Relating instrumental and sensory data of German Riesling wines by means of PLS

Anita Busch

Ulrich Fischer

Michael Meyners

The Sixth sense

Content

- Instrumental and sensory Data collection: „Does organic viticulture have an impact on the aroma of German Riesling wines?“
- PCA and PLS
- Data-Analysis
- Interpretation

The Data:

56 Riesling wines from

4 viticultural areas,

3 vintages and

2 production styles

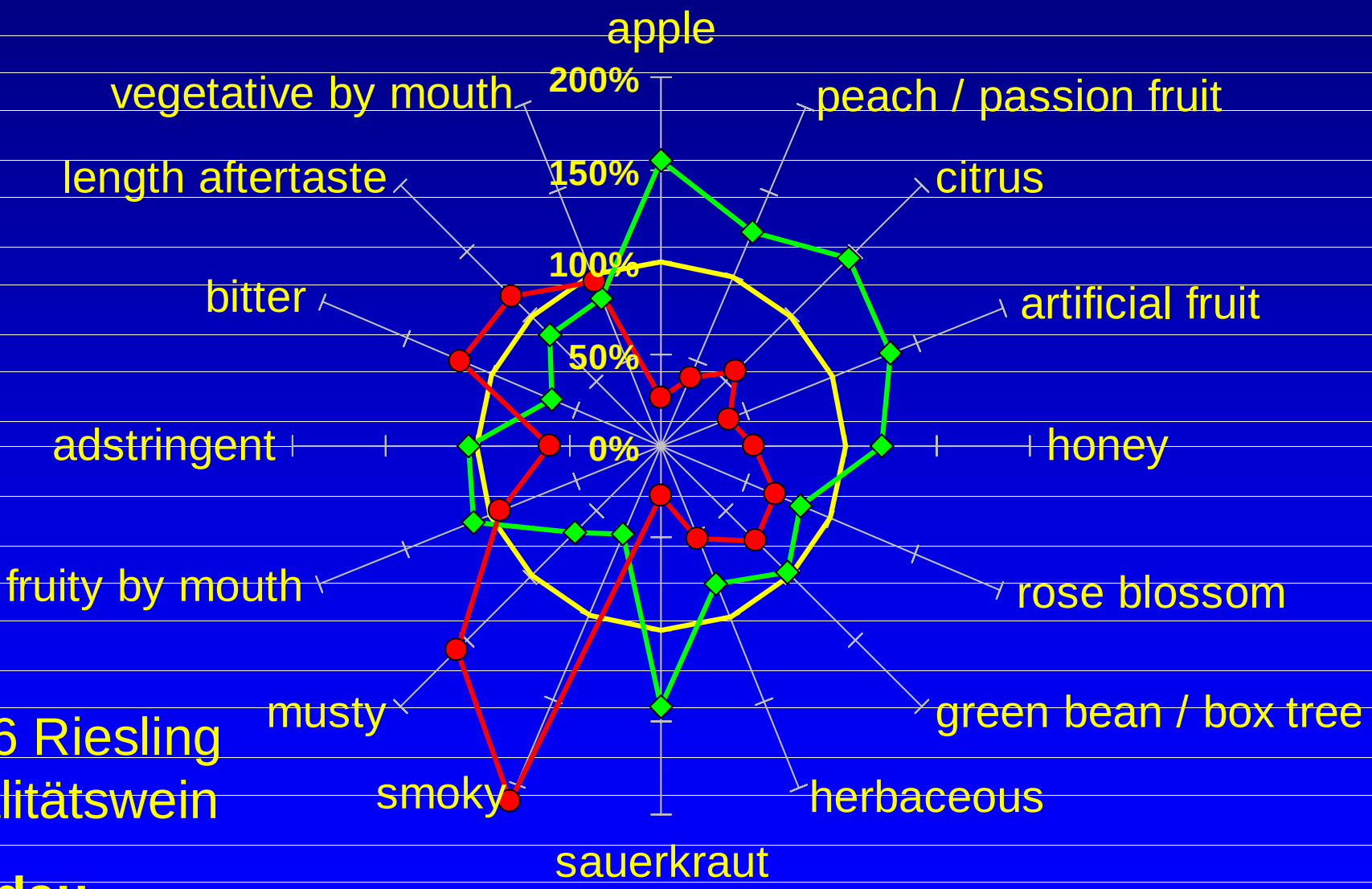
aroma profile, GC-MS, 55 flavour compounds:

production style	3 compounds p<0.05
geographic region	15 compounds p<0.05

sensory profile,15 descriptors and 14 judges 2 Reps

production style	1 attribute p<0.05
geographic region	7 attributes p<0.05
wine estate	0 attributes p<0.05

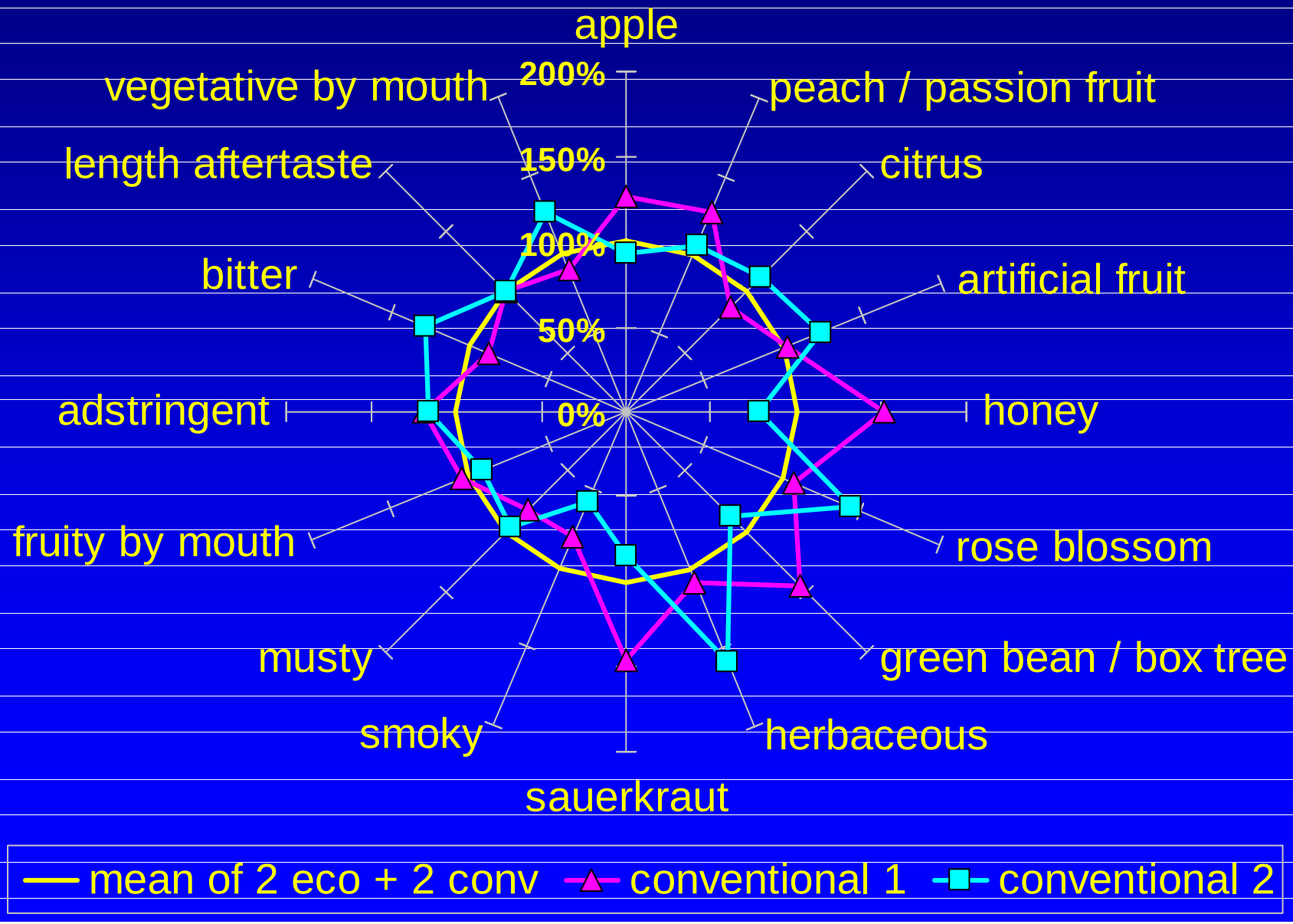
Comparison of wines from 2 different organic wine estates (n=14J x 2R)



