PLS regression and PARAFAC2

Less used but useful

Rasmus Bro

Chemometrics Group
Dept. Food Science
Royal Veterinary & Agricultural University (KVL)
rb@kvl.dk

Outline

The short version!

1. Multiway regression
2. PARAFAC2

Chemometric Regression

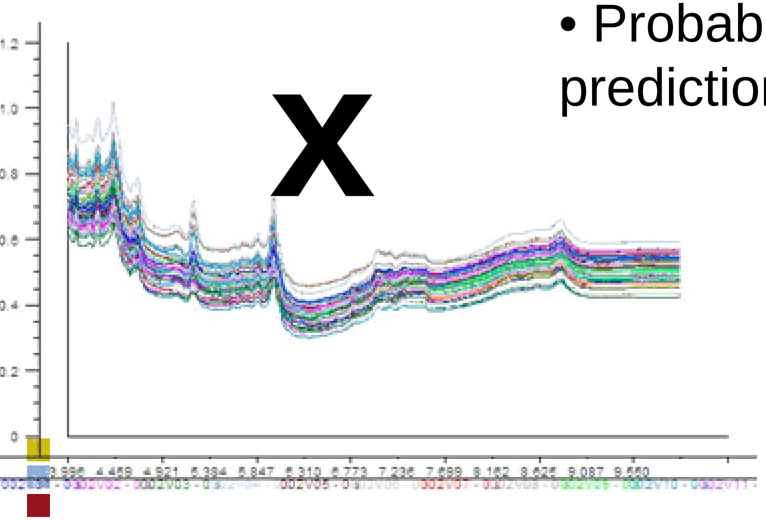
$$\min_{\mathbf{b}} \left\| \mathbf{y} - \mathbf{Xb} \right\|_F^2$$

***Example*: Online determination of activity of antidepressant**

b = X⁺y does not work

- Collinearity
- X is probably around rank 10-20 plus some 40 pure noise components
- Probably only a rank 4-10 part is needed for optimal predictions

X



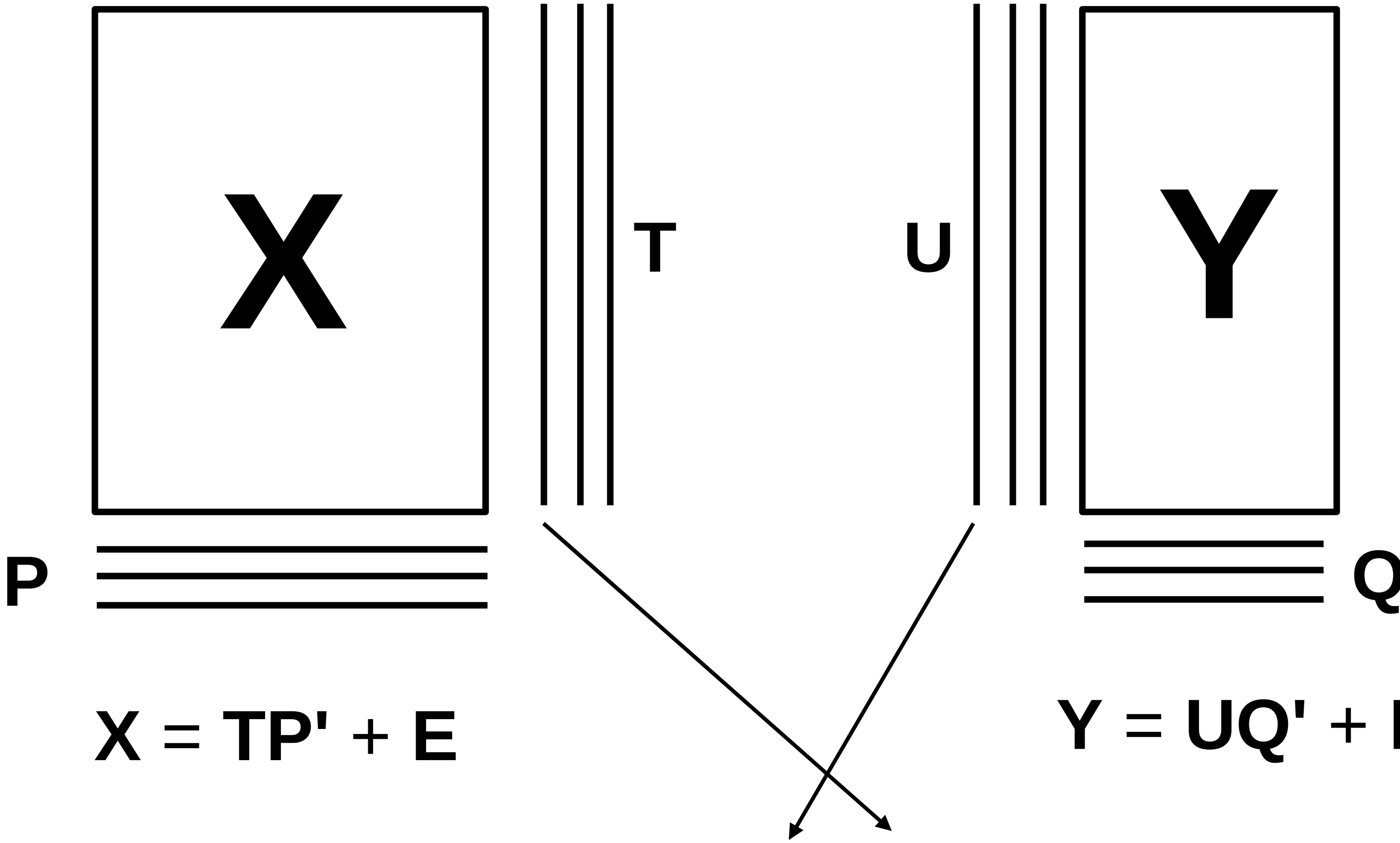
Fast method for determination measuring directly on the tablet
Near infrared spectra (X, 780 variables) - 10000-4000 cm^{-1} (1000-2500 nm), Water content from lab (y)
Courtesy, Lundbeck A/S

PLS regression

Two-way PLS principle

- Find $\mathbf{w}$ ($||\mathbf{w}||=1$) and $\mathbf{q}$ ($||\mathbf{q}||=1$) such that the one-component models of $\mathbf{X} = \mathbf{t}\mathbf{w}' + \mathbf{E}_X$ and $\mathbf{Y} = \mathbf{u}\mathbf{q}' + \mathbf{E}_Y$ have maximal covariance ($\mathbf{t}'\mathbf{u}$)
- Predict $\mathbf{u}$ from $\mathbf{t}$ ($\mathbf{t}r$)
- Subtract the model from $\mathbf{X}$ ($\mathbf{t}\mathbf{w}'$) and predictions from $\mathbf{Y}$ ($\mathbf{t}r\mathbf{q}'$)
- Proceed with the next component from residuals

PLS – Partial Least Squares regression



$$X = TP' + E$$

$$Y = UQ' + F$$

$$\mathbf{u_1} = r_1 \mathbf{t}_1$$
$$\mathbf{u_2} = r_2 \mathbf{t}_2$$
etc.

PLS – Partial Least Squares regression

Two-way PLS: how to do it - one y.

Max covariance

LS model

$$\max_{\mathbf{w}}\left[\operatorname{cov}(\mathbf{t},\mathbf{y}) \;\middle|\; \mathbf{t}=\mathbf{X}\mathbf{w}\right]=$$

$$\max_{\mathbf{w}}\left[\mathbf{t}'\mathbf{y} \;\middle|\; \mathbf{t}=\mathbf{X}\mathbf{w}\right]=$$

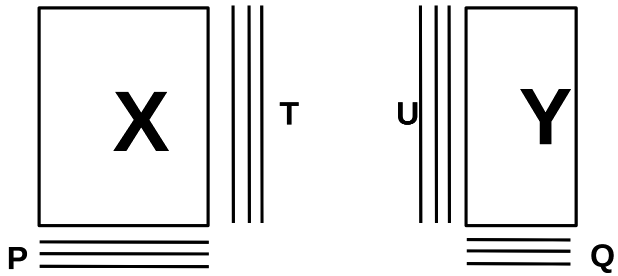
$$\max_{\mathbf{w}}\left[\mathbf{y}'\mathbf{X}\mathbf{w}\right]=$$

$$\max_{\mathbf{w}}\left[\mathbf{z}'\mathbf{w}\right]\Rightarrow$$

$$\mathbf{w}=\frac{\mathbf{X}'\mathbf{y}}{\left\|\mathbf{X}'\mathbf{y}\right\|}$$

