



UNIVERSIDAD
POLITECNICA
DE VALENCIA

Multiway analysis: A practitioner's perspective



***José Manuel Prats-Montalbán
&
Alberto Ferrer***

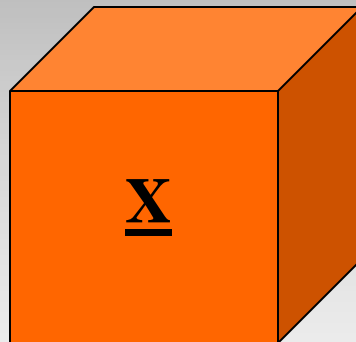
Multivariate Statistical Engineering Group
<http://mseg.webs.upv.es/>

Dp. of Applied Statistics, Operations Research & Quality

Universidad Politécnica de Valencia, Spain



Why and **when** should we use multiway models when dealing with multiway data?



Our “client” is a **practitioner**, not a psychometrician nor a chemometrician (or a statistician)



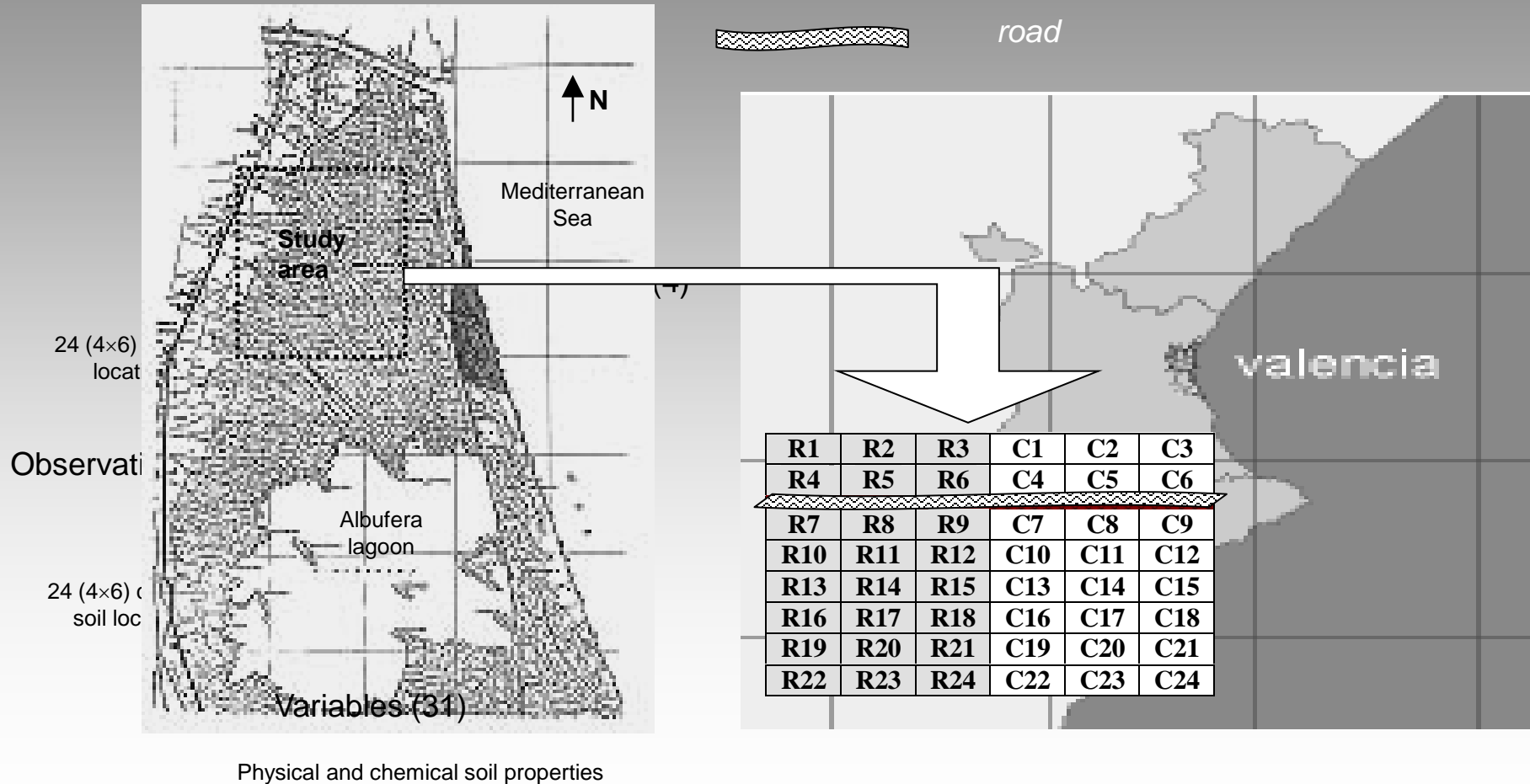
3 different applications of N-way models
based on professor Bro's N-way Toolbox
<http://www.models.kvl.dk/source/>

