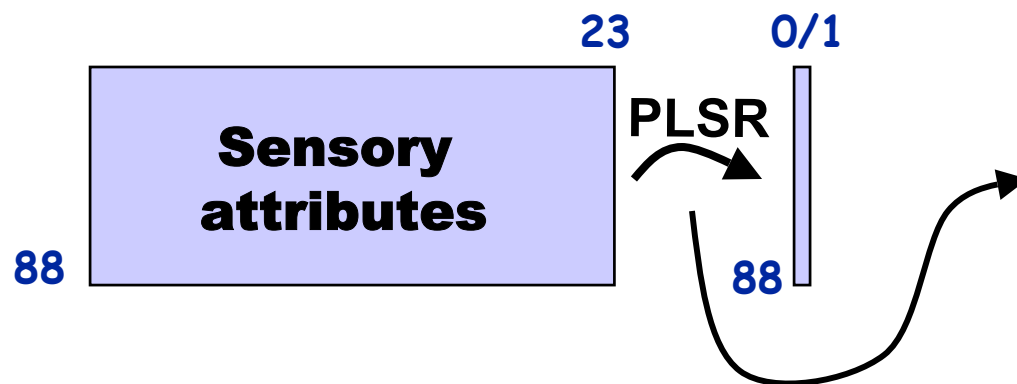
“Femaleish”

# Score plot - how good is the discrimination?



# Removing effect of replicates

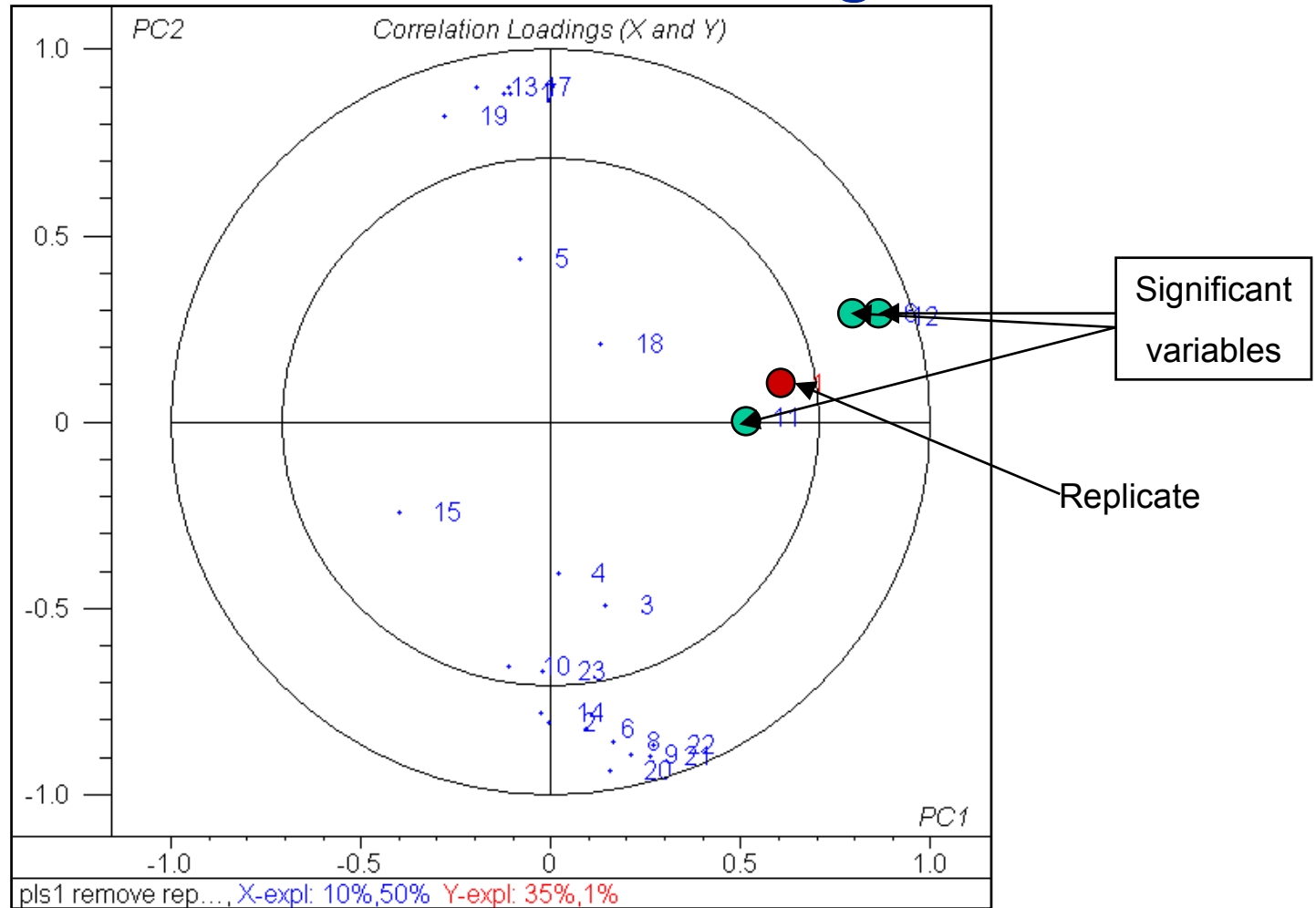
- ◆ Product: Ciabatta
- ◆ 88 samples generated from an experimental design with replicates
- ◆ 23 sensory attributes
- ◆ PLSR with replicate as response variable
  - Decide on the optimal number of components
  - Keep X-residual, **E**, for further modeling



$$\begin{matrix} N & & K \\ \boxed{X} & = & \begin{matrix} A \\ T \end{matrix} \cdot \begin{matrix} A & K \\ P^T & \end{matrix} + \boxed{E} \\ & & & & N & & K \end{matrix}$$
$$\begin{matrix} N & & J \\ \boxed{Y} & = & \begin{matrix} A \\ T \end{matrix} \cdot \begin{matrix} A & J \\ Q^T & \end{matrix} + \boxed{F} \\ & & & & N & & J \end{matrix}$$

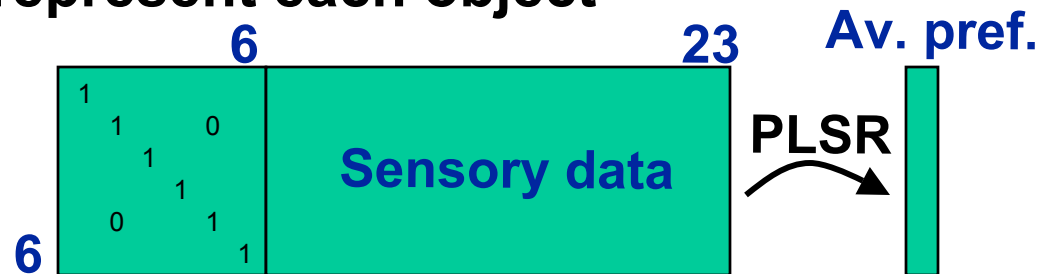
The matrix  $E$  in the first equation is highlighted with a red border.