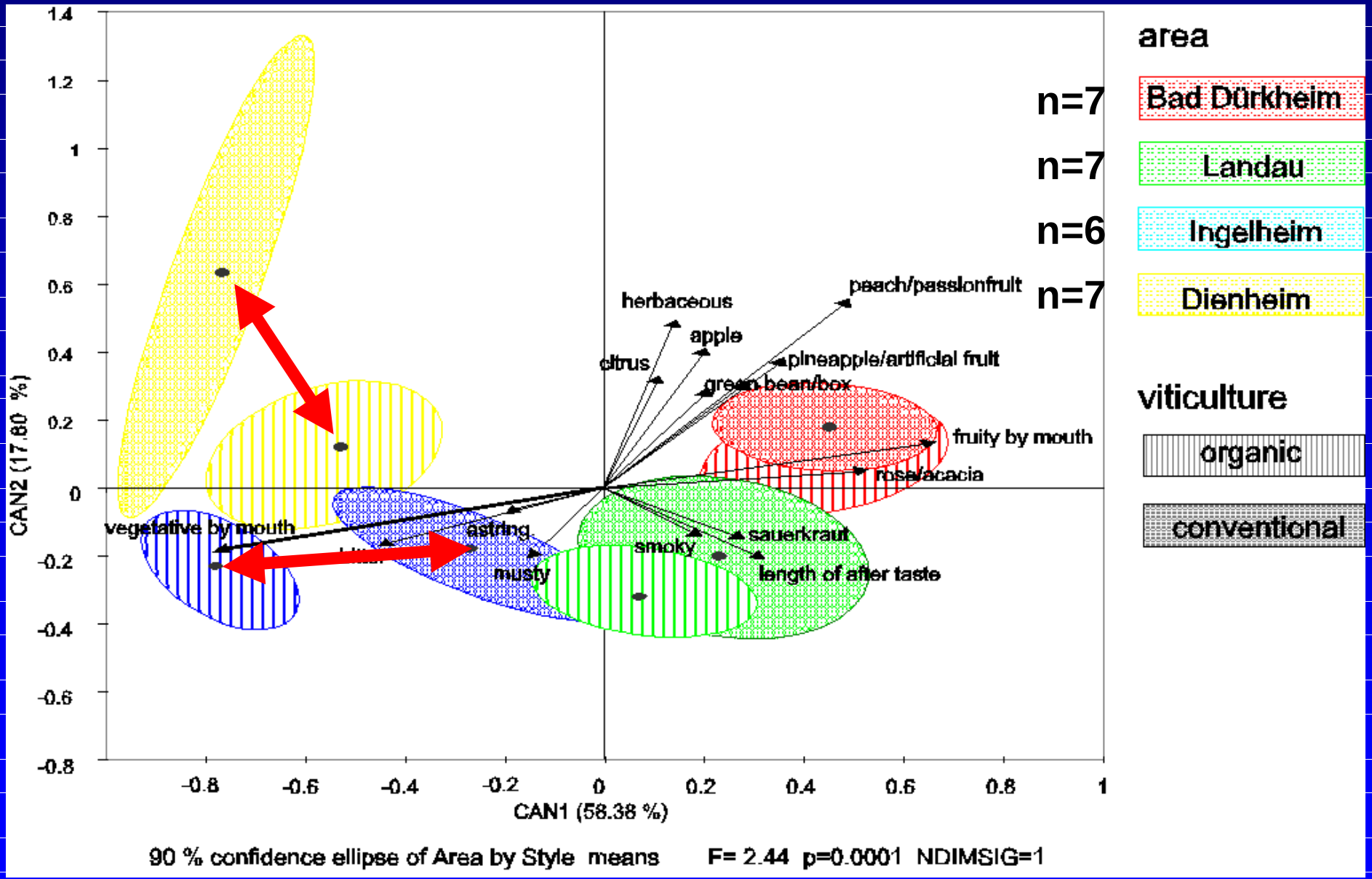
1996 Riesling
Qualitätswein

Landau

Comparison of wines from 2 different convent. wine estates (n=14J x 2R)

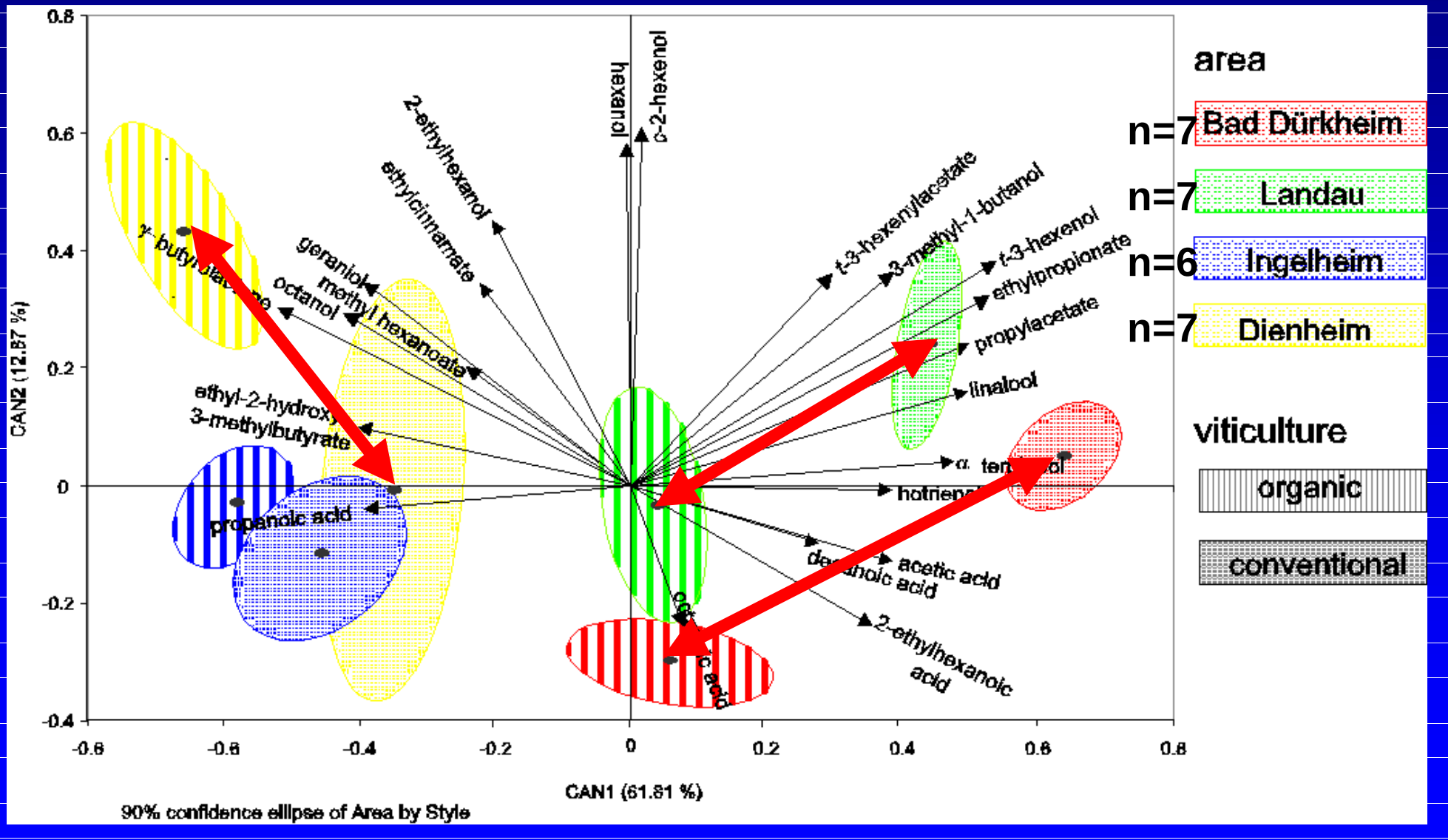


CVA: discrimination of area and production style based on sensory variables



90 % confidence ellipse of Area by Style means F= 2.44 p=0.0001 NDIMSIG=1

CVA: discrimination of area and production style based on volatiles



Relating instrumental and sensory variables

Objective: Are we able to explain and predict the sensory of wines by volatiles?