1. Enviromental data application: characterization of soils irrigated with wastewater. Tucker3
2. Bioinformatics application:
 - Tucker3 for integrative descriptive analysis of gene expression data
 - N-PLS for relationships between genes expression, physiological and metabolomic data
3. Industrial application: comparison of Unfold-PCA vs Tucker3 for fault detection and diagnosis in a SBR



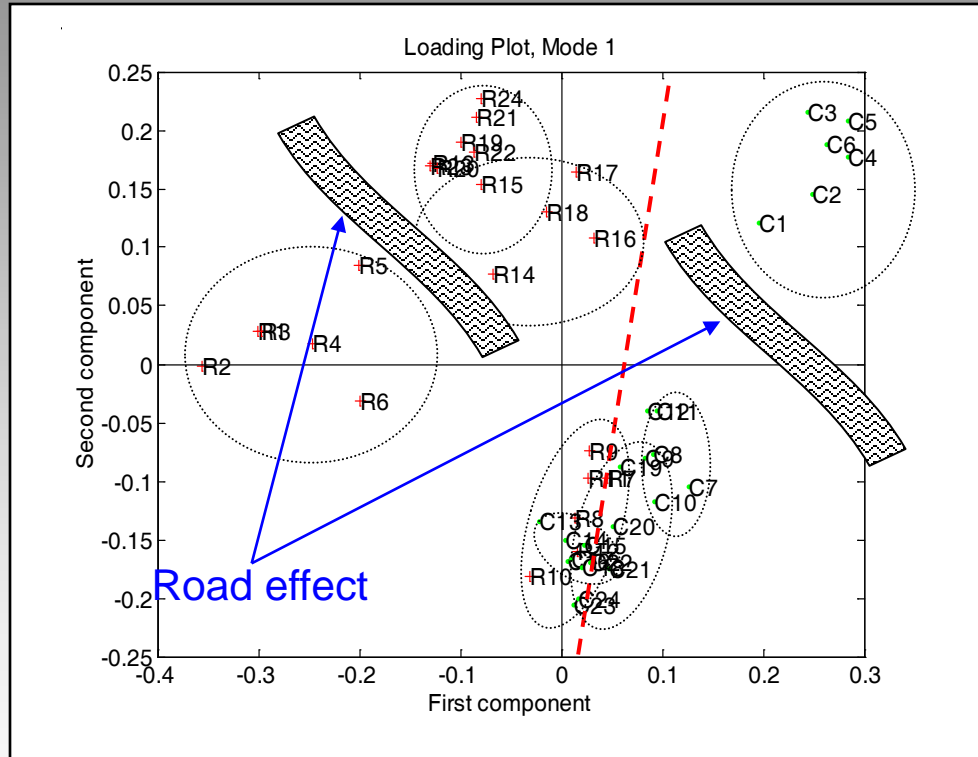
Characterization of soils irrigated with untreated urban wastewater

Goal: to characterize the different types of soils corresponding to citric fruits and rice, with the target of determining those variables which affect them