# Correlation loadings



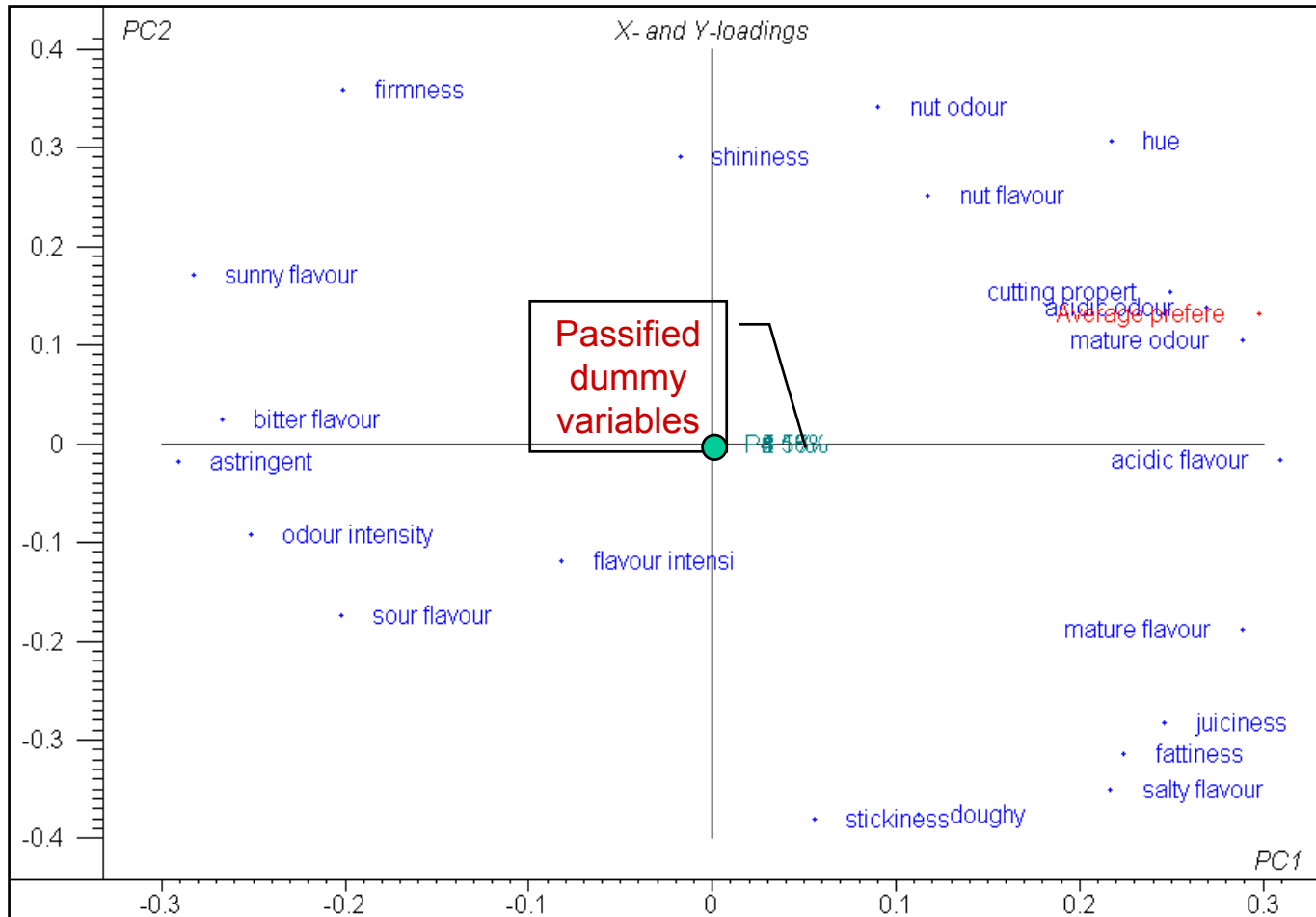
# Objects as dummy variables

- ◆ Case: Cheese data with average preference as response variable
- ◆ Augment sensory data with an identity matrix to represent each object



- ◆ Downweigh (passify) these variables
- ◆ Alternative to biplot
  - Different ways to scale scores and loadings by eigenvalues (Gabriel, Gilbert poster #9)
  - But it is still a vector model (Ennis, Sensometrics 2004)