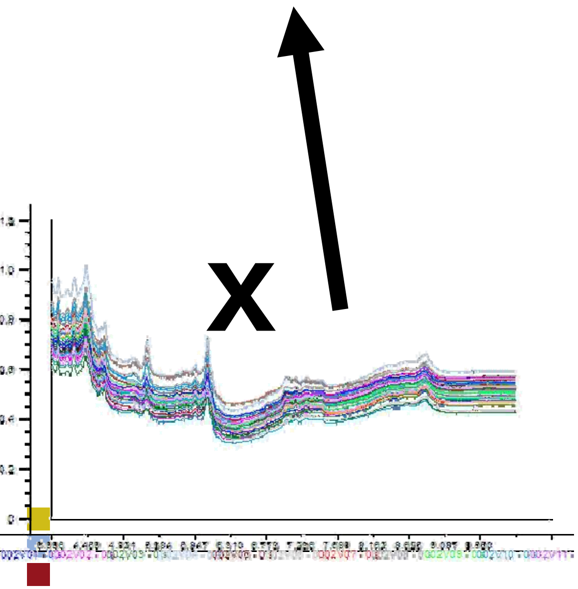
PLS – what's the advantage

Outlier check
Plus variable importance
Plus much more

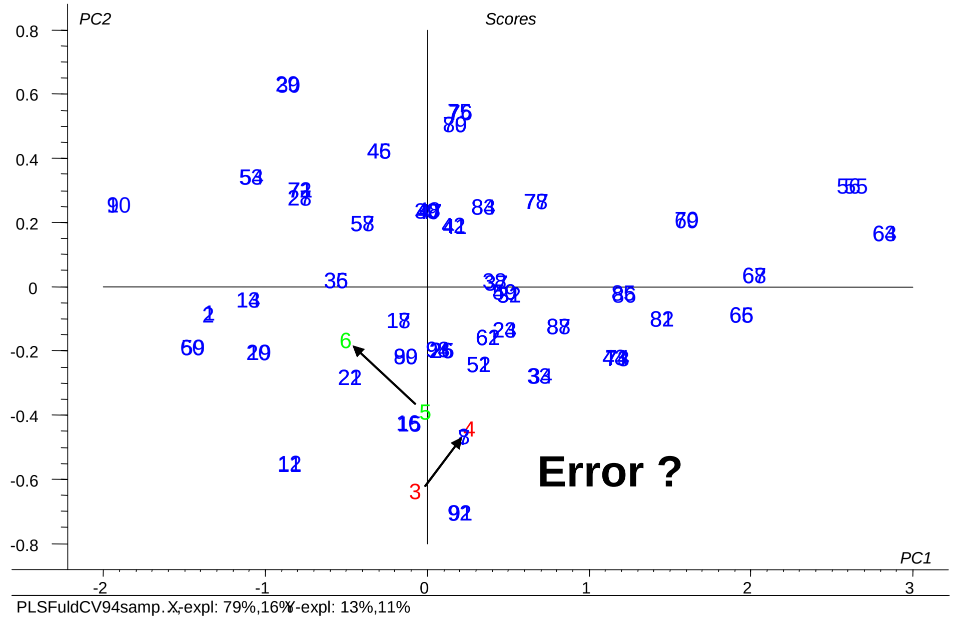


X=TP'+E

Y=UQ'+F



T(:,2)



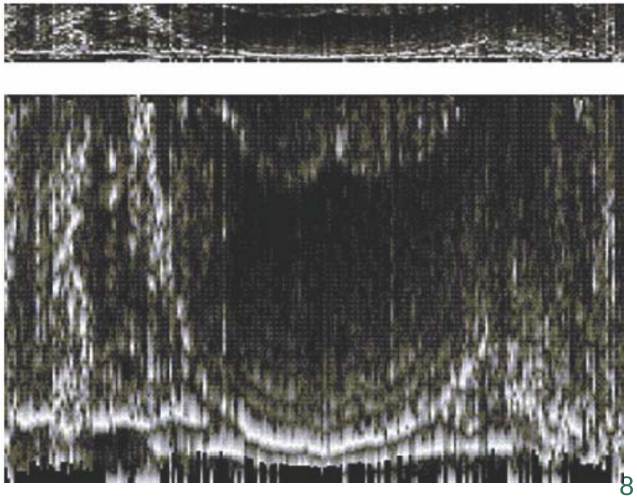
T(:,1)

Chemometric Regression

Another typical example

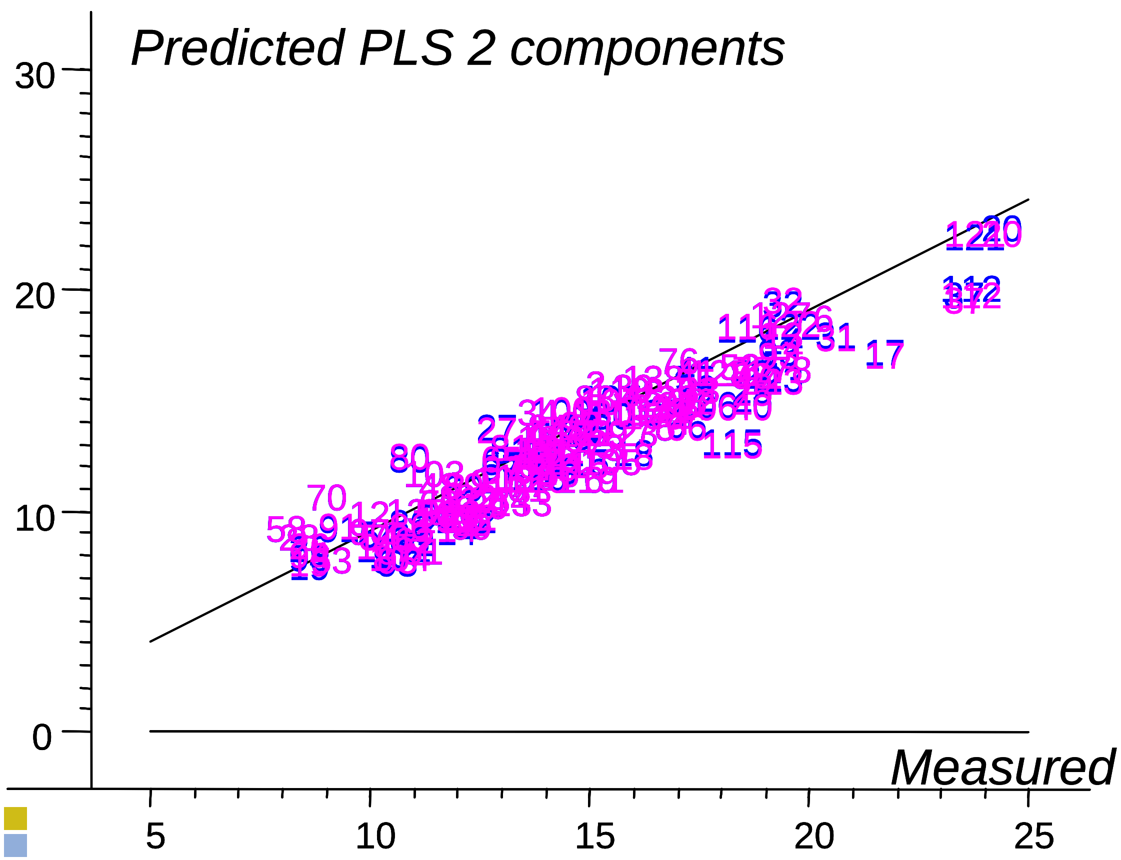
Up to 1200 carcasses per hour

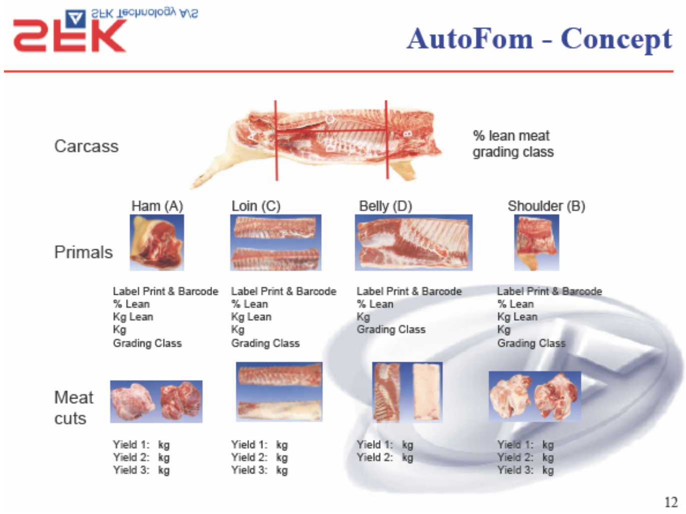
One sample of ??