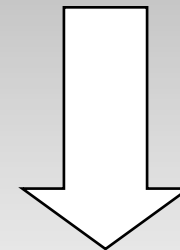


Characterisation of soils irrigated with untreated urban wastewater

Descriptive study: diagonalize, two components



First component separates the two types of soils



One model for each type of soil

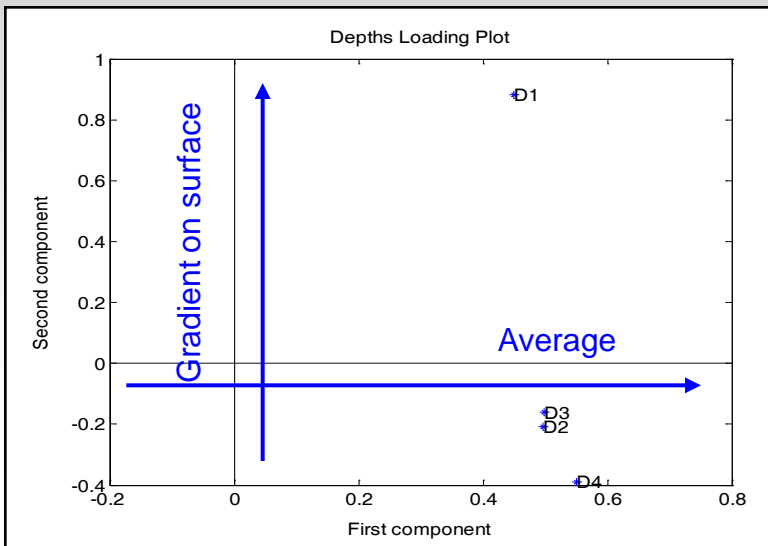
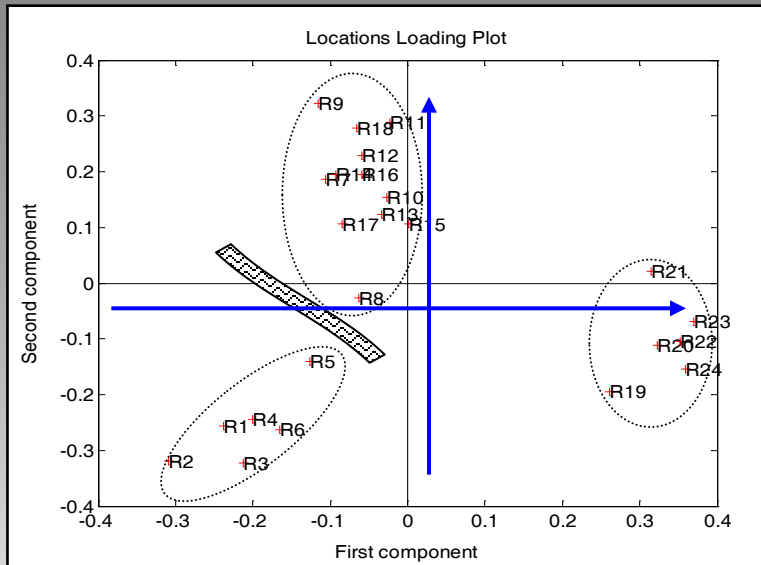
Locations loading plot for the first two diagonalized components.

R = rice soils, *C* = citric fruit soils.