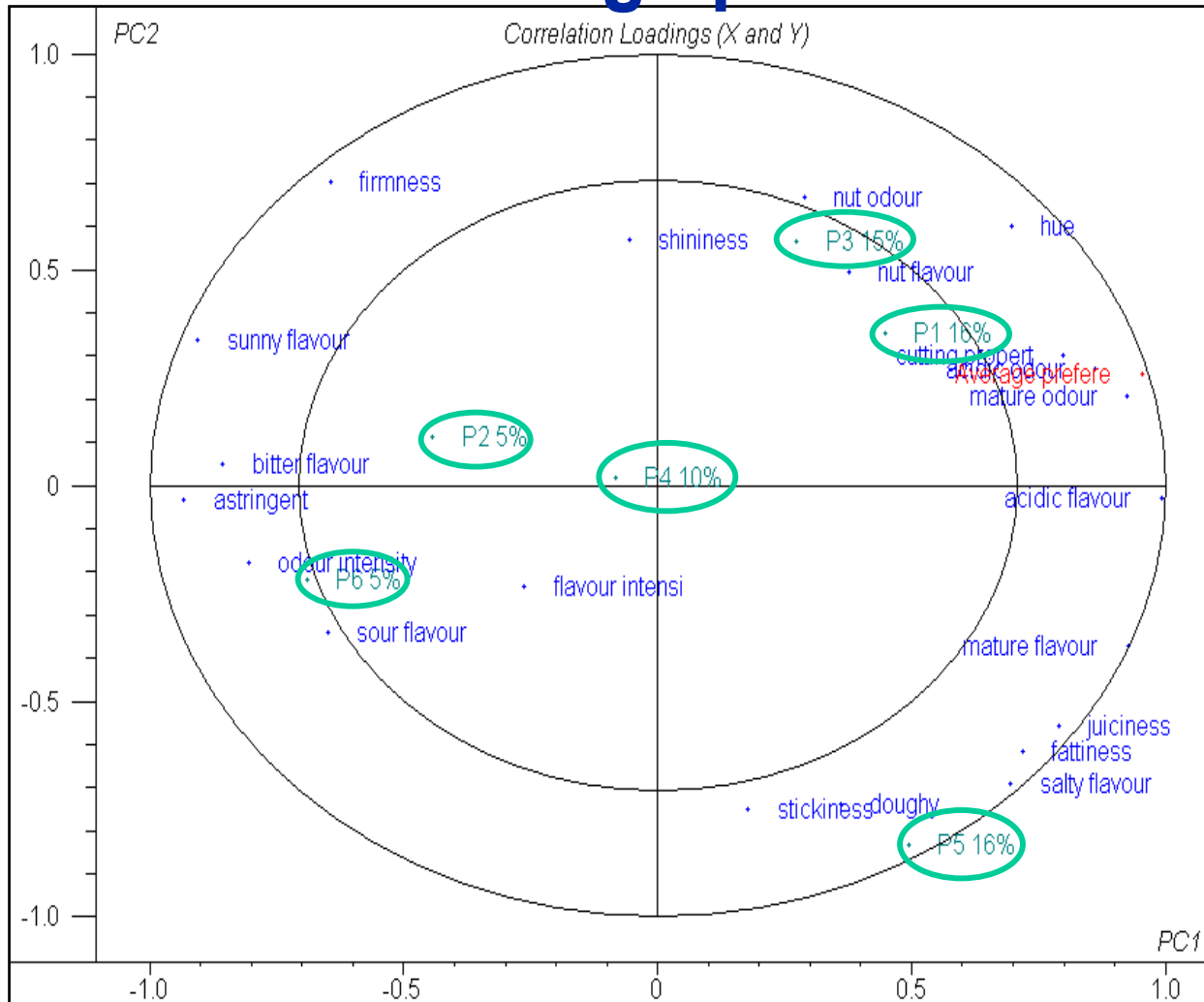
# Cheese data - average preference mapping



**3 % , 7%**

**50 % , 91%**

# Cheese data - average preference mapping



23 % , 7%

50 % , 91%

Products' correlation and direction interpretable correlation loadings plot

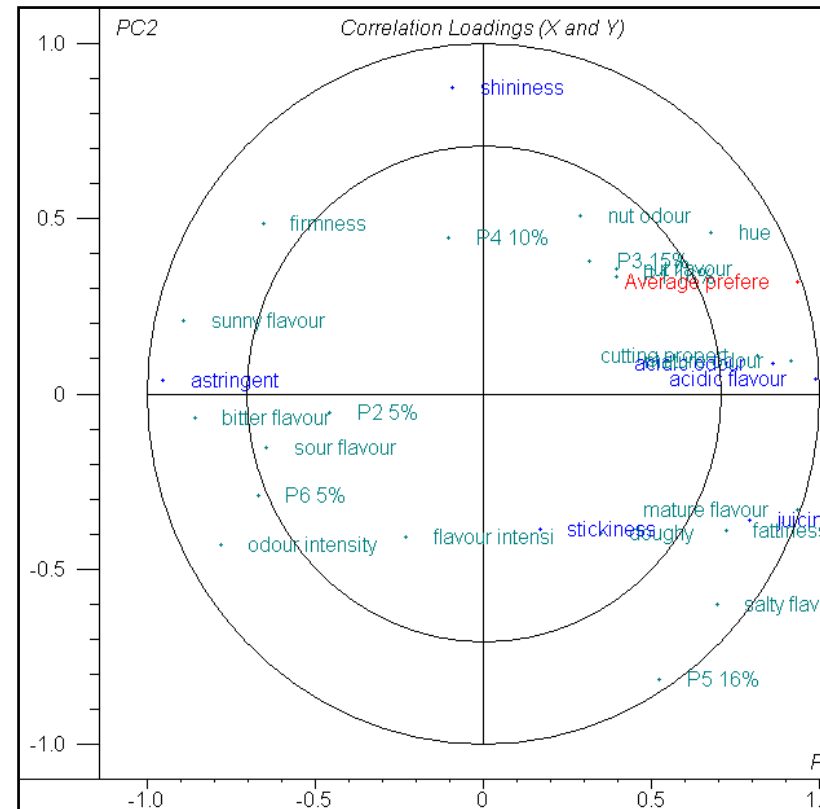
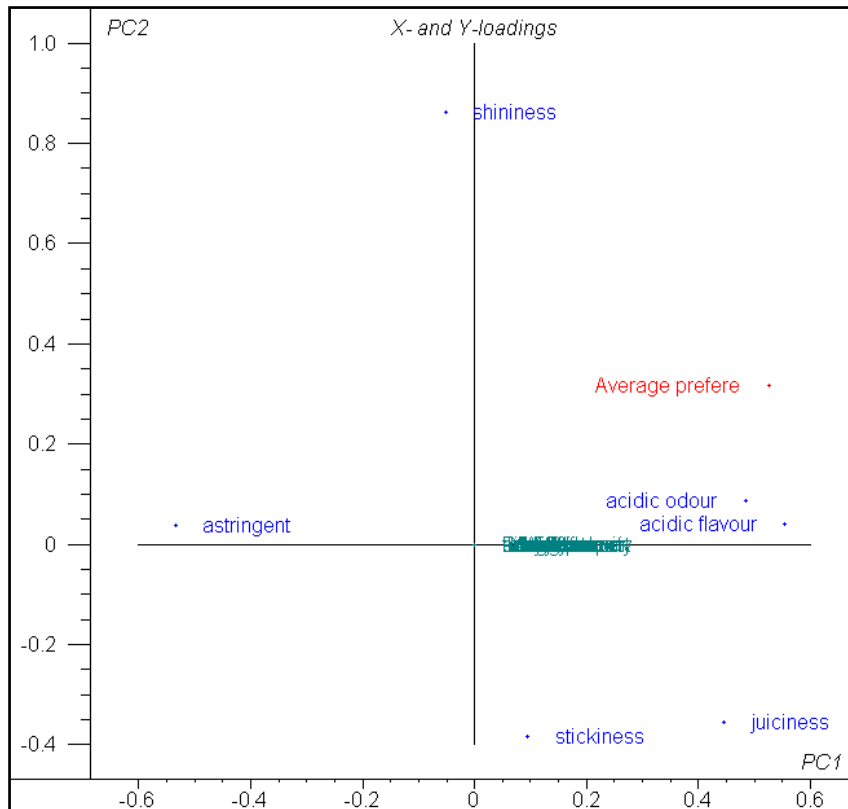
# Variable selection/subset selection in regression