Instrumental variables: K=55 volatiles	Mean sensory variables: L=12 descriptors
$X_1, \ldots, X_K$	$Y_1, \ldots, Y_L$

Aim: building a linear relationship

$$\vec{Y} = \vec{\mu}_Y + \mathbf{B}\vec{X} + \vec{\varepsilon}$$

Sample: N=56 Riesling Wines

Principal Component Analysis

Tool for reduction of dimensionality of a variable set with

- high dimensionality, high amount of correlation
- many variables, few observations

Idea : Variables originate from mixture of few unknown latent variables and noise

Components

→ maximizing variance

→ pairwise uncorrelated

$$T_j = \vec{p}'_j \, \vec{X}, \quad j = 1, \ldots, A$$

Projection of K-dim. space to A-dim. subspace

$$\begin{pmatrix} X_1 \\ \vdots \\ X_K \end{pmatrix} = \begin{pmatrix} p_{11} & \cdots & p_{A1} \\ \vdots & \ddots & \vdots \\ p_{1K} & \cdots & p_{AK} \end{pmatrix} \begin{pmatrix} T_1 \\ \vdots \\ T_A \end{pmatrix} + \vec{E}_A$$

Partial Least Squares - Regression

Tool for relating two variable sets

→ biased estimator for regression coefficients

Idea: Ascribe space of explanatory and response variables separately to latent variables

Close relation between X and Y

→ close relation between latent variables

Partial Least Squares - Regression

$$\vec{X} = \vec{p}_1 T_1 + \vec{E}_1 \qquad \vec{Y} = \vec{q}_1 U_1 + \vec{F}_1^{\,*} \qquad \vec{Y} = \vec{d}_1 T_1 + \vec{F}_1$$

1) Components: Solve

$$\max\{\mathrm{cov}^2(T_1, U_1)\} = \max_{\vec{w}'\vec{w}=1, \vec{c}'\vec{c}=1} \{\mathrm{cov}^2(\vec{w}'\vec{X}, \vec{c}'\vec{Y})\},$$

2) Loading vectors: $\vec{p}_1$ and $\vec{d}_1$ by linear regression,

3) Error: Go on with $\vec{E}_1$ and $\vec{F}_1$ instead of $\vec{X}$ and $\vec{Y}$.

Partial Least Squares - Regression

$$\vec{E}_1 = \vec{p}_2 T_2 + \vec{E}_2 \qquad \vec{F}_1 = \vec{q}_2 U_2 + \vec{F}_2^{\,*} \qquad \vec{F}_1 = \vec{d}_2 T_2 + \vec{F}_2$$

1) Components: Solve

$$\max\{\mathrm{cov}^2(T_2, U_2)\} = \max_{\vec{w}'\vec{w}=1, \vec{c}'\vec{c}=1} \{\mathrm{cov}^2(\vec{w}'\vec{E}_1, \vec{c}'\vec{F}_1)\},$$

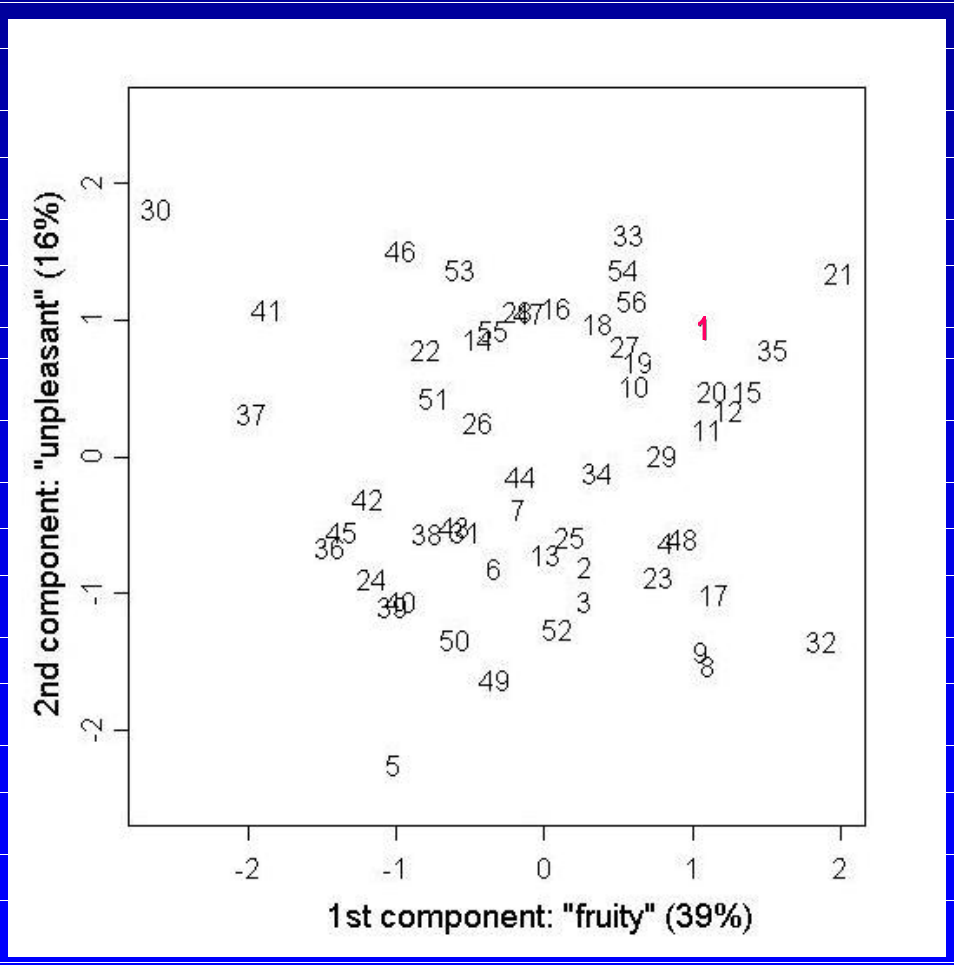
2) Loading vectors: $\vec{p}_2$ and $\vec{d}_2$ by linear regression,

3) Error: Go on with $\vec{E}_2$ and $\vec{F}_2$ instead of $\vec{E}_1$ and $\vec{F}_1$.