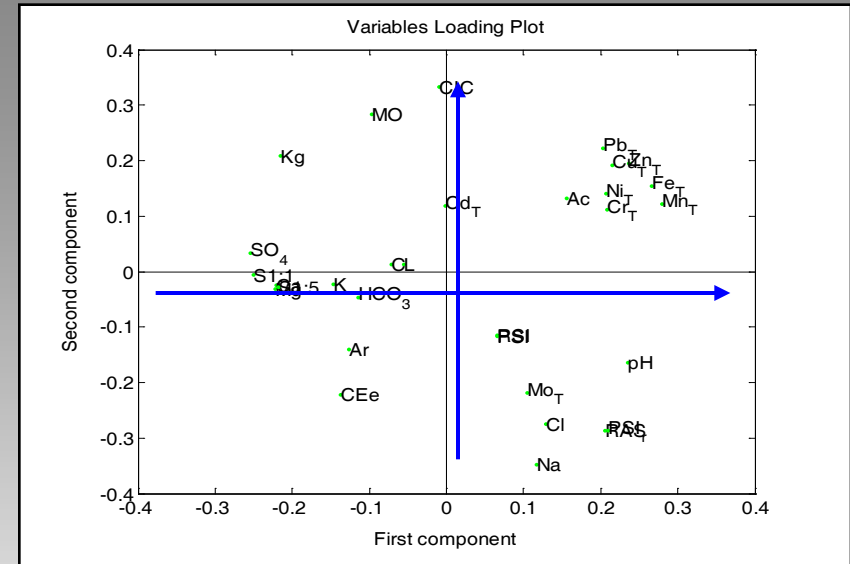


Characterisation of soils irrigated with untreated urban wastewater

Model for the rice soil



(2, 2, 2) Tucker3 model, 45.54%



D1 correspond to 0-10 cm; D2 to 10-20 cm; D3 to 20-40 cm and D4 to 40-60 cm depths.

Elements	Expl. Var. (SS)	Core
(1, 1, 1)	58.13182%	28.22815
(2, 2, 1)	23.59816%	-17.98517
(2, 2, 2)	11.58465%	-12.60135
(1, 1, 2)	4.69699%	8.02390



Characterisation of soils irrigated with untreated urban wastewater

Model for the rice soil