Chemometric Regression

Fat Thickness



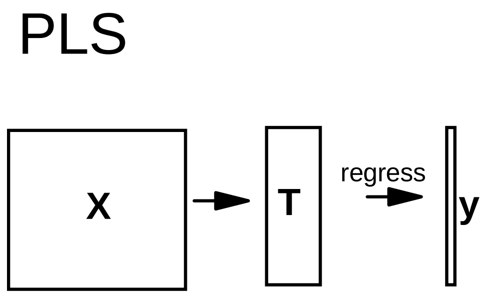


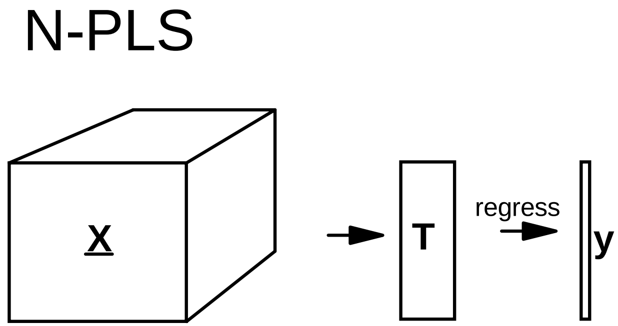
Extending PLS to multi-way

N-PLS (N-way PLS)

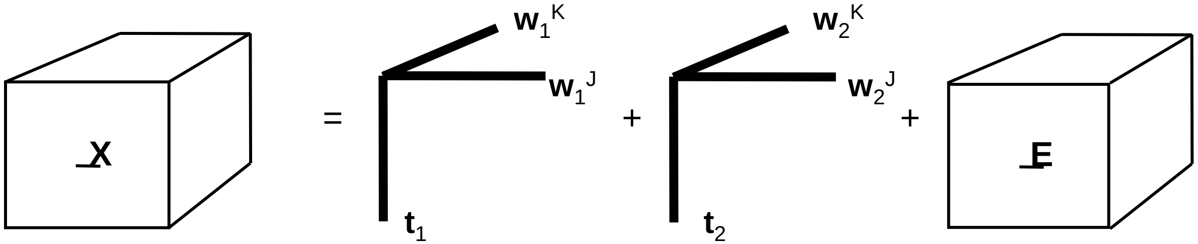
N-PLS

- For two-way data a bilinear model is used
- For three-way data a trilinear model is used

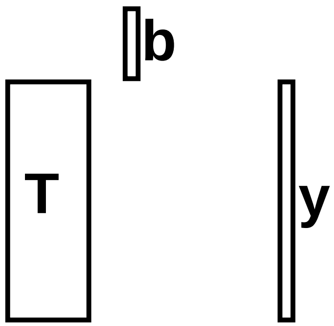




Three-way PLS regression



- Use a trilinear (PARAFAC-like) model of **X** but such that the scores are predictive of **y**.



Bro. Multiway calibration. Multi-linear PLS. Journal of Chemometrics 10:47-61, 1996.

PLS – Partial Least Squares regression

Two-way PLS: how to do it - one y.

Max covariance

$$\max_{\mathbf{w}} \left[\operatorname{cov}(\mathbf{t}, \mathbf{y}) \;\middle|\; \mathbf{t} = \mathbf{X}\mathbf{w} \right] =$$

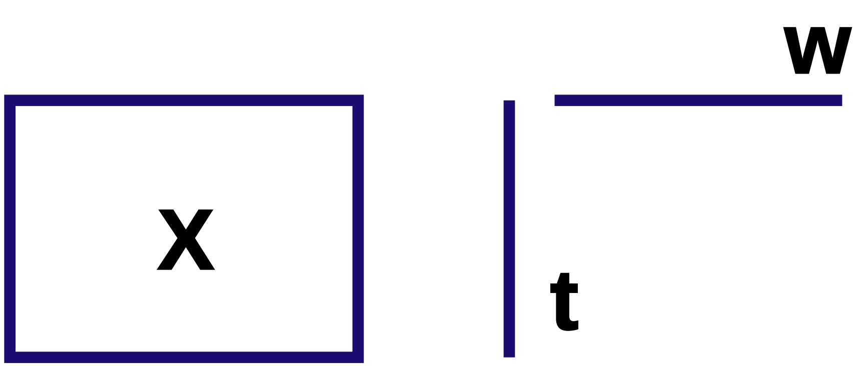
LS model

$$\max_{\mathbf{w}} \left[\mathbf{t}'\mathbf{y} \;\middle|\; \mathbf{t} = \mathbf{X}\mathbf{w} \right] =$$

$$\max_{\mathbf{w}} \left[\mathbf{y}'\mathbf{X}\mathbf{w} \right] =$$

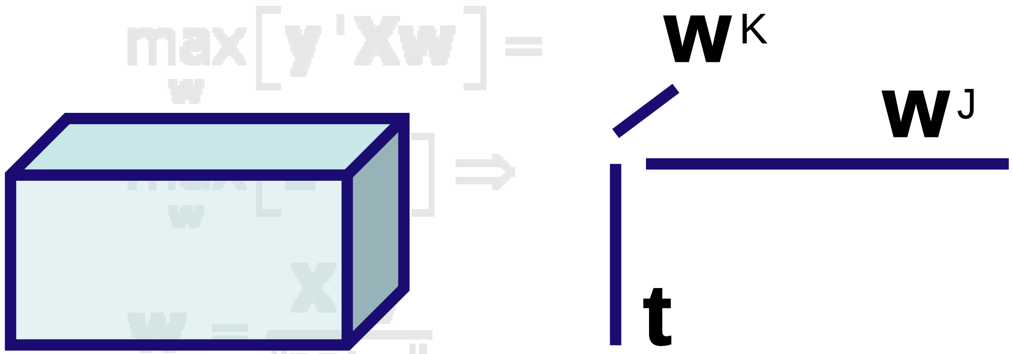
$$\max_{\mathbf{w}} \left[\mathbf{z}'\mathbf{w} \right] \Rightarrow$$

$$\mathbf{w} = \frac{\mathbf{X}'\mathbf{y}}{\left\| \mathbf{X}'\mathbf{y} \right\|}$$



Three-way PLS

Same thing but now there are two weight vectors for each component, one for each variable mode



Three-way PLS regression

Three-way PLS: how to do it - one y.

$$\max_{\mathbf{w}^J\mathbf{w}^K}\left[\operatorname{cov}(\mathbf{t},\mathbf{y}) \;\middle|\; t_i = \mathbf{w}^J\mathbf{X}_i\mathbf{w}^K\right] =$$

$$\max_{\mathbf{w}^J\mathbf{w}^K}\left[\mathbf{t}^T\mathbf{y} \;\middle|\; t_i = (\mathbf{w}^J)^T\mathbf{X}_i\mathbf{w}^K\right] =$$

$$\max_{\mathbf{w}^J\mathbf{w}^K}\left[(\mathbf{w}^J)^T\mathbf{Z}\mathbf{w}^K \;\middle|\; z_{jk} = \mathbf{y}^T\mathbf{x}_{jk}\right] \Rightarrow$$

SVD on $\mathbf{Z}$

Three-way PLS: how to do it - one y.