- ◆ **The first thing to ask is: what is the objective?**
  - **Remove “useless” variables to simplify model interpretation**
    - ◆ E.g. find which sensory attributes that are relevant in preference mapping
  - **Improve prediction ability**
    - ◆ E.g. “Find the best combination of GC compounds in wine to model specific sensory attributes”
- ◆ **Some methods**
  - **Stepwise regression/Best combination search**
  - **Principal variates ( $\max(X'yy'X)$ )**
  - **Genetic or evolutionary algorithms**
  - **Uncertainty estimates from resampling methods (e.g. Jack-knifing or bootstrapping, ref. J.-F. Meullenet, Session 5)**
- ◆ **... but are there alternatives to removing variables (weight = 0)?**
  - **Give variables low or high weights depending on the objective**
  - **Interpret their correlation loadings and thereby possible impact in the model**

# Illustration of downweighing - cheese data

- ◆ **Assume a variable selection procedure selected variables:**
  - shininess
  - astringent
  - acidic flavor
  - acidic odor
  - juiciness
  - stickiness
- ◆ **Run a regression with other variables passified**
- ◆ **Visualize as correlation loadings**

# Don't remove - keep them up your sleeve!

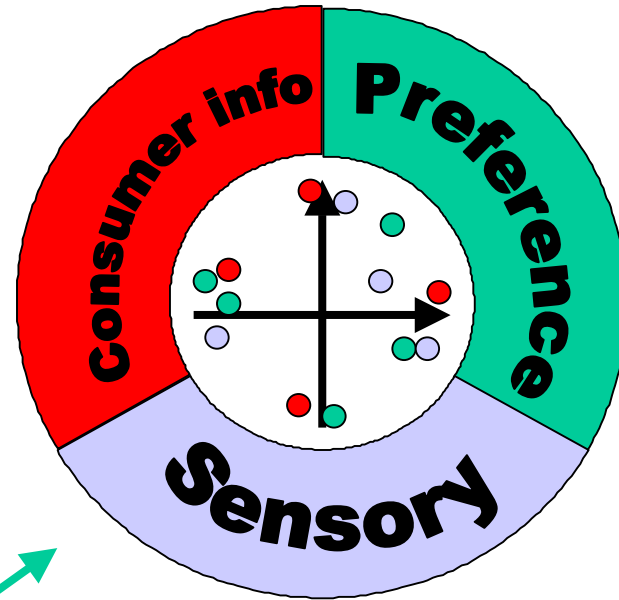


**%, 10 %**

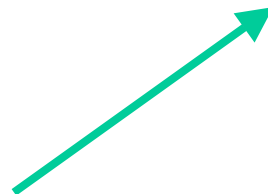
**55 %, 88 %**

Who is the consumer?

Consumer



Consumer

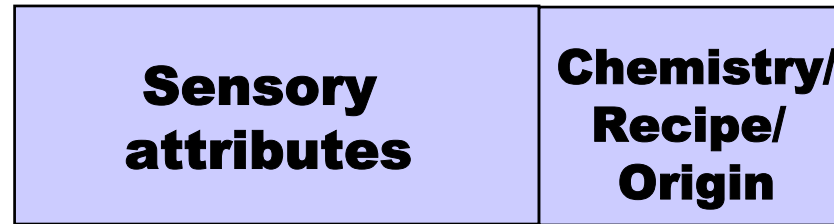


Sensory attributes + other

Product



Product

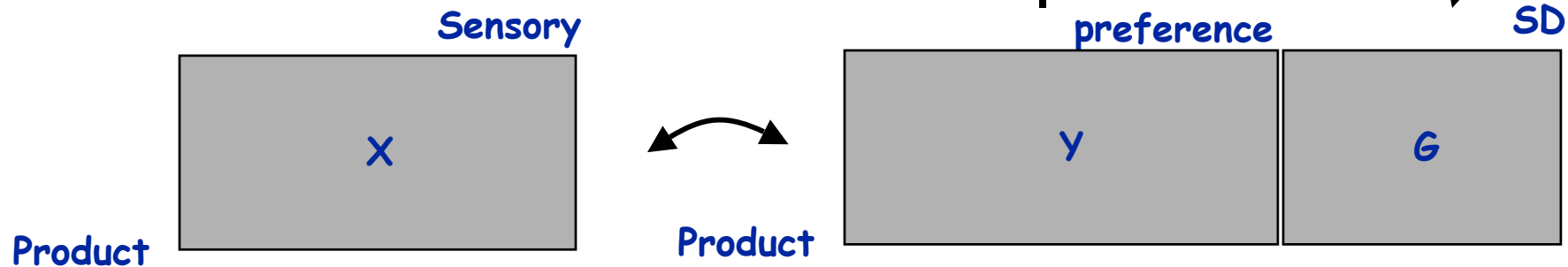


Product characteristics

Z: Table with consumers characterised by sociodemographic data

Procedure:

1.  $G$  = PLS Regression or correlation between  $(Z, Y)$
2.  $G$  as new response matrix linking sociodemographic data to preference
3. Model  $[Y \ G] = f(X)$