- Heavy metals positively correlated to Ac (Clay), and with CIC (Cation exchange capacity) and MO (Organic matter)

Heavy metals + MO \longrightarrow Organic-metallic complexes

- High correlation between Fe_T, Mn_T and Pb_T, Cr_T
Highly reduced conditions (rice farming) \longrightarrow Immobilization and precipitation of heavy metals

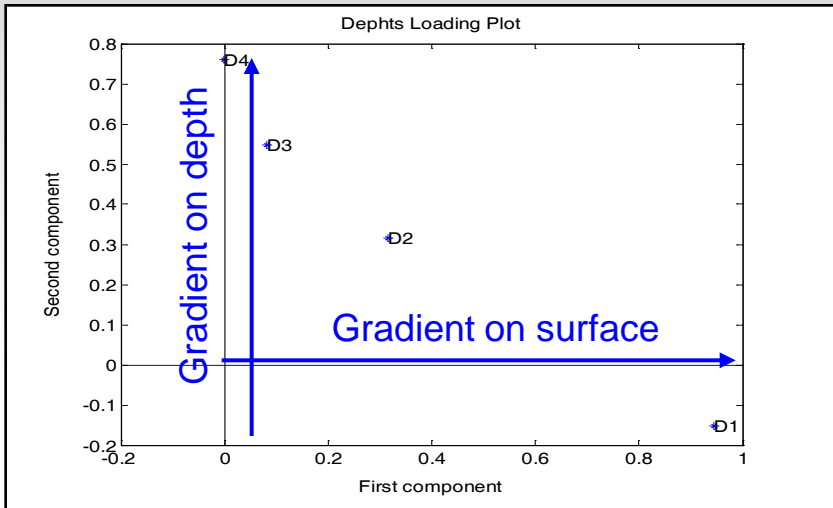
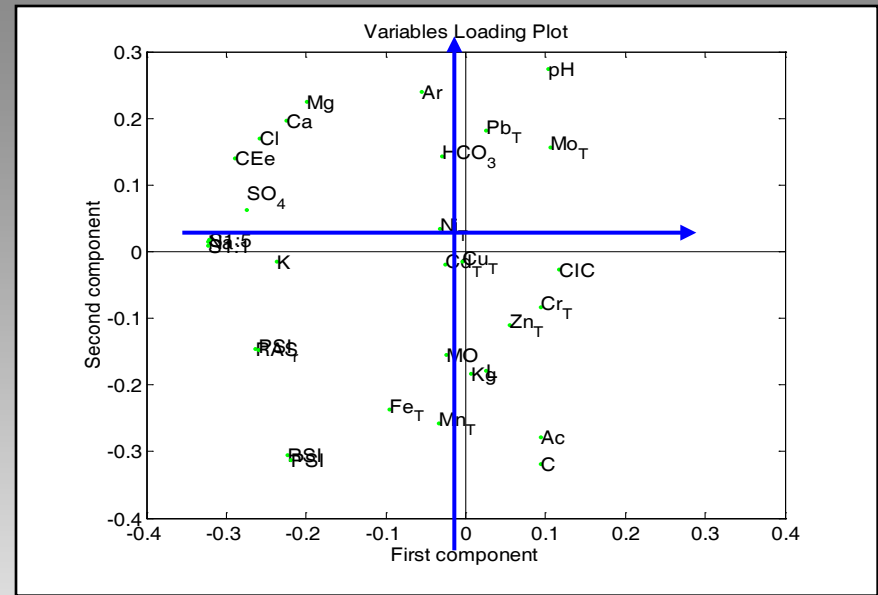
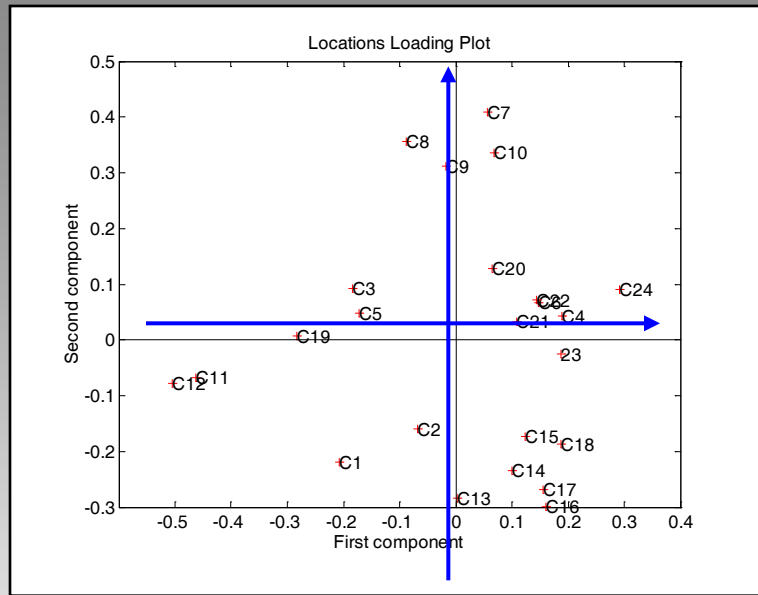
- Difference between the surface and the rest of soil strata (gradient on surface)