Table 1. Matlab code for tri-PLS1 regression of **y** ($I \times 1$) on (centred) **X** ($I \times JK$)

```
e=y;                                          % initialize
for lv=1:LV                                   % for each factor
   Z=reshape(e'*X,J,K);                       % vector of covariances→matrix
   [wJ, wK]=svd(Z);                           % find weights maximizing covariance
   WJ=[WJ wJ(:,1)];                           % save weights J-mode
   WK=[WK wK(:,1)];                           % save weights K-mode
   T=[T X*kron(wK(:,1),wJ(:,1))];             % save scores I-mode
   b=inv(T'*T)*T'*y;                          % y loadings wrt T
   e=y−T*b;                                   % residual y
end
bNPLS=0;
for 1v=1:LV
   bNPLS=bNPLS+kron(WK(:,1v),WJ(:,1v))*b(1v); %regression coeffs
end
```

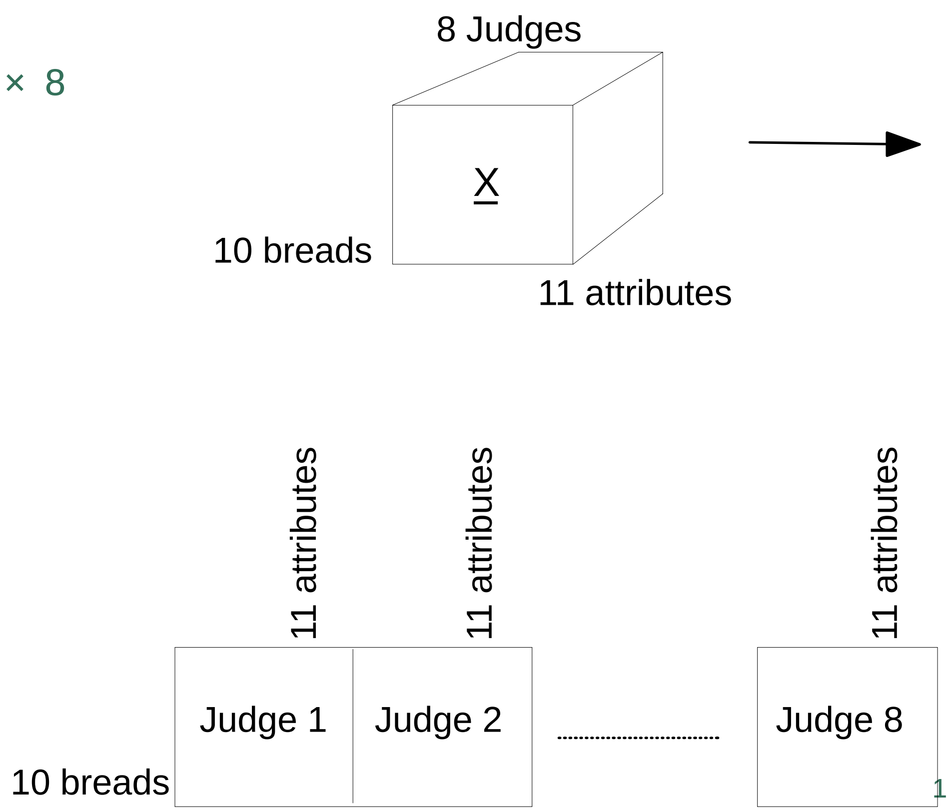
 J. Chemometrics, Vol. 12, 77–81 (1998)

N-PLS regression
Example sensory data

Three-way, two-way:

Does it make a difference?

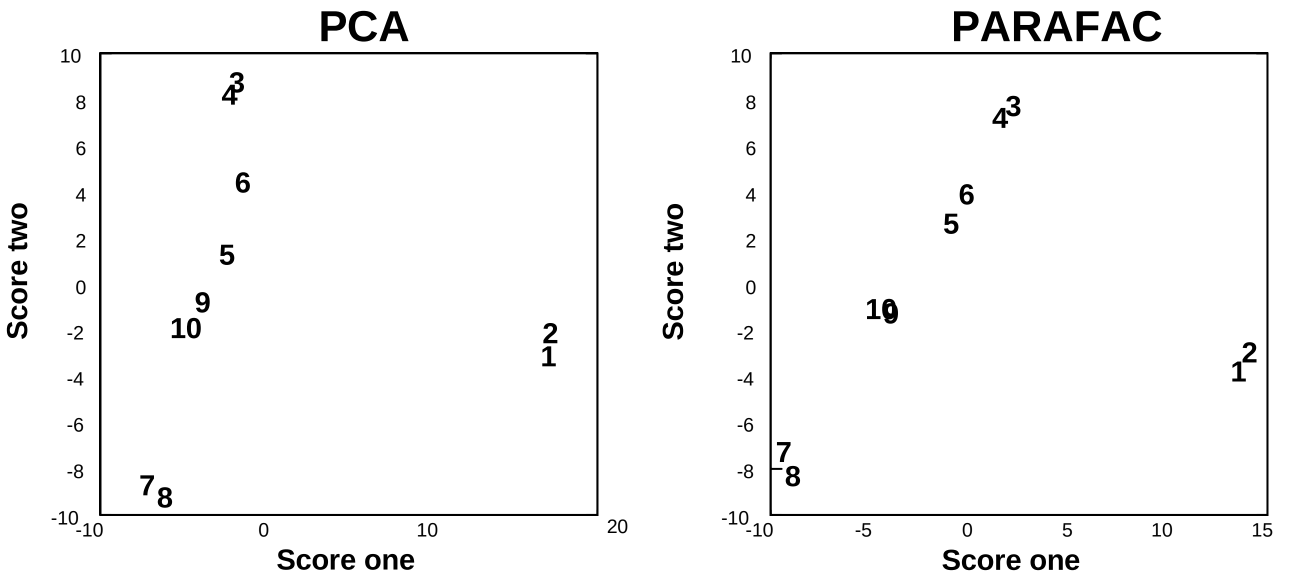
- 5 breads (in replicates) × 11 attributes × 8 judges
- Data due to Magni Martens



Sensory data

Scores from bilinear PCA and trilinear PARAFAC

Three-way more robust because of 'stronger' structural model

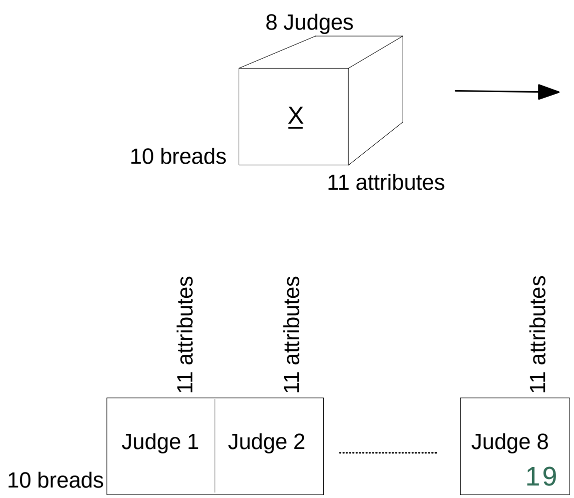


Similar but note that replicates are closer for PARAFAC

Cross-validation

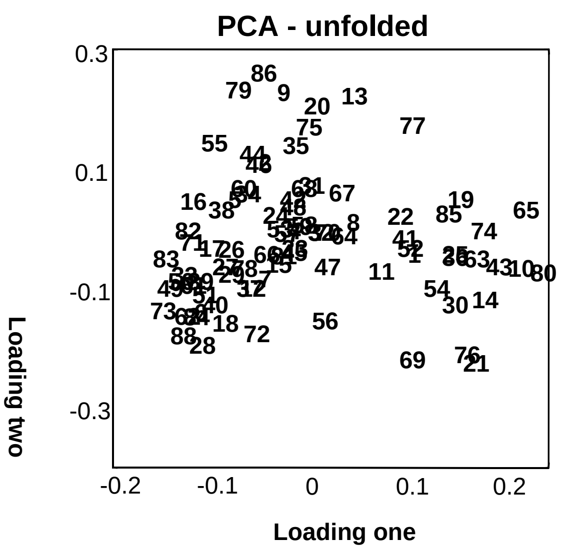
- PARAFAC fits worse but provide best predictions
- Thus nothing gained going to more complex PCA

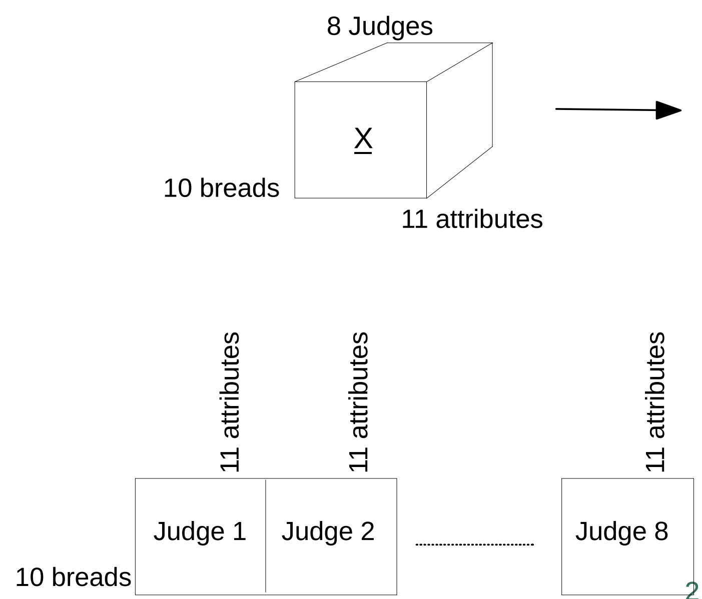
Number of components	PARAFAC		PCA	
	Fit	*Cross-va*	*Fit*	*Cross-val*
1	35.3	14.5	44.6	13.2
2	49.2	26.2	65.8	26.5
3	57.4	32.9	74.3	18.6