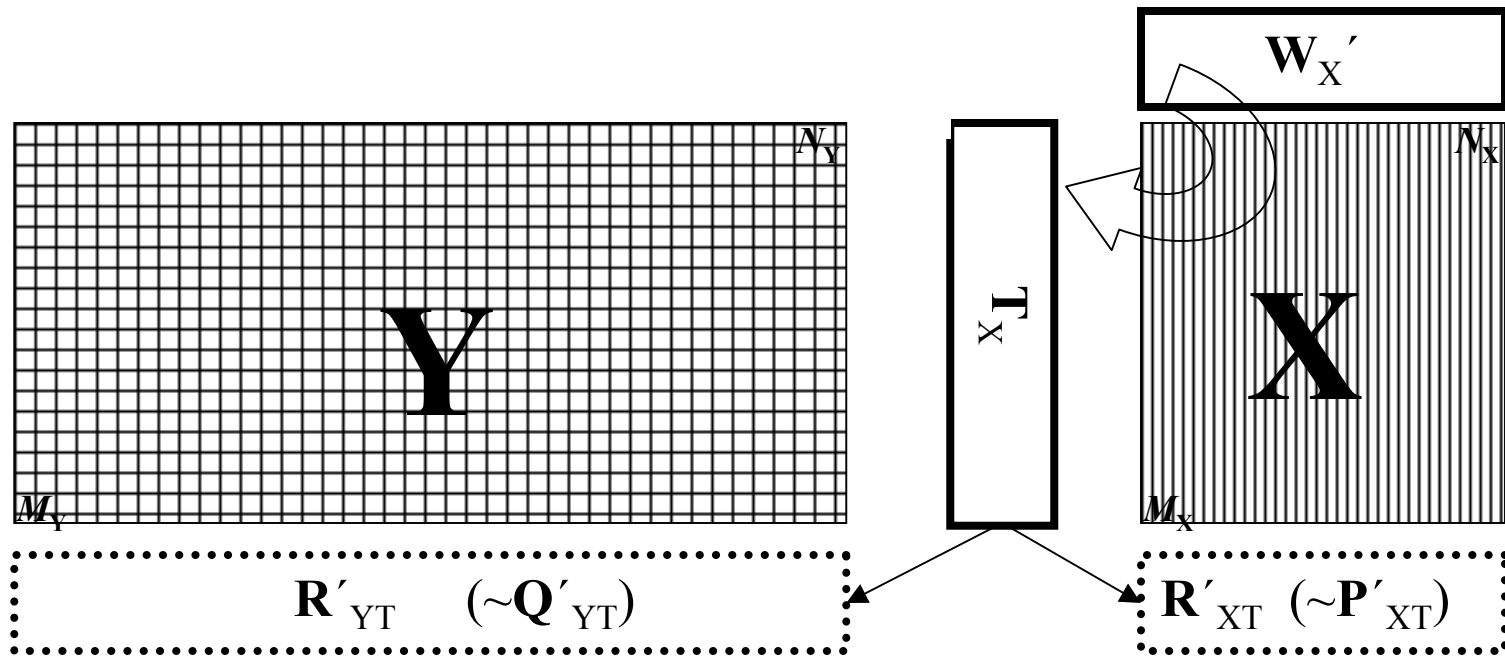
# L-PLSR modelling in consumer research

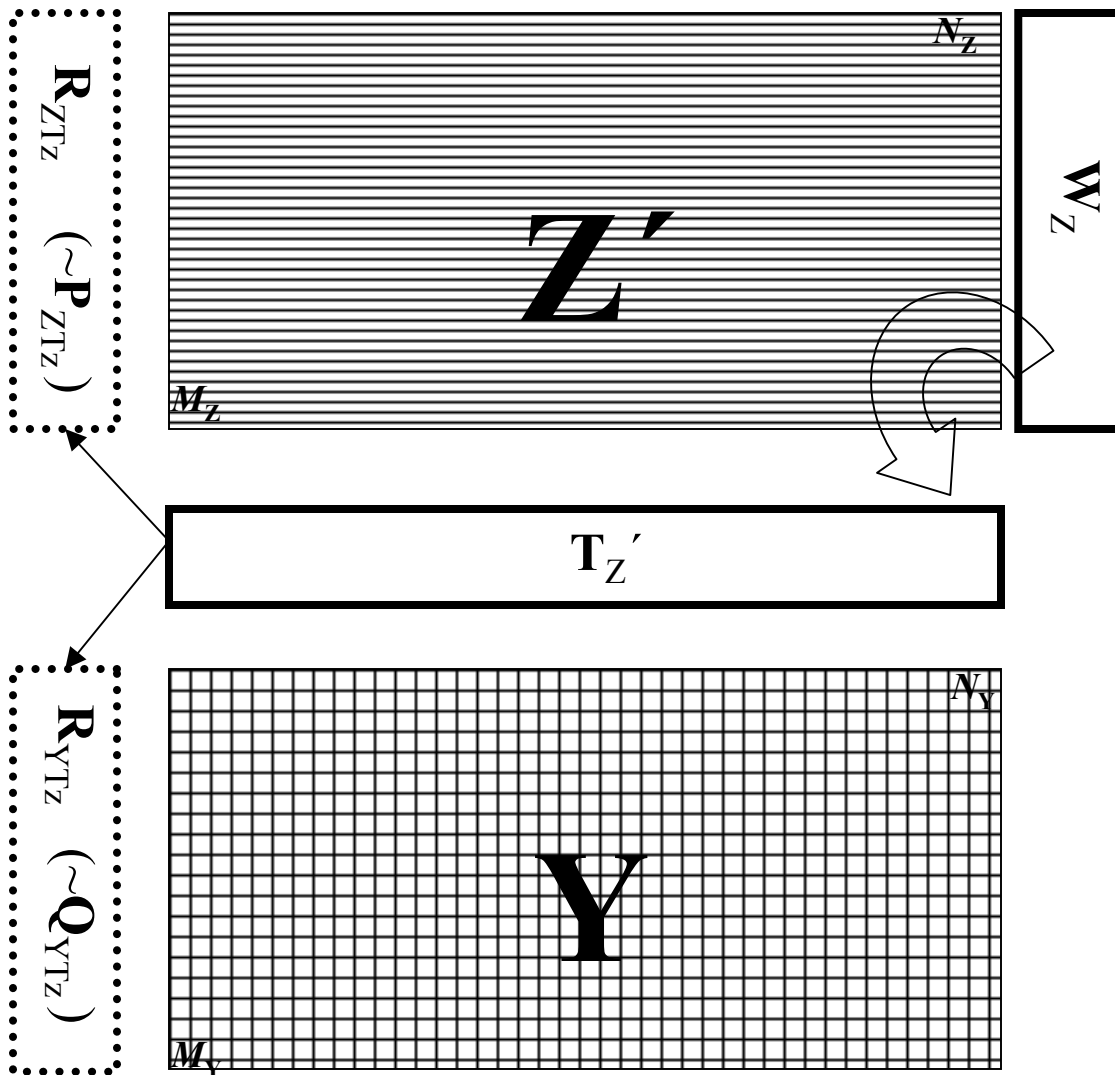
## Martens et al 2003

- ◆ The X matrix = product variables
- ◆ The Y matrix = consumer responses e.g. preference
- ◆ The Z matrix = background characteristics for consumers
  