Characterisation of soils irrigated with untreated urban wastewater

Model for the citric fruit soil

(2, 2, 2) Tucker3 model, 36.66%

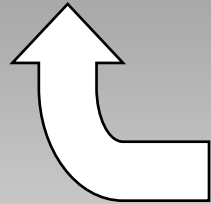


Elements	Expl. Var. (SS)	Core entry.
(1, 1, 1)	55.00200%	24.36756
(2, 2, 2)	24.15224%	16.14736
(2, 2, 1)	12.50809%	11.62033
(1, 1, 2)	5.49021%	7.69870



Model for the citric fruit soil

- *Evolution of the variables with depth (gradient also in depths)*



Citric fruit soils remain without mixing

- *Segregation within one parcel (locations C7-C10 vs C11-C12) due to high and low grounds.*
- *Again, Fe_T and Mn_T are highly correlated*



Conclusions

Tucker3 provides easily interpretable loading plots for environmentalists.

Interpretation requires core inspection, even when maximal variance or diagonality is asked for.

In this environmental study:

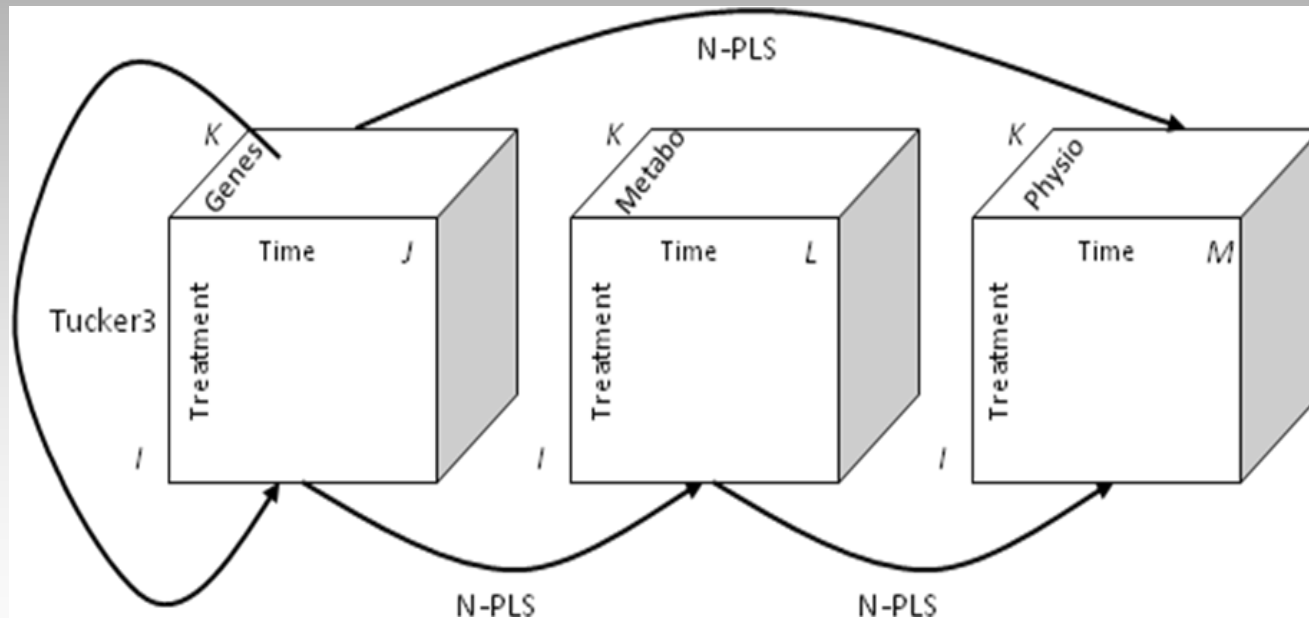
- *Rice parcels seem to be more homogeneous than the citric fruit parcels (first mode)*
- *Rice soils do not evolve with depth, while for citric fruit soils this evolution does exist (third mode)*
- *Most important variables have been isolated (second mode)*

Monitoring?



Goals

- to understand evolution in time and correlation between genes
- to describe relationships between genes & metabolomic & physiological data
- to show up the amount of information provided by metabolomic over transcriptomic data for describing physiological data



Data

2665 transcriptomic, 310 metabolomic and 19 physiological data
5 treatments (UT, CO, Low, Med, High)
3 times (6, 24 and 48 hours)

Models

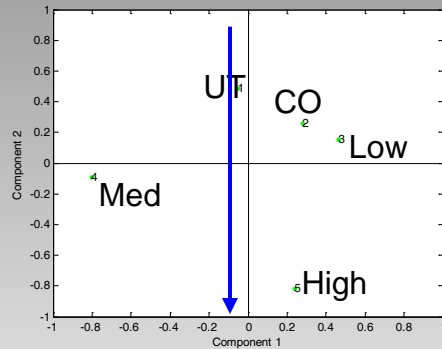
Tucker3
N-PLS



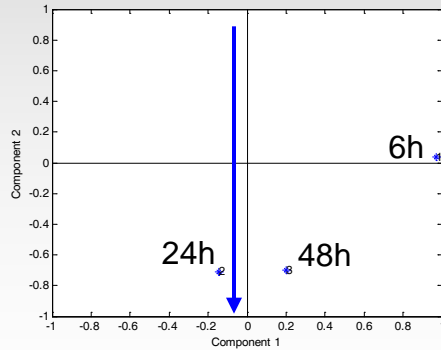
Integration of high dimensional arrays in omics data