Sensory data

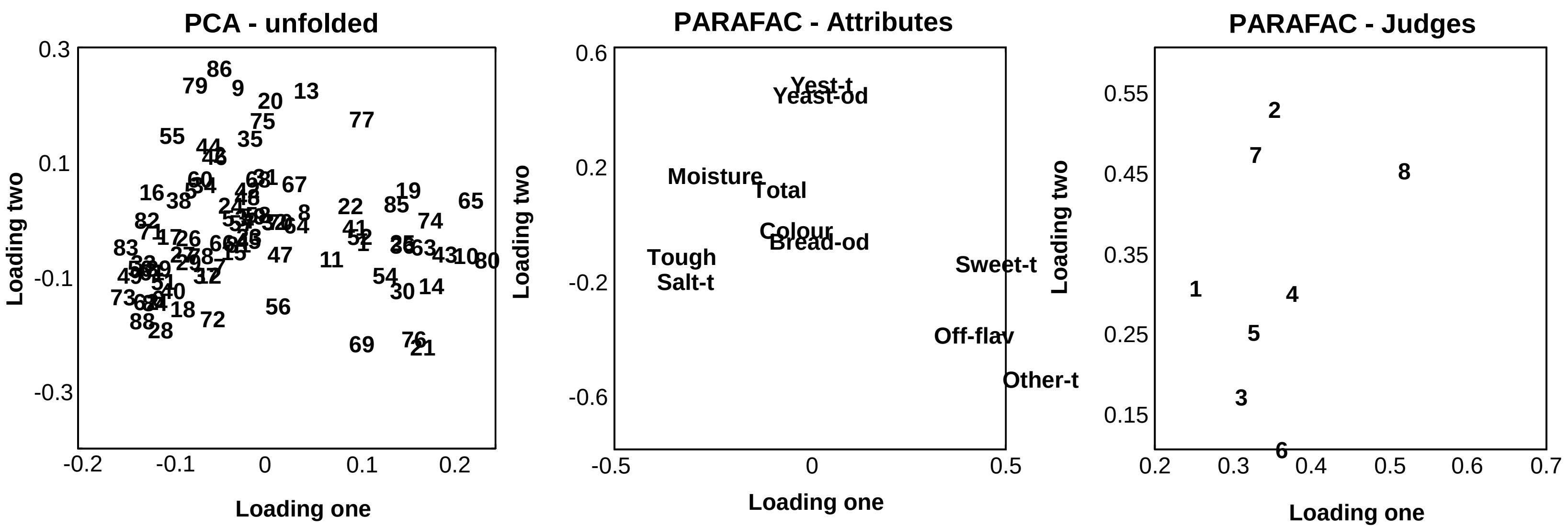
Loadings from bilinear PCA





Sensory data

Loadings from bilinear PCA and trilinear PARAFAC

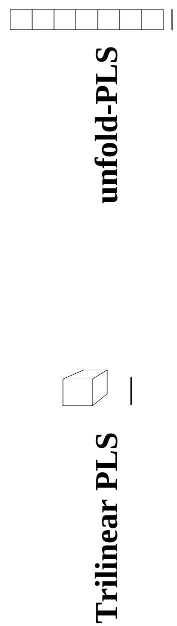


PARAFAC 19 loading-elements per component

PCA 88 loading-elements per component!

Sensory data

- Calibration - predict salt content



Sensory data

- Calibration - predict salt content
  - 25% improvement!!

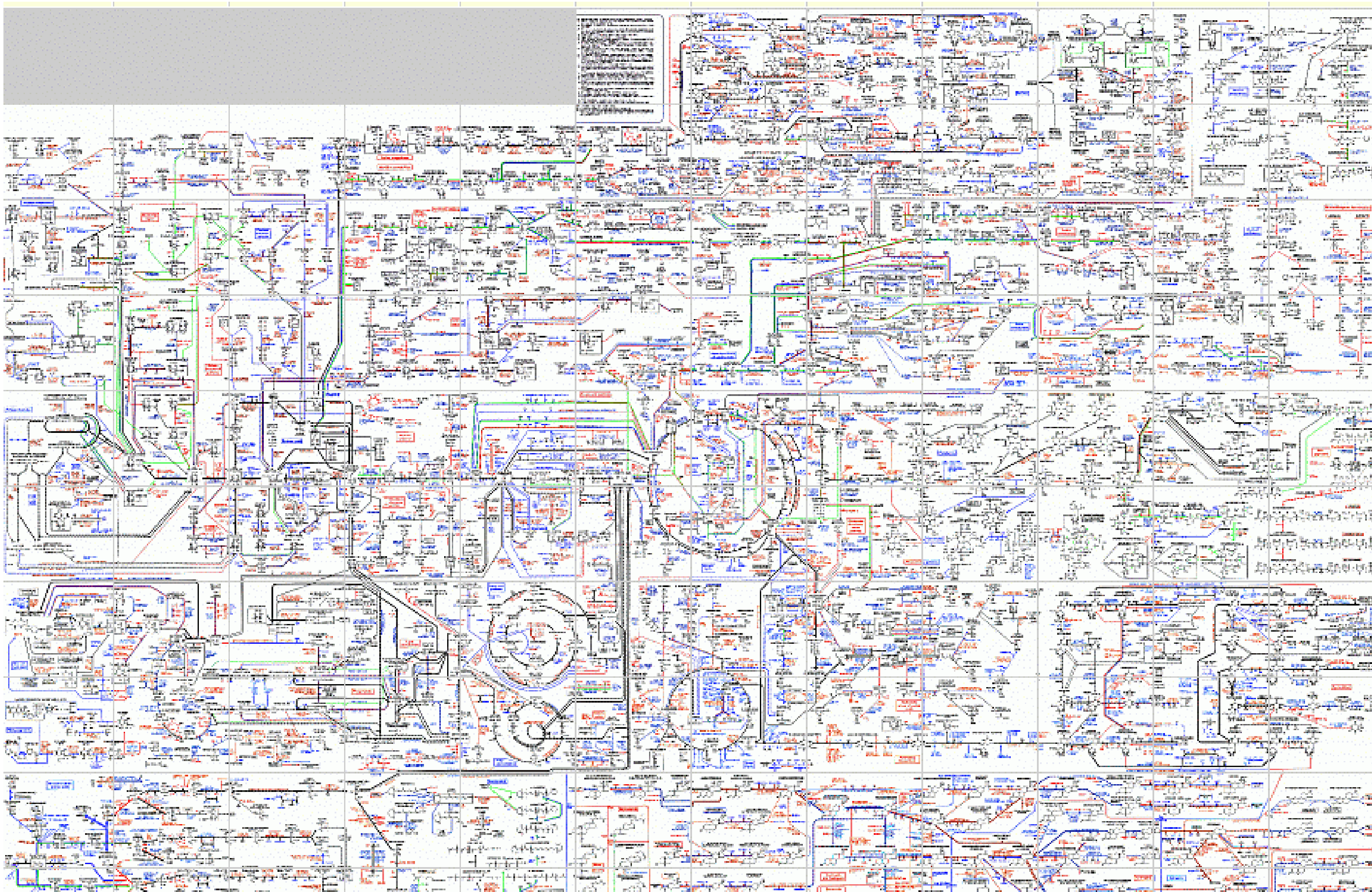
	LV	Variation explained /%				RMSE	
		X cal.	X val.	Y cal.	Y val.	Y cal.	Y val.
unfold-PLS	1	43	25	80	62	0.21	0.29
	2	61	38	95	76	0.10	0.23
	3	74	49	100	84	0.03	0.19
Trilinear PLS	1	31	22	75	60	0.23	0.30
	2	46	36	93	82	0.12	0.20
	3	54	44	98	91	0.07	0.15

Three-way more predictive
because less overfit

Multi-way analysis in metabonomics

Metabonomics “the systematic study of the unique chemical fingerprints that specific cellular processes leave behind”.

A simplified view on metabolic pathways



From Roche

Herbal medicine

Exploring hypericum

The mechanism by which St. John's Wort acts as an antidepressant is not fully understood.