- ◆ The procedure can bring out the structures in Y that are seen both from X and from Z.
  - The consumer responses (Y) are interpreted by the product variables (X) and the background information (Z)
  - The model parameters are checked by cross validation
- ◆ Singular value decomposition of matrix  $XY'Z$  gives  $w_x$  and  $w_z$

**Two-block**

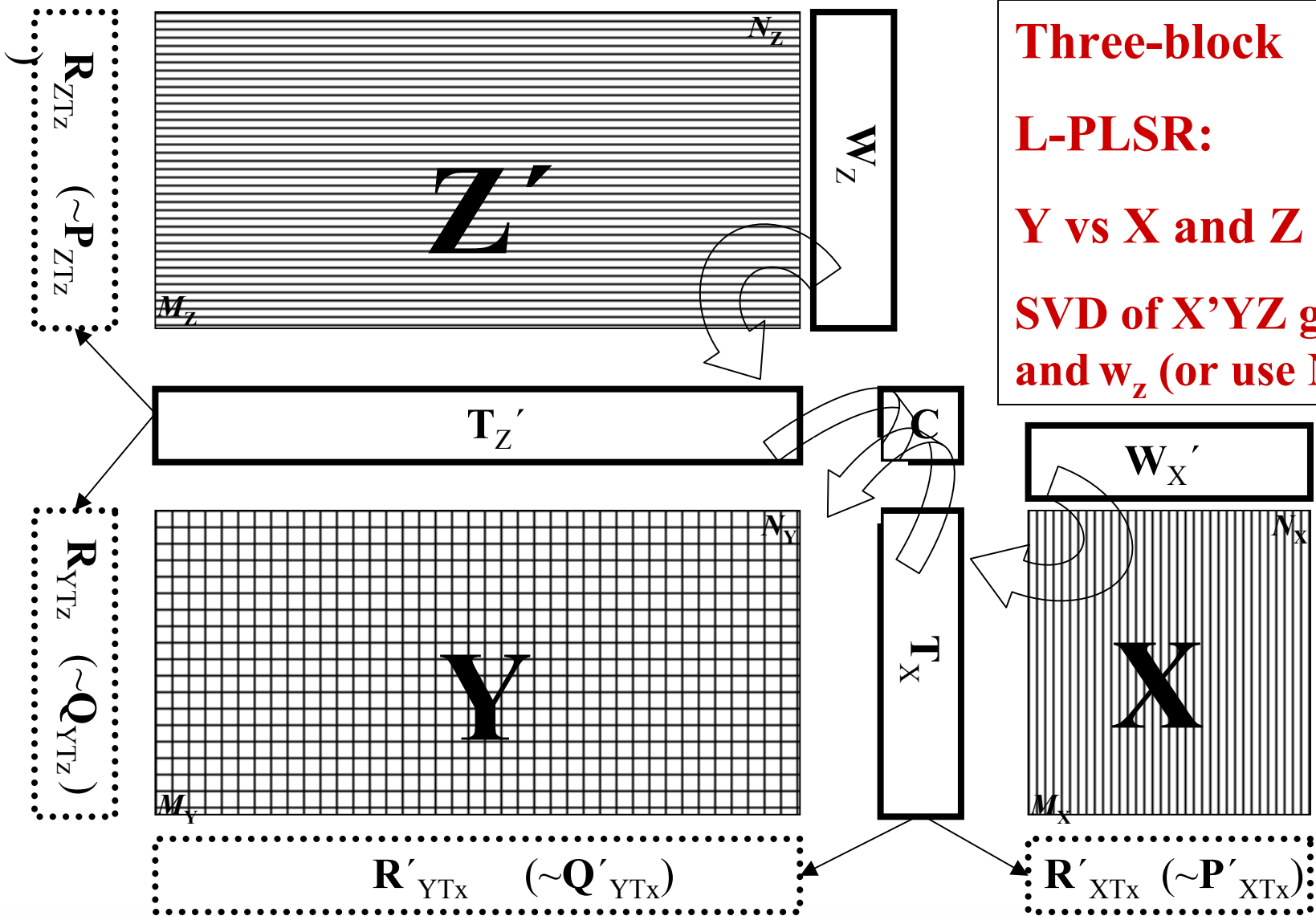
**PLSR: Y vs X**





**Two-block**

**PLSR:  $Y'$  vs  $Z$**



**Three-block  
L-PLSR:  
Y vs X and Z**

**SVD of  $X'YZ$  gives  $w_x$   
and  $w_z$  (or use NIPALS)**

# Some words about consumer segments...

- ◆ **Consumer segments are often generated from cluster analysis**
  - Long-thin orientation of the data
  - Short-fat orientation (like in preference mapping)
  - Mean centring and scaling issue
- ◆ **Segment on preference or consumer background variables?**
- ◆ **How unique are the segments? Validate with**
  - Local PCA models
  - Support Vector Machines
- ◆ **With the L-model there is no requirement to segment consumers *a priori***

# Pitfalls in L-PLSR

- ◆ **Correlations between Z (sociodemographic) and Y (preference) might be low**
  - ◆ **...but the correlations to preferences for individual Z-variables can be still be systematic**
  - ◆ **And, as always, with 200-300 consumers, correlations might be significant... but too weak to justify a specific marketing strategy**
- ⇒ **Perform PLS regression between Z and Y to find significant variables as a screening step**

# O-PLSR - Motivation

- ◆ O-PLSR removes the *part* of the predictor variables that is orthogonal to Y (Trygg & Wold, 2001)
  - ◆ Must decide on how many components to extract from e.g. cross-validation
    - Number of predictive components (*one* if only one response)
    - Number of orthogonal components
  - ◆ Has been extended to O2-PLSR (Trygg & Wold, 2003)
  - ◆ May use jack-knifing to estimate uncertainties and thereby find which variables that are significant
- ⇒ Get optimal interpretation and prediction in *one* model