PCA of mean sensory profile

$$\mathbf{Y} = \vec{t}_1 \vec{p}'_1 + \vec{t}_2 \vec{p}'_2 + \mathbf{E}$$

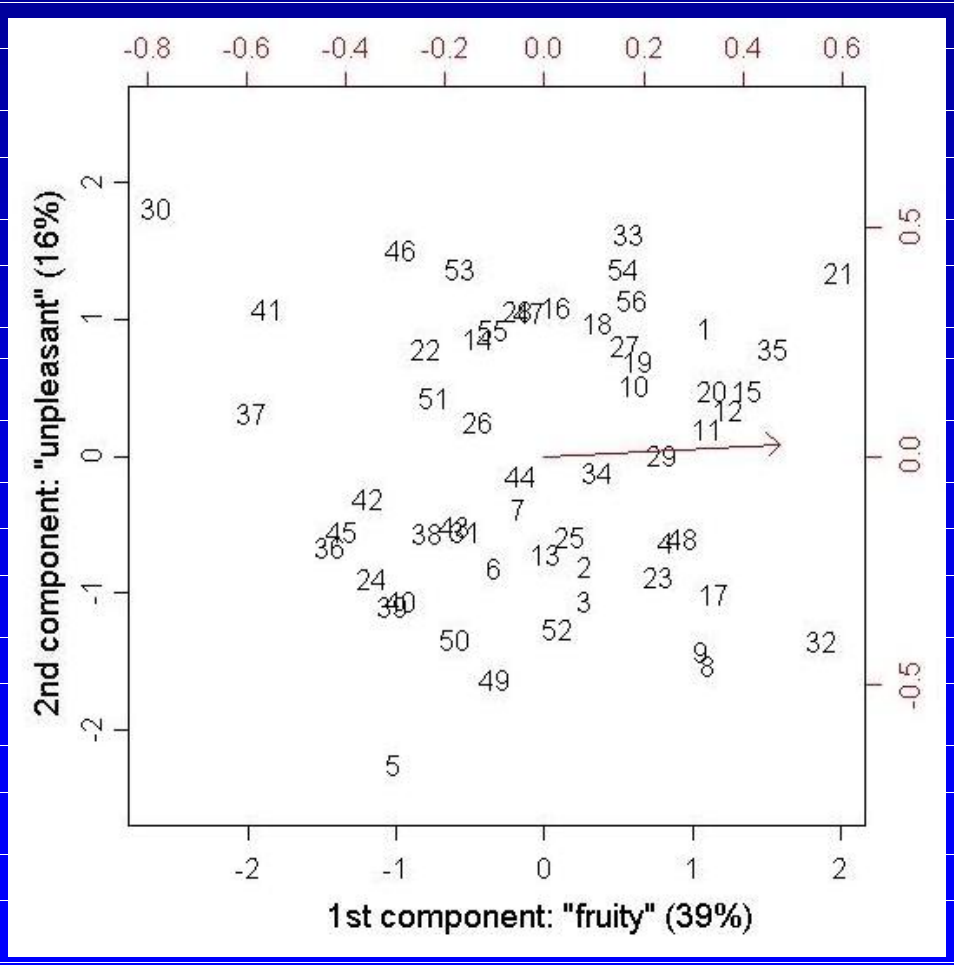
$$= \begin{pmatrix} t_{11} \\ \vdots \\ t_{1N} \end{pmatrix} \begin{pmatrix} p_{11} & \cdots & p_{1K} \end{pmatrix} + \begin{pmatrix} t_{21} \\ \vdots \\ t_{2N} \end{pmatrix} \begin{pmatrix} p_{21} & \cdots & p_{2K} \end{pmatrix} + \mathbf{E}$$



PCA of mean sensory profile

$$\mathbf{Y} = \vec{t}_1 \vec{p}'_1 + \vec{t}_2 \vec{p}'_2 + \mathbf{E}$$

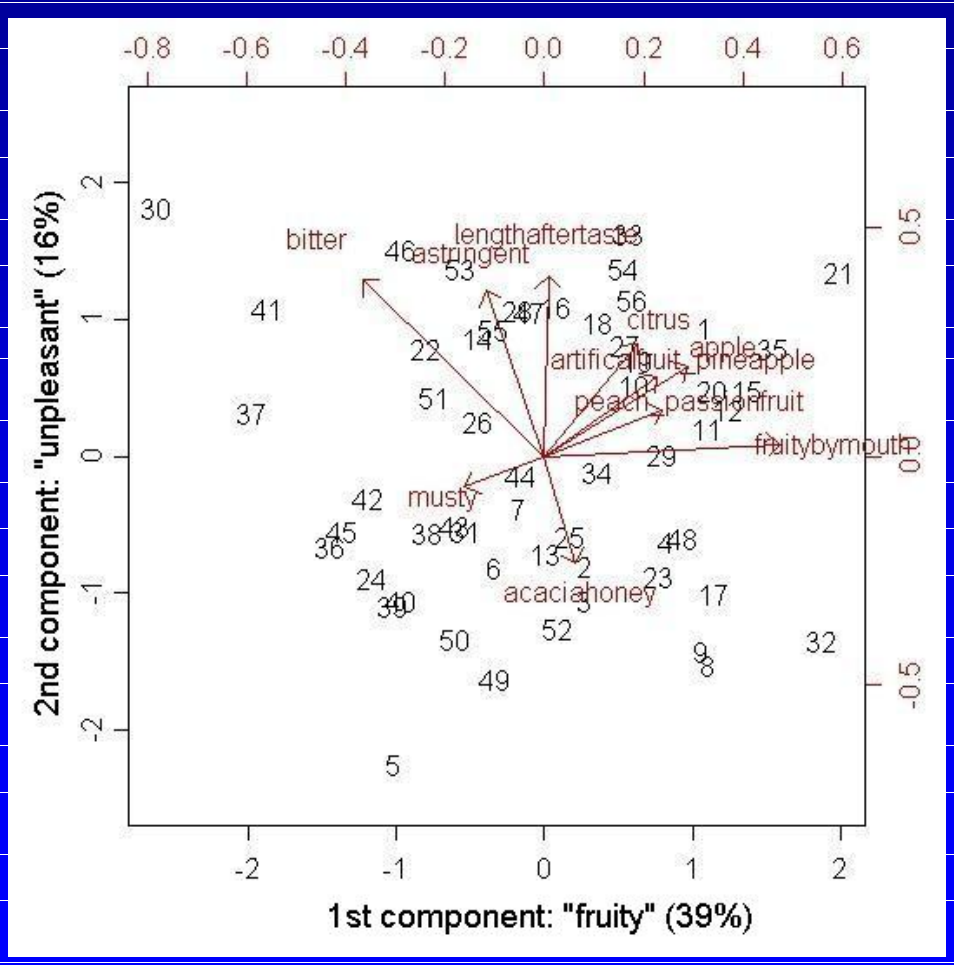
$$= \begin{pmatrix} t_{11} \\ \vdots \\ t_{1N} \end{pmatrix} \begin{pmatrix} p_{11} & \cdots & p_{1K} \end{pmatrix} + \begin{pmatrix} t_{21} \\ \vdots \\ t_{2N} \end{pmatrix} \begin{pmatrix} p_{21} & \cdots & p_{2K} \end{pmatrix} + \mathbf{E}$$