The antidepressant or mood elevating effects of St. John's Wort were originally thought to be due solely to hypericin

http://www.holistic-online.com/Herbal-Med/_Herbs/h20.htm

Advocated by Hippocrates

Hypericum perforatum = St. John's wort

Herbal medicine

Available online at www.sciencedirect.com

SCIENCE DIRECT

Life Sciences 73 (2003) 627–639

ELSEVIER

Life Sciences

www.elsevier.com/locate/lifescie

Several compounds seem essential

Step by step removal of hyperforin and hypericin: activity profile of different *Hypericum* preparations in behavioral models

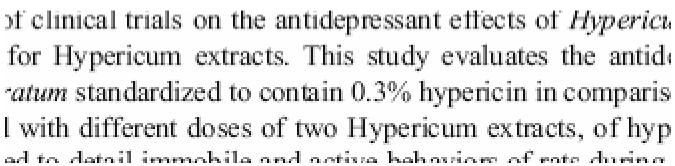
Veronika Butterweck[a,*], Volker Christoffel[c], Adolf Nahrstedt[b], Frank Petereit[b], Barbara Spengler[c], Hilke Winterhoff[a]

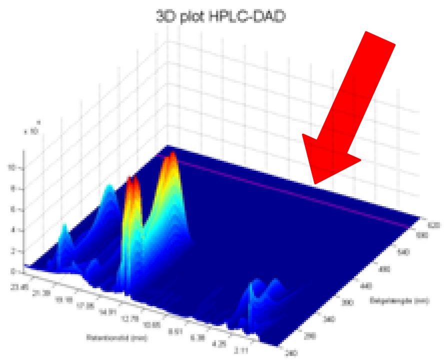
time in a dosage of 500 mg/kg after acute as well as after repeated treatment. The present results clearly show that an SJW extract free of hyperforin and hypericin exerts antidepressant activity in behavioral models, supporting our working hypothesis that flavonoids are part of the constituents responsible for the therapeutic

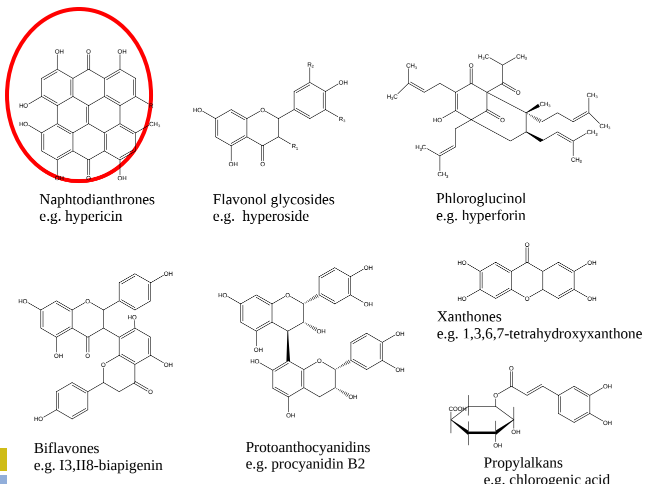
Rutin is Essential for the Antidepressant Activity of *Hypericum perforatum* Extracts in the Forced Swimming Test

Rapid Co

Herbal medicine


ɔf clinical trials on the antidepressant effects of *Hypericu*
for Hypericum extracts. This study evaluates the antid
ratum standardized to contain 0.3% hypericin in comparis
l with different doses of two Hypericum extracts, of hyp
ed to detail immobile and active behaviors of rats during





Still standardization of hypericum is done *only* with respect to hypericin!!

Problem:
Find a more satisfactory model of the relation between biochemistry and clinical effect

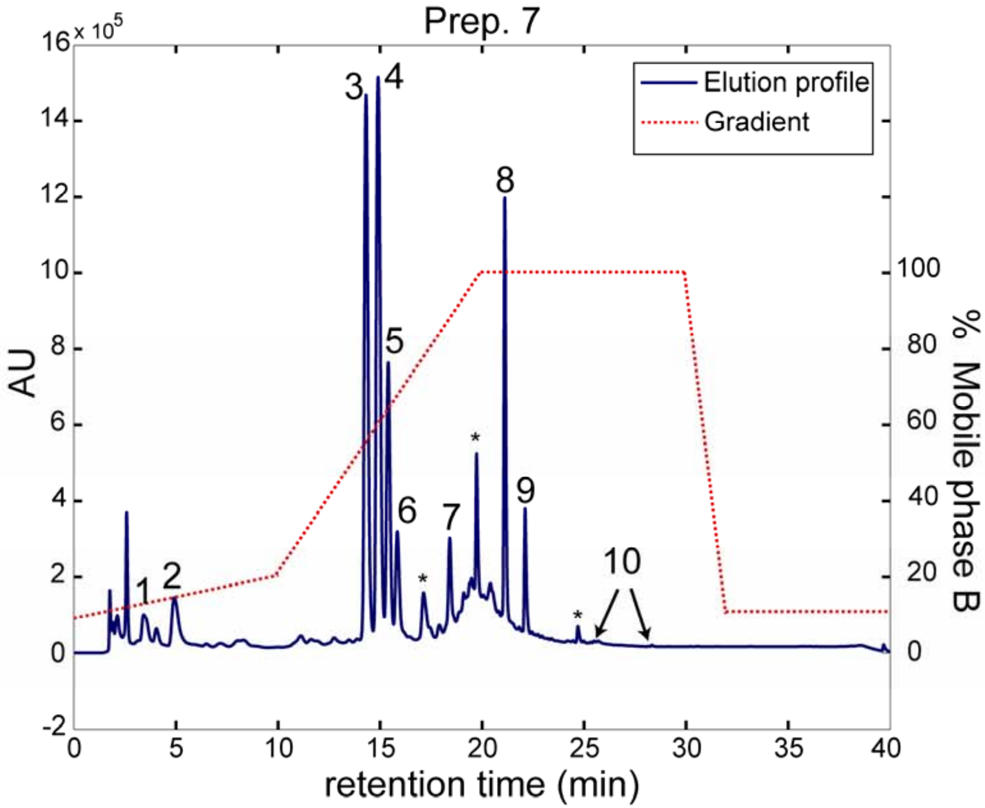
Solution:
Measure many different extracts with chemical fingerprinting (NMR,

chromatography, etc.) and correlate to physiological effect

Here:
Chromatographic data are complicated and need special care

Herbal medicine

HPLC-UVVis for separation and identification



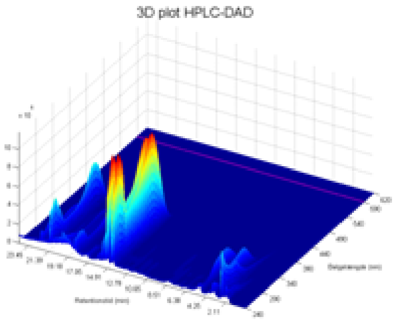
Mobile phases.
A: MeCN:H_2O 5:95 + 0.1% HCOOH
B: MeCN:H_2O 95:5 + 0.1% HCOOH

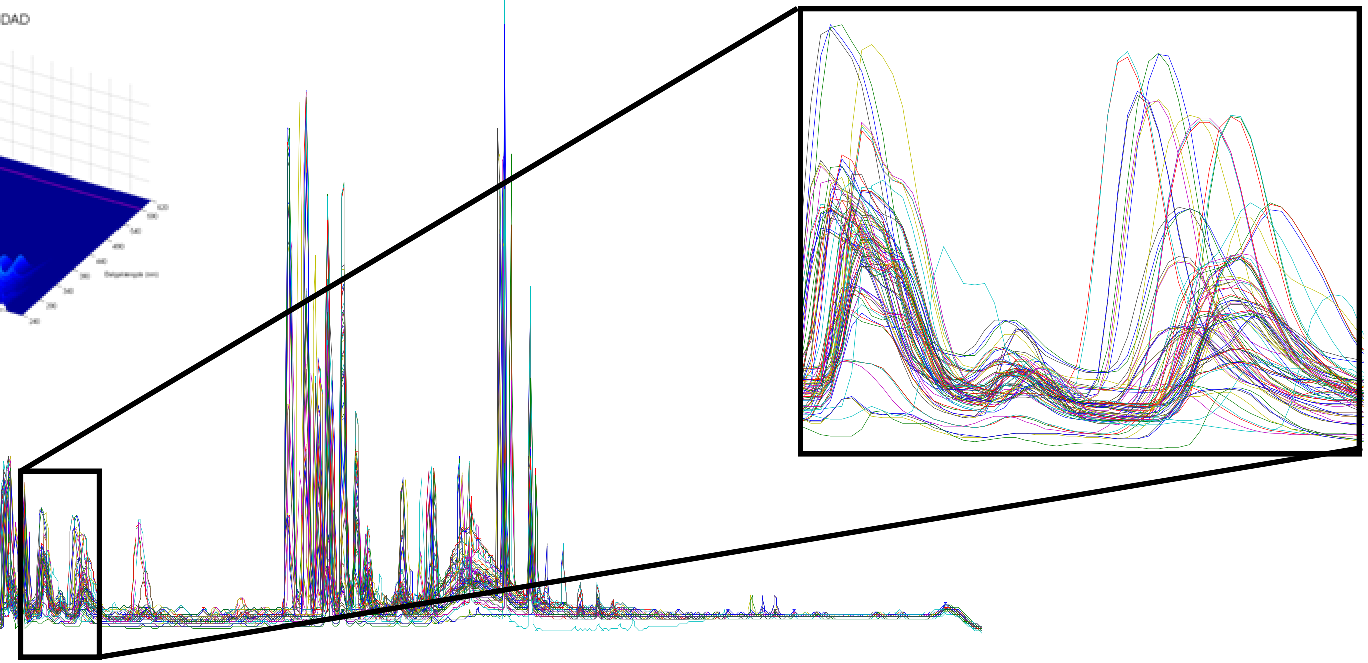
Assignment.
1. Chlorogenic acid derivative (extact structure unknown)
2. Chlorogenin acid
3. Rutin
4. Hyperoside
5. Isoquercetin
6. Quercetrin
7. Quercetin
8. I3-II8-biapigenin
9. Hypericins (extacts structures unknown)