# O2-PLS regression

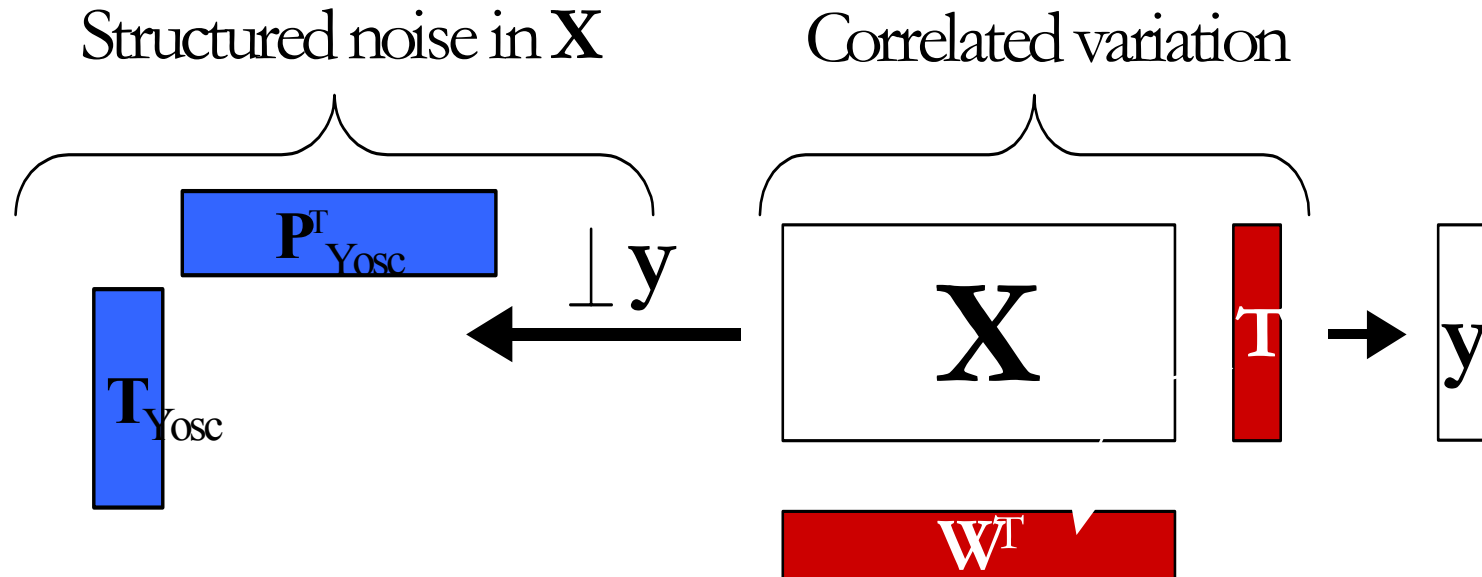
The general O2-PLS regression has the following structure:

$$\begin{array}{l}
 \text{Model of } \mathbf{X}: \quad \mathbf{X} = \mathbf{T}\mathbf{W}^T + \mathbf{T}_{Y_{osc}}\mathbf{P}_{Y_{osc}}^T + \mathbf{E} \\
 \text{Model of } \mathbf{Y}: \quad \mathbf{Y} = \underbrace{\mathbf{U}\mathbf{C}^T}_{\text{Predictive part}} + \underbrace{\mathbf{U}_{X_{osc}}\mathbf{P}_{X_{osc}}^T}_{\text{Structured noise}} + \underbrace{\mathbf{F}}_{\text{Pure residual}}
 \end{array}$$

For PLS1,  $\mathbf{w}_{Y_{osc}}$  is the difference vector ( $\mathbf{p}-\mathbf{w}$ ) between the first PLSR component loadings:

$$\mathbf{t}_{Y_{osc}} = \mathbf{X}\mathbf{w}_{Y_{osc}} = \mathbf{X}(\mathbf{p}-\mathbf{w}) / \|\mathbf{p}-\mathbf{w}\|$$

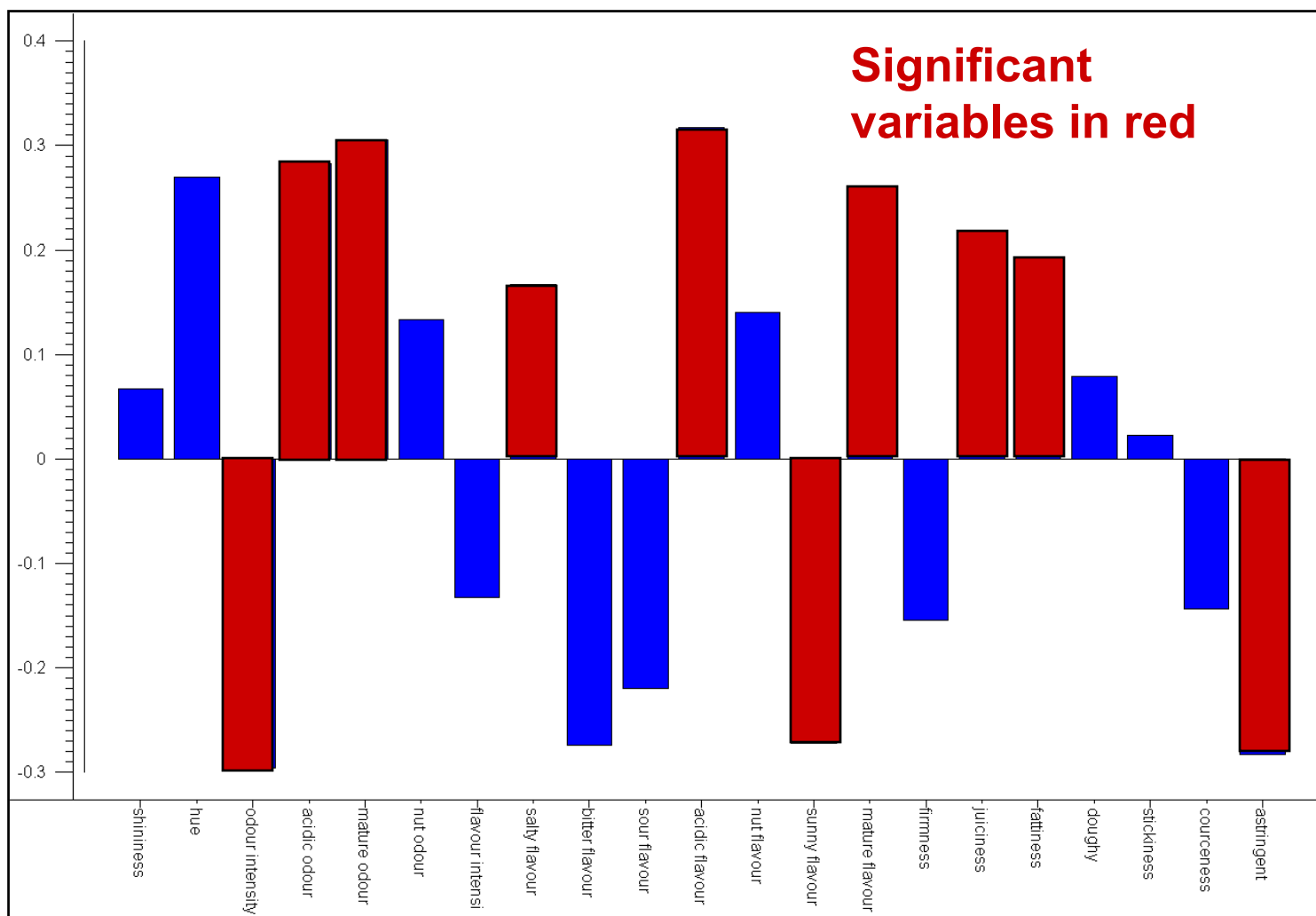
# Separation of X in O-PLSR (correspondingly also for Y in O2-PLSR)