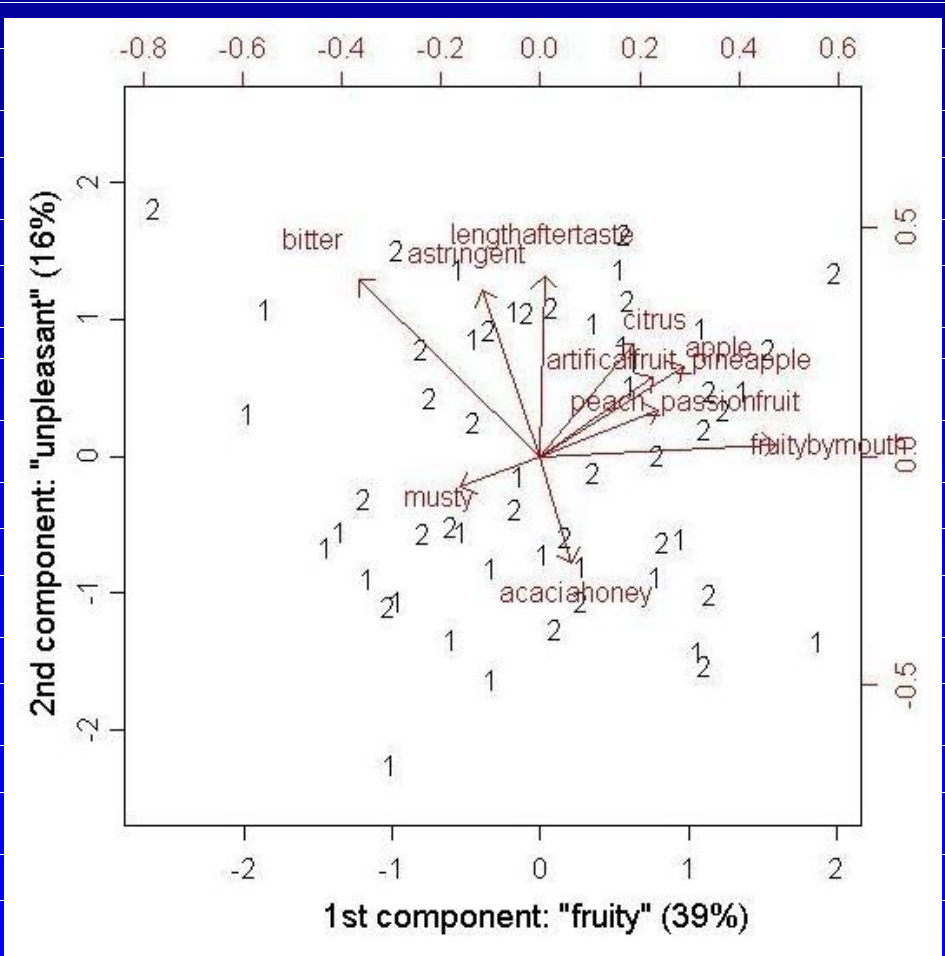
PCA of mean sensory profile

$$\mathbf{Y} = \vec{t}_1 \vec{p}'_1 + \vec{t}_2 \vec{p}'_2 + \mathbf{E}$$

$$= \begin{pmatrix} t_{11} \\ \vdots \\ t_{1N} \end{pmatrix} \begin{pmatrix} p_{11} & \cdots & p_{1K} \end{pmatrix} + \begin{pmatrix} t_{21} \\ \vdots \\ t_{2N} \end{pmatrix} \begin{pmatrix} p_{21} & \cdots & p_{2K} \end{pmatrix} + \mathbf{E}$$



PCA of mean sensory profile

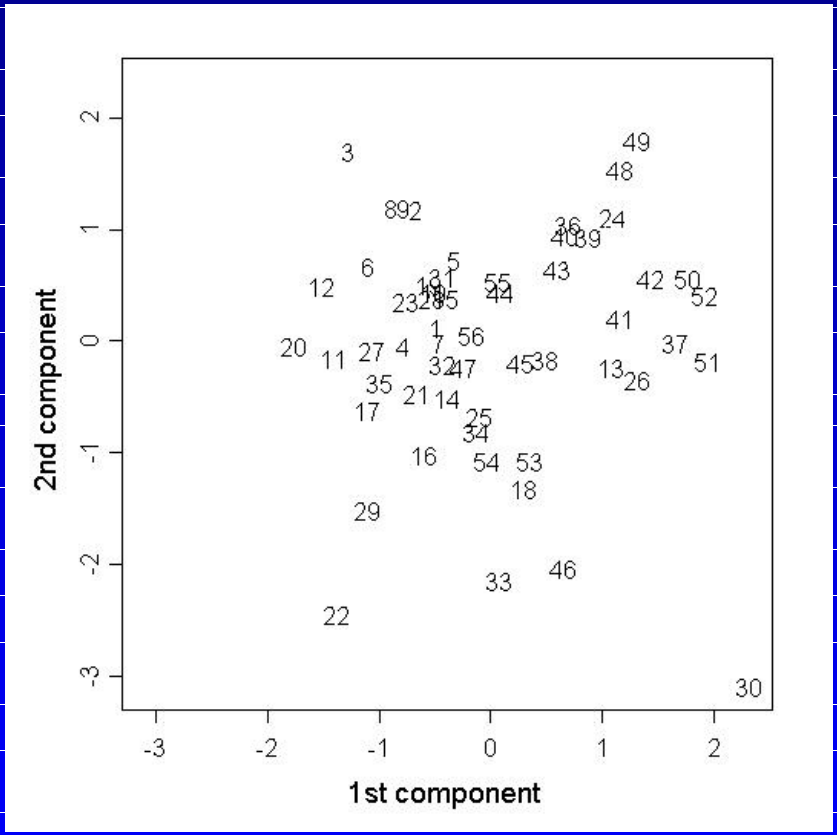


production type:

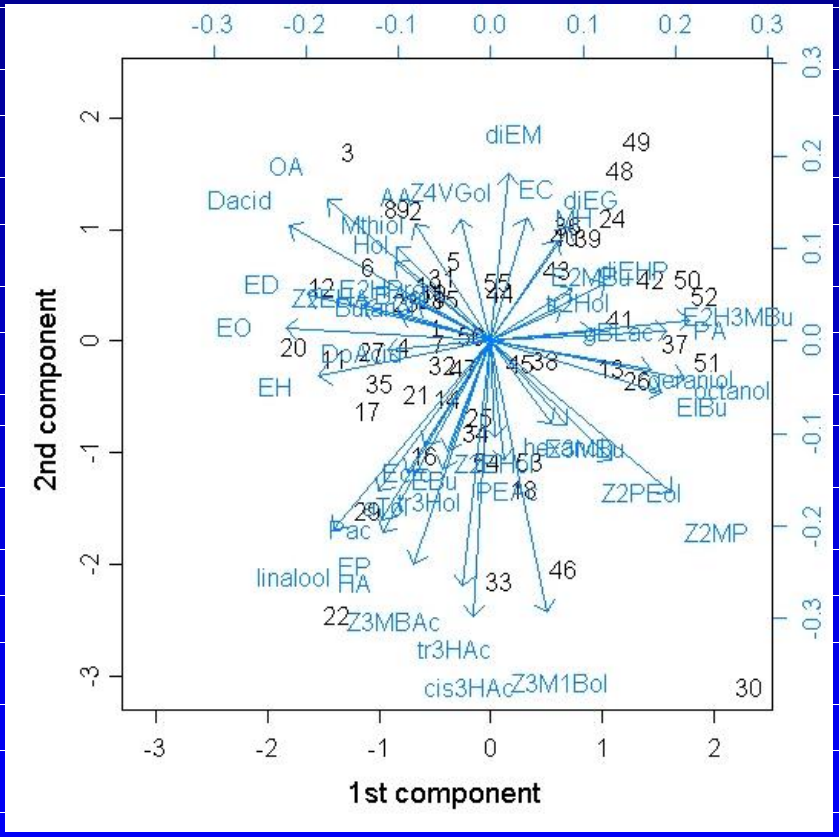
1: organic

2: conventional

PLS-Regression

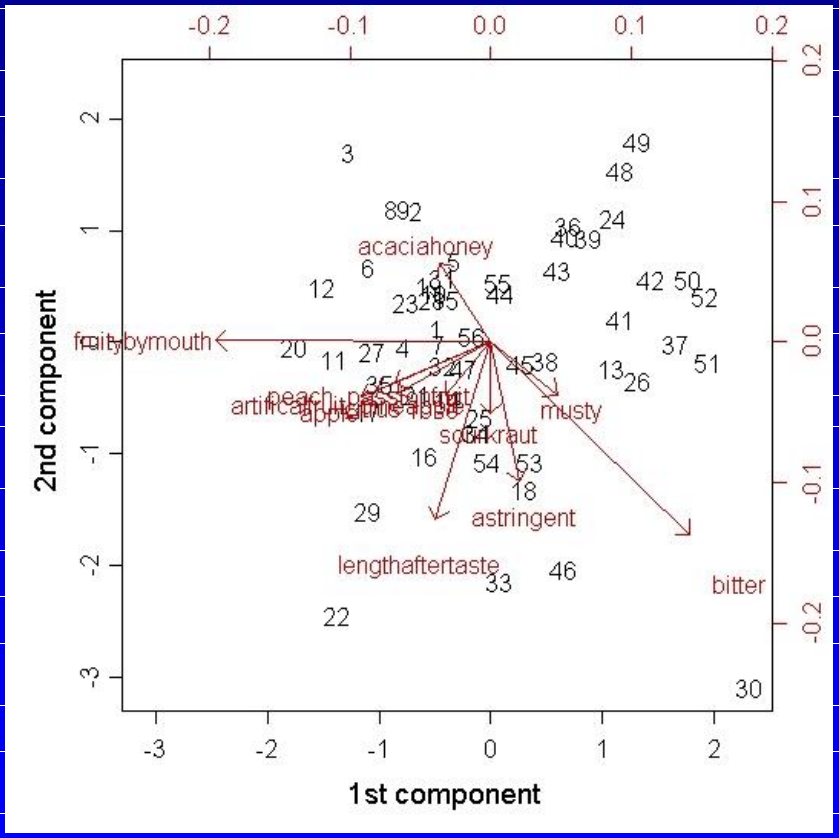


PLS-Regression



loadings of volatiles (X)