* Structure unknown

Herbal medicine

Most peaks are identified but some are difficult

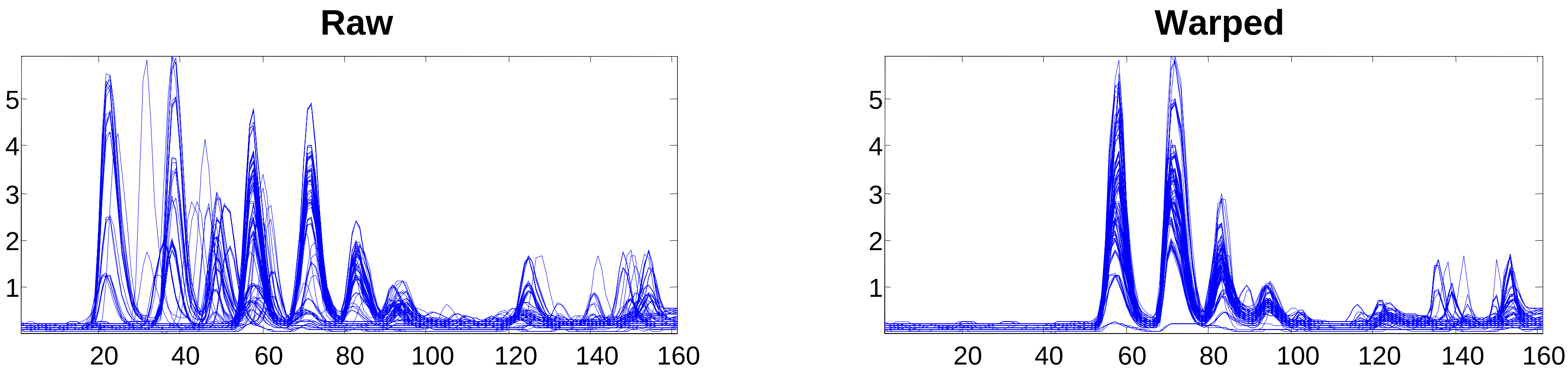




Summed profile of all samples

Herbal medicine

Warping to eliminate shift problems



JOURNAL OF CHEMOMETRICS
J. Chemometrics 2004; 18: 231–241
Published online 16 July 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/cem.859

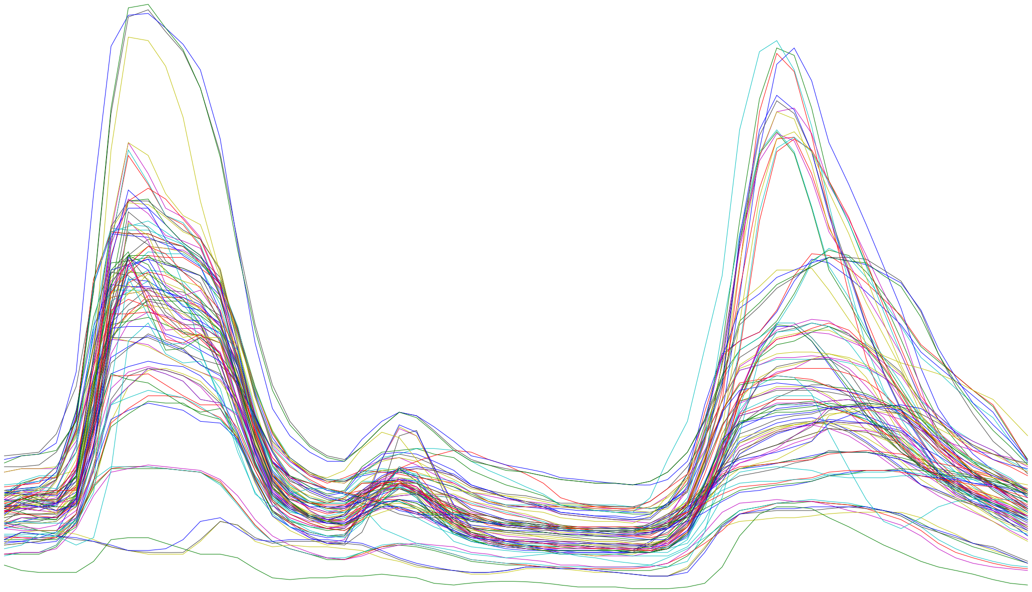
Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data

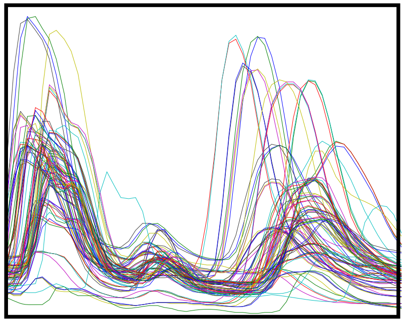
Giorgio Tomasi*, Frans van den Berg and Claus Andersson

Herbal medicine

Warping not always enough

2D warped data still have shape-difference





Herbal medicine

PARAFAC can not handle shifts and shape changes

PARAFAC(1) $\mathbf{X}_k = \mathbf{A}\mathbf{D}_k\mathbf{B}^T$

Herbal medicine

**Actually it is more general than shifts but it's a feasible approximation to what PARAFAC2 can handle*

PARAFAC2 for handling shifts*

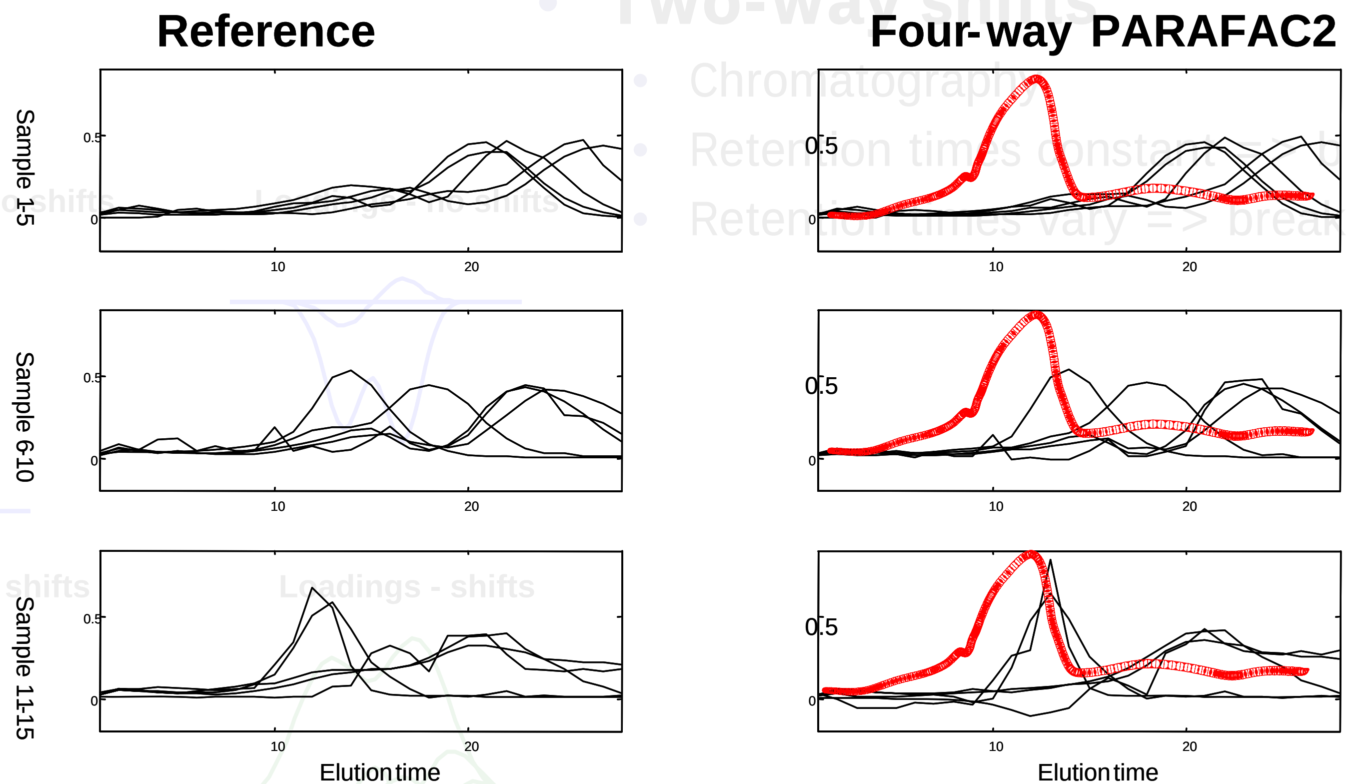
PARAFAC2 $\quad \mathbf{X}_k = \mathbf{A}\mathbf{D}_k\mathbf{B}_k^T \quad$ subject to $\mathbf{B}_k^T\mathbf{B}_k$ constant