Describing transcriptomic data

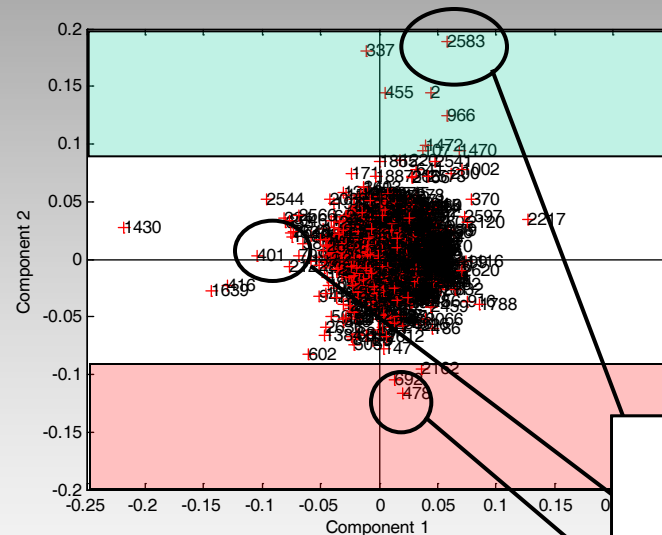
(2,2,2) Tucker3 model (rotated), 42.87% explained, centred across treatments, not scaled



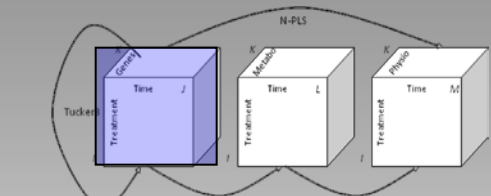
Treatments



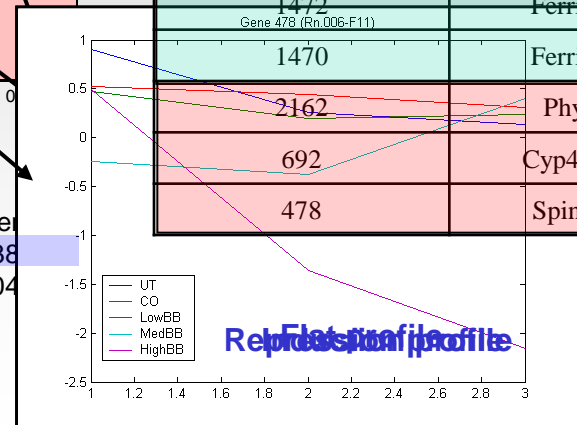
Time



Element	Exp. Var. (SS)	Core entry Sq.	Core ex
(2, 2, 2)	62.85422%	19.15303	366.838
(1, 1, 1)	35.29682%	14.35285	206.004



Gene Number	Gene name
2583	Gsta2
337	Ephx1
455	Akr7a3
2	Gsta2
966	Gstm1
1472	Ferritin
Gene 478 (Rn.D06-F11)	Ferritin
1470	Ferritin
2162	Phyh
692	Cyp4a22
478	Spin2b

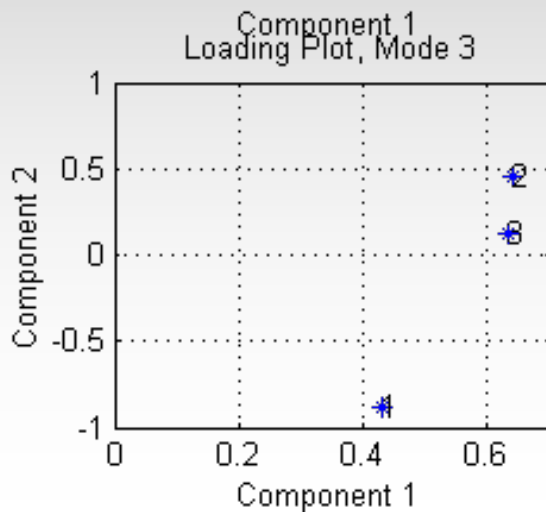
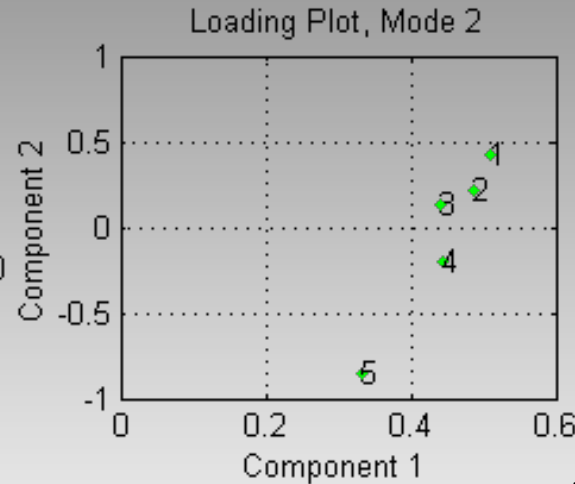
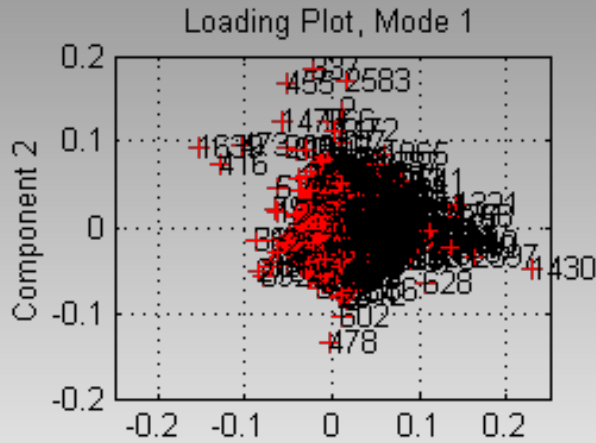
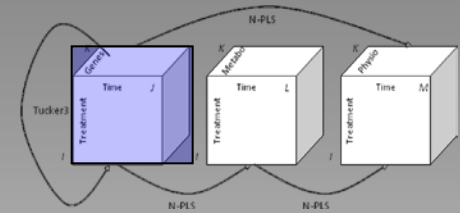


Integration of high dimensional arrays in omics data

Other pre-processings

No centring nor scaling, Tucker 3 [2 2 2]

63.92% explained variance



No biological meaning

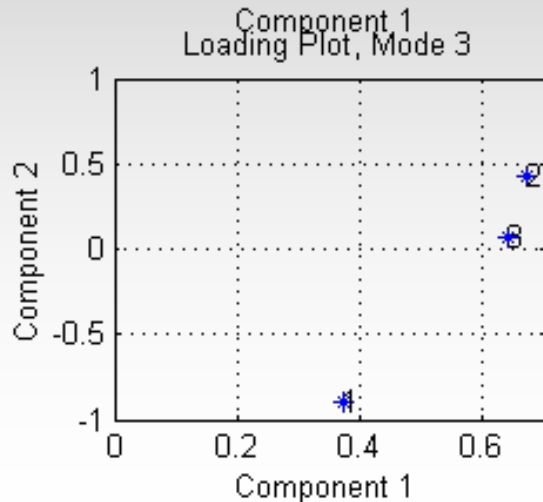
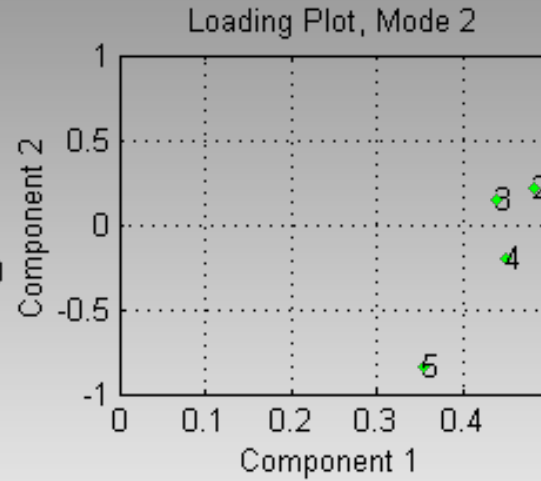
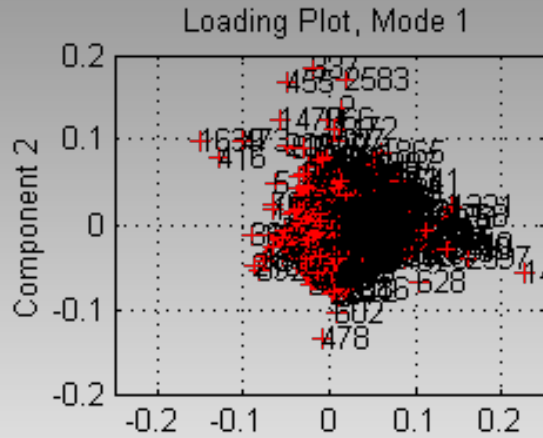