# O-PLSR - Cheese data

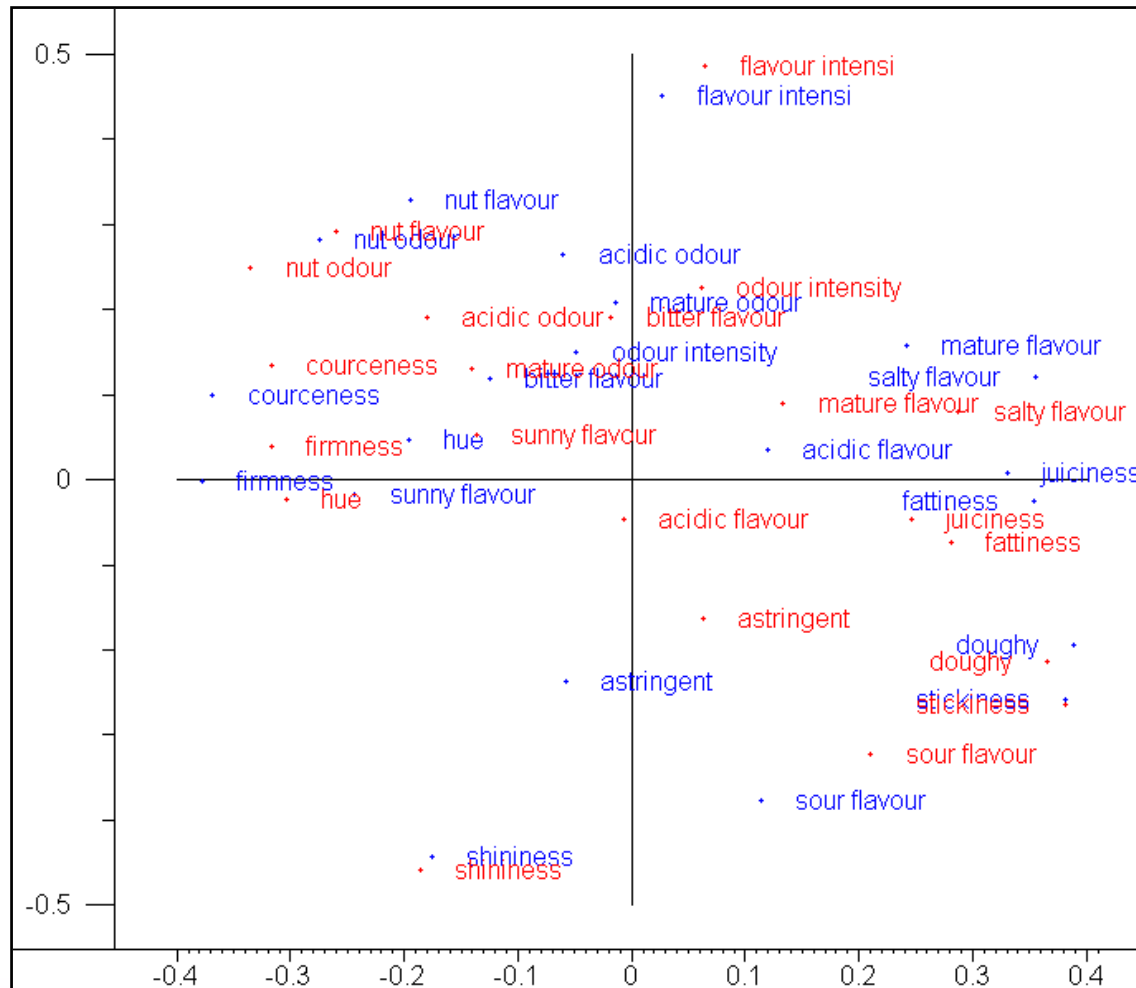
- ◆ Run O-PLSR with average preference as response variable
- ◆ Estimate uncertainties for all model parameters
  - When the structured noise components (e.g.  $P_{Y_{osc}}$ ) are extracted during cross-validation, their orientation may be mirrored. To handle this one can use:
    - ◆ Orthogonal Procrustes rotation
    - ◆ Flip and order vectors based on correlation between main and sub-models
- ◆ Compare results

# Predictive component in O-PLSR; cheese data



# Comparison of loadings

Blue: Portho 1 and 2    Red: P2 and P3



# Prediction in O-PLSR

Center and scale  $X_{\text{new}}$

For  $a = 1:A_{\text{ortho}}$

$$t_{\text{new}} = X_{\text{new}} * W_{\text{ortho}(a)}$$

$$X_{\text{new}} = X_{\text{new}} - t_{\text{new}} * P_{\text{ortho}(a)}^T$$

end

$$\hat{Y} = B_{\text{pred}} * X_{\text{new}} \text{ (filtered)}$$

$$X_{\text{new,ortho}} = \sum_{a=1}^{A_{\text{ortho}}} t_{\text{new}} * P_{\text{ortho}}^T$$

# Summary

- ◆ **Chemical, sensory, preference and consumer data can be combined with quantitative models**
  - validation
  - significance testing
  - passify; don't throw away
  - orthogonal components
- ◆ **Correlation loadings are useful**
- ◆ **We (the PLS clan) should lend our ears to maximum likelihood, logit models, GLM and other ways to model binary data and data from distributions other than the (multi)normal**

# Acknowledgements

**Harald Martens, Norway (for borrowing some of his slides)**  
**Tine Norwegian Dairies**

**Thanks for your attention!**

***...and may your data be with you***