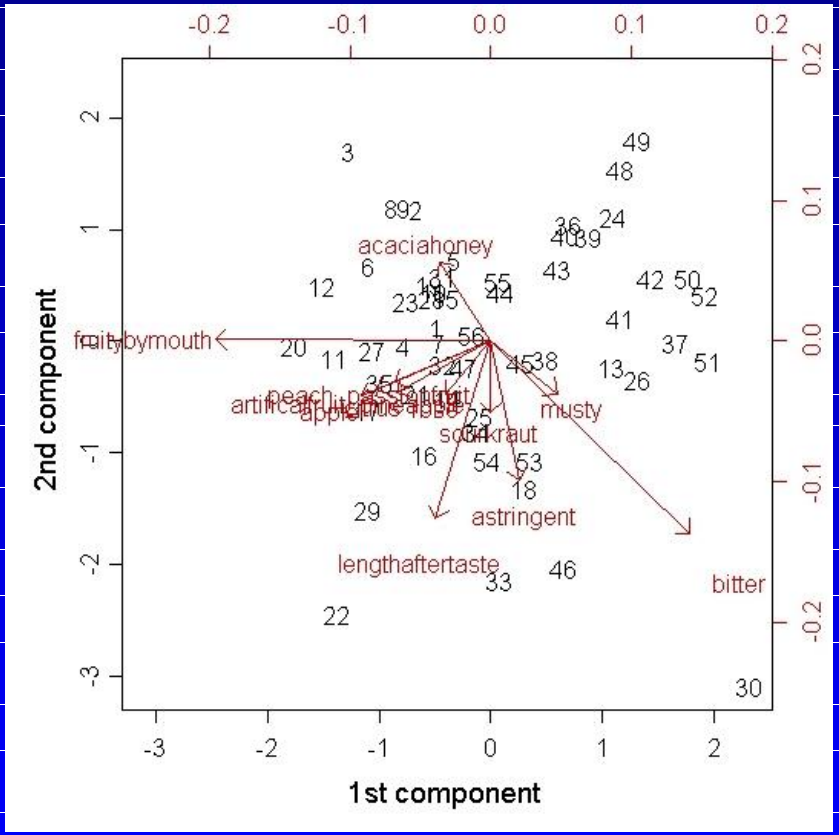
PLS-Regression



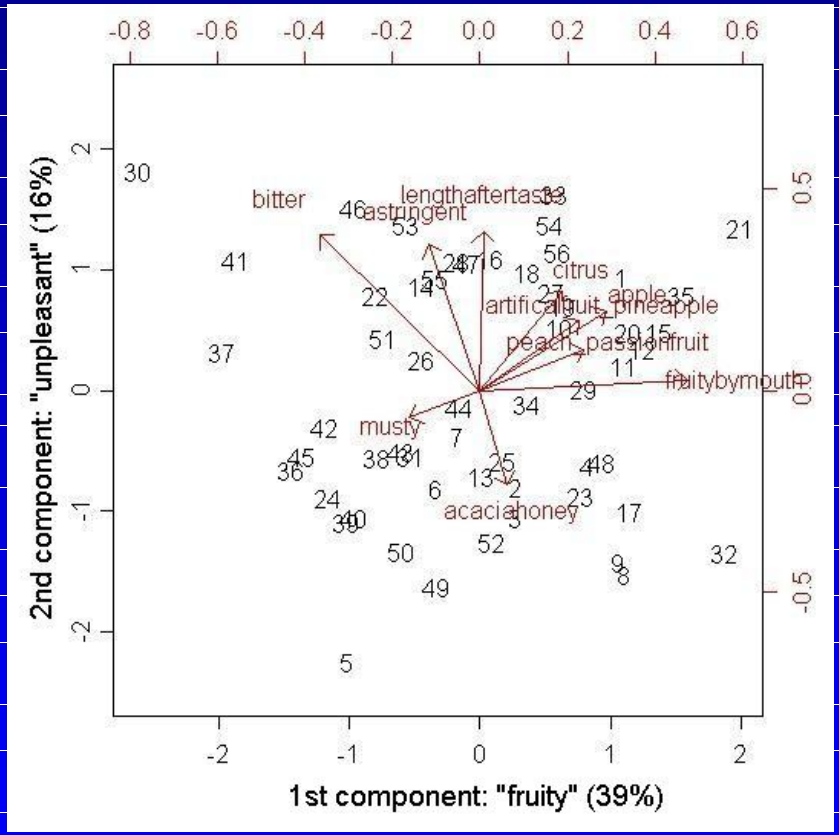
loadings of sensory (Y)