PARAFAC(1) $\quad \mathbf{X}_k = \mathbf{A}\mathbf{D}_k\mathbf{B}^T$

R. A. Harshman. *UCLA working papers in phonetics* 22:30-47, 1972.
H. A. L. Kiers, J. M. F. ten Berge, R. Bro. *J. Chemom.* 13:275-294, 1999.
R. Bro, C. A. Andersson, H. A. L. Kiers. *J. Chemom.* 13:295-309, 1999.

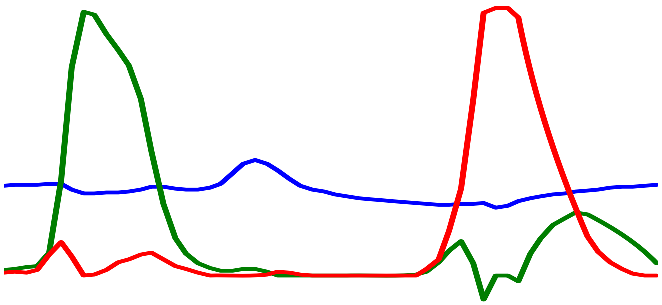
Herbal medicine

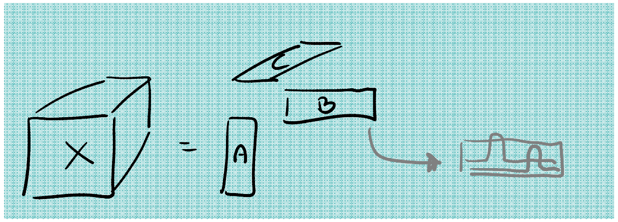
- **Two-way shifts**
  - Chromatography
  - Retention times constant => bilinear data
  - Retention times vary => breakdown

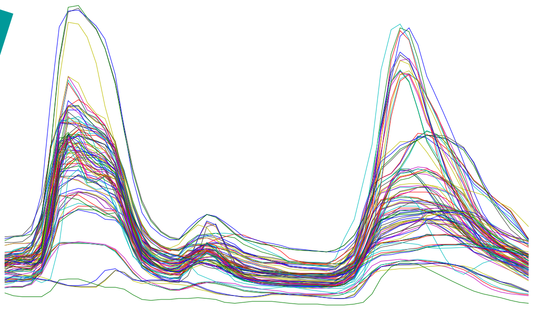


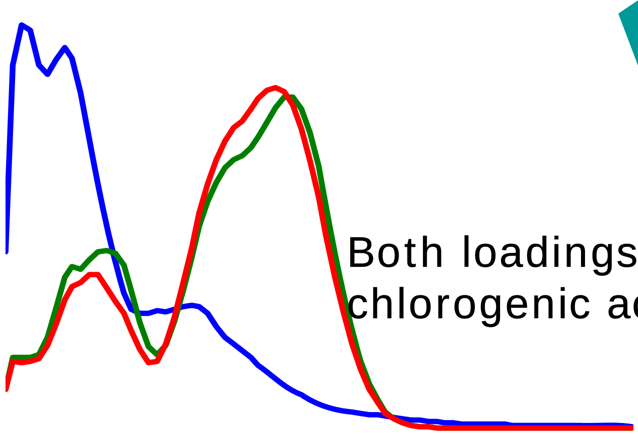
Herbal medicine

PARAFAC shows one eluent and two almost identical spectral peaks



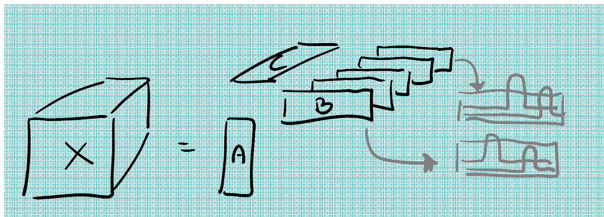




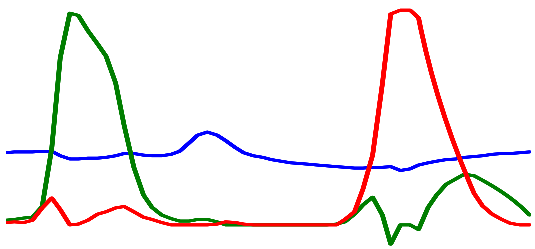


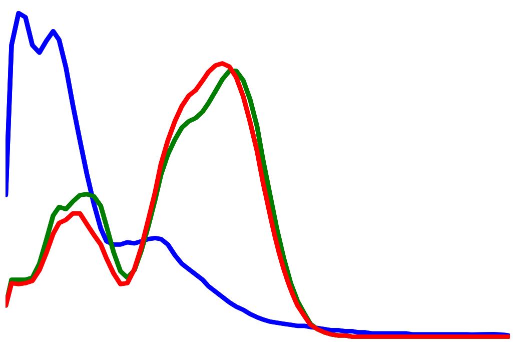
Both loadings similar to chlorogenic acid

Herbal medicine

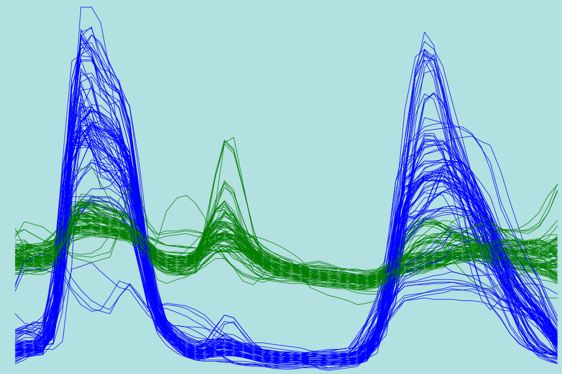


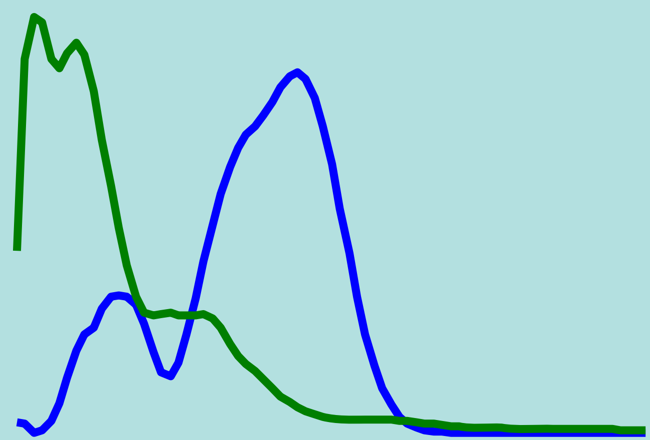
PARAFAC result from 93 samples





PARAFAC2 result from 93 samples





Herbal medicine

Next step:

Use PARAFAC, Warping, PARAFAC2 to get concentrations of specific chemicals

Perform clinical tests and relate to measured patterns

Perform verifying tests based on results

The end

N-PLS Regression

Combines good predictions with powerful visual tools

PARAFAC2

Makes it possible to fit PARAFAC even when it's impossible!!!

Multi-way papers, dissertations etc. available at www.models.kvl.dk/tricap2006.zip