PLS-Regression vs. PCA of sensory

PLS

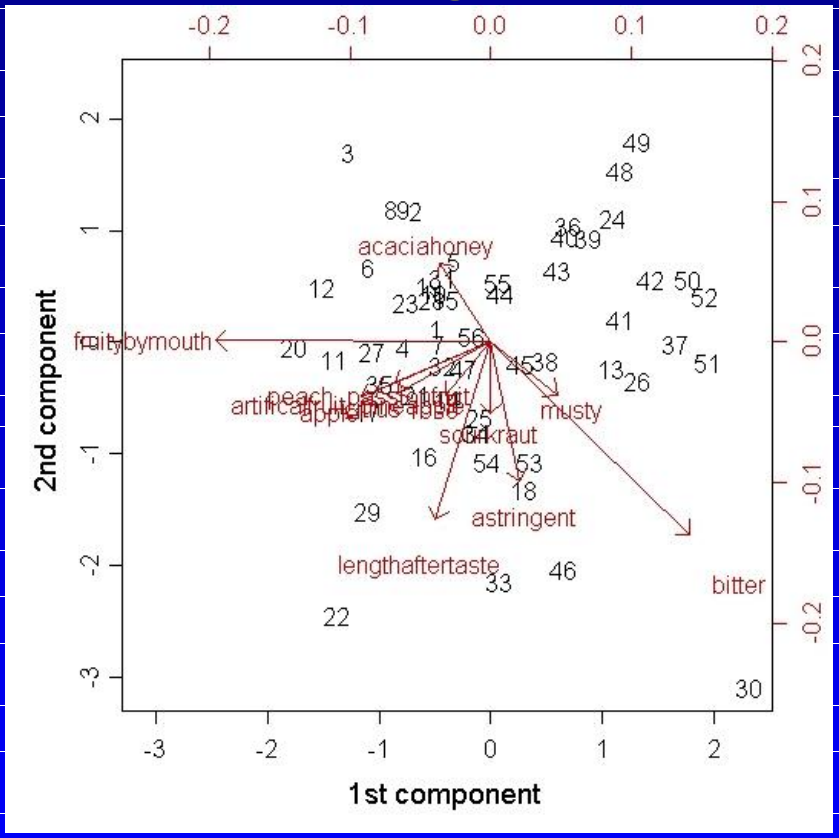


PCA

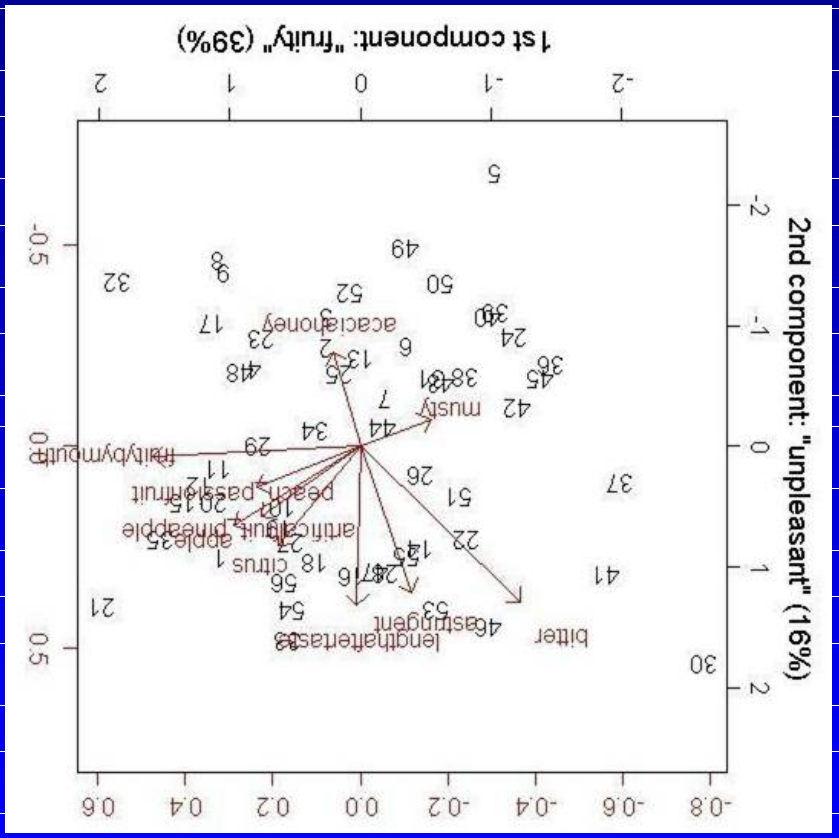


PLS-Regression vs. PCA of sensory

PLS



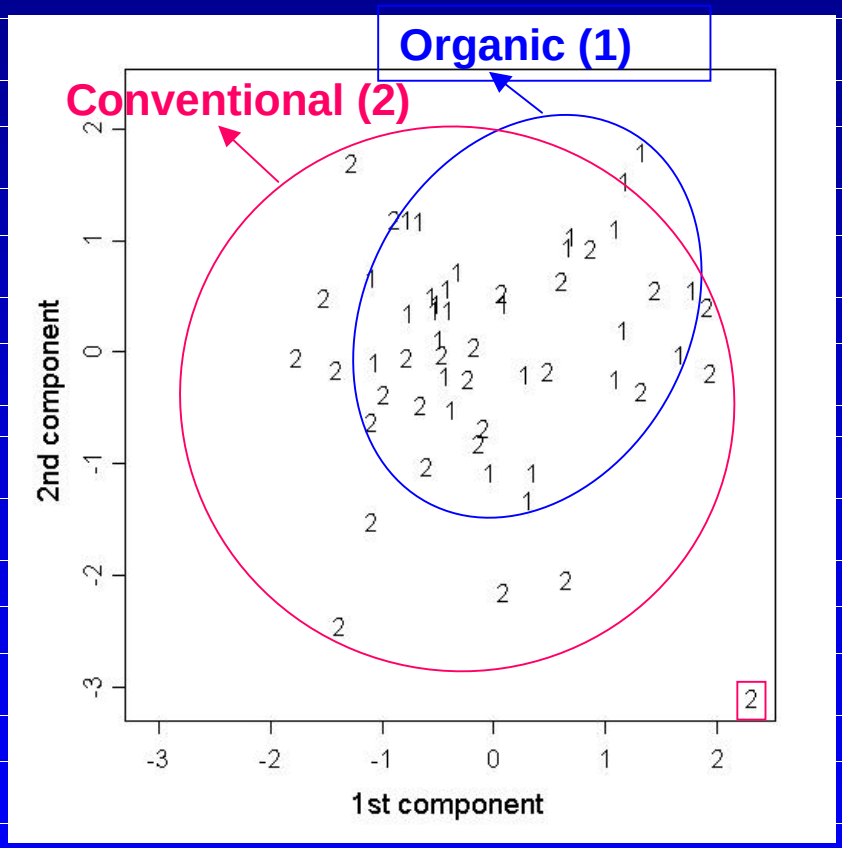
PCA

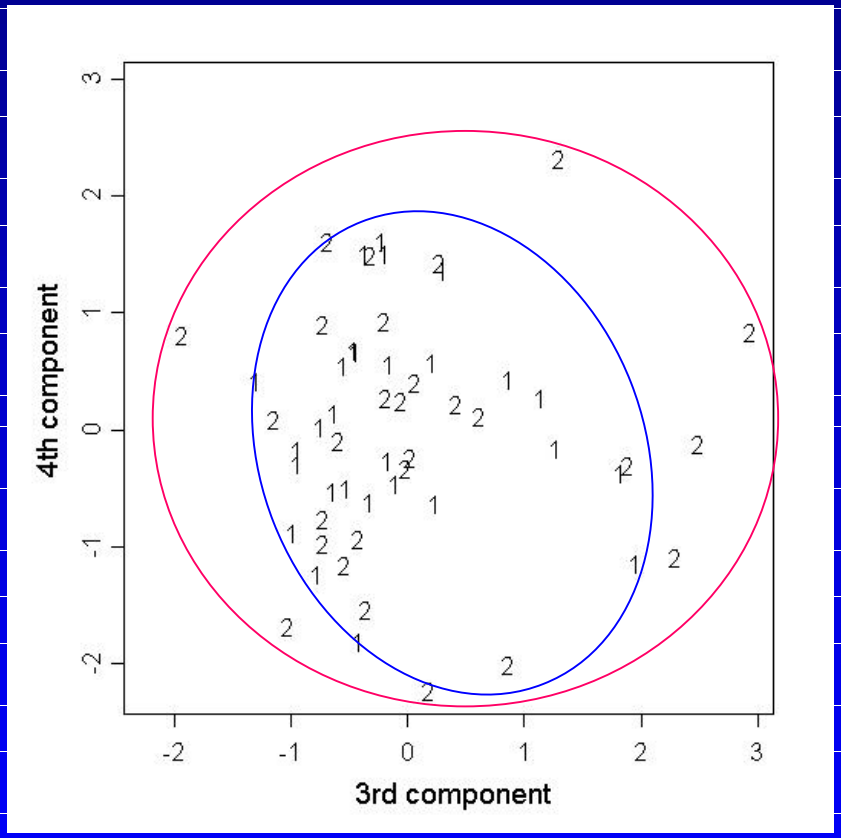


1st/2nd component: Regression-Loadings like in PCA; Scores differ

→ Relation in first components, but high dimensional problem

PLS-scores by production type





Wider spread for conventionally produced wines (2)

Number of PLS-components

Measures of Goodness:

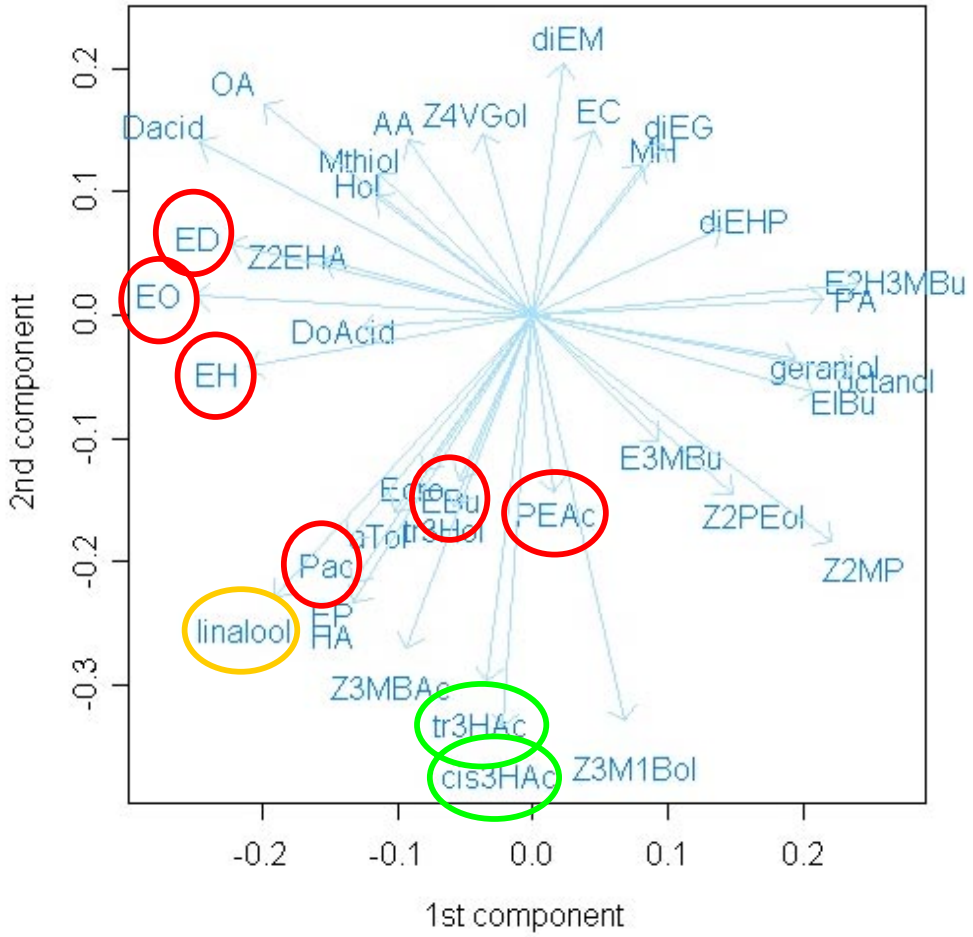
$$R^2 = 1 - RSS/TSS$$

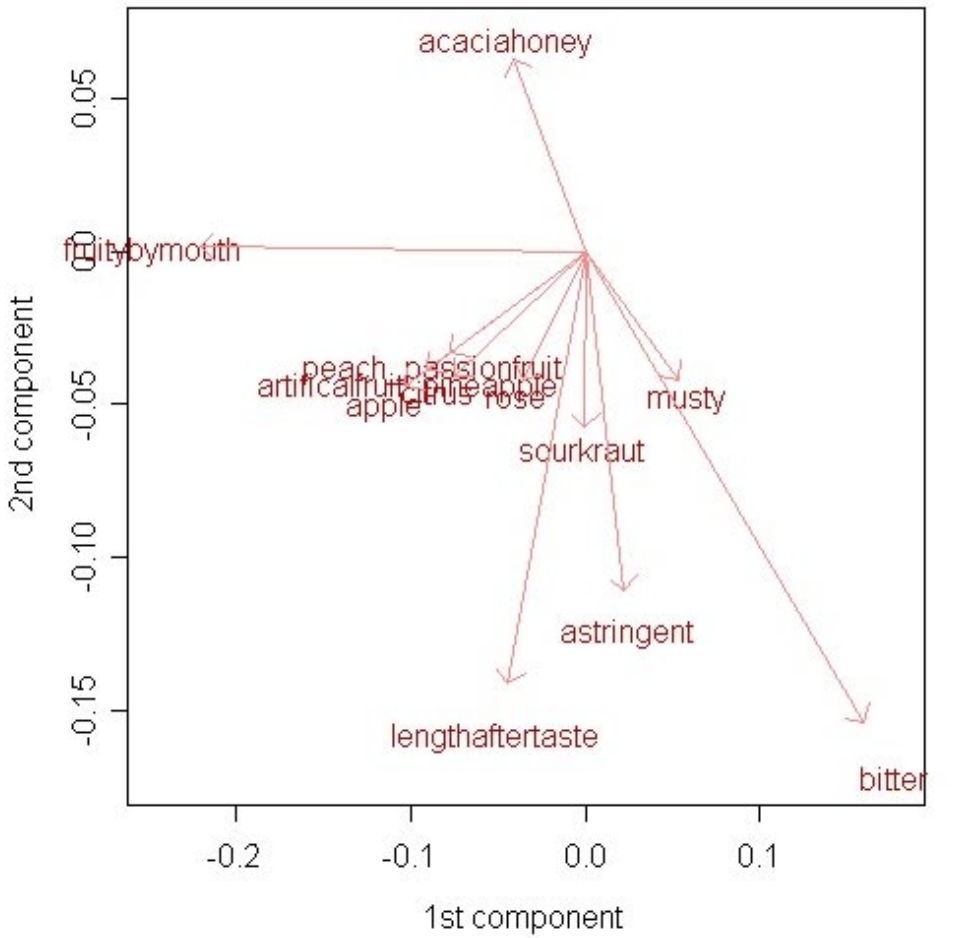
$$R^2_{crossvalidated} = 1 - PRESS/TSS$$

multivariate: maximize mean (median etc.) of $R^2_{crossvalidated}$

- **Here**: maximum of mean with **2** components
- 1st quantile below value for prediction with mean
- *bitter* and *fruity by mouth:* $R^2_{crossvalidated} \approx 0.2$.

Loadings from PLS - Regression





Conclusions

- Bad prediction → larger sample size required
- PC1 explains 18%, PC2 8%, however covariance 60%
- Relationship is probably more complex, non-linear
- Are sensory variables non-explainable by volatils ?
- Other approaches: variable selection,
  nonlinear relationship,
  more sophisticated models such as
  DPLSR, PLS-path analysis?

- One presentation cannot replace the fruitful discussion in a complete workshop

References

- Dupin, I.; Schlich, P.; Fischer, U. (1999). *Differentiation of wines produced by organic or conventional viticulture according to their sensory profiles and aroma composition*. Tagungsband des 24ten Weltkongreß für Rebe und Wein, Seite 128-135.
- Frank, I.E. and Friedman, J.H. (1993). *A Statistical View of some Chemometrics Regression Tools*, Technometrics 35(2), 109-135.
- Busch, A.S. (2001). *Verknüpfung von instrumentellen, sensorischen und Konsumentendaten.* Diplomarbeit, Universität Dortmund

F-test based on ANOVA model region + production style + region * production style + wine estate(region * production style). wine estate(region * production style) was treated as a randomized variable

	production style		region		wine estate	
degree of freedom	1		3		30	
sensory attributes	**F**	**p**	**F**	**p**	**F**	**p**
apple	0.45	0.5101	3.92	0.0206	2.03	0.0427
peach / passion fruit	2.73	0.1100	6.03	0.0029	1.00	0.4936
honey	0.09	0.7606	2.37	0.0898	0.50	0.9476
grean bean / box tree	3.98	0.0579	2.75	0.0663	4.54	0.0002
herbaceous	6.17	0.0200	1.68	0.1976	1.79	0.0788
rose / acacia blossom	3.32	0.0790	3.31	0.0352	0.88	0.6147
sauerkraut	2.13	0.1558	2.74	0.0635	1.08	0.4186
fruity by mouth	0.76	0.3904	7.18	0.0011	1.02	0.4769
bitter	0.49	0.4910	3.85	0.0219	1.75	0.0864
length aftertaste	1.50	0.2308	3.18	0.0395	0.70	0.7971
vegetative by mouth	0.57	0.4559	8.44	0.0005	1.43	0.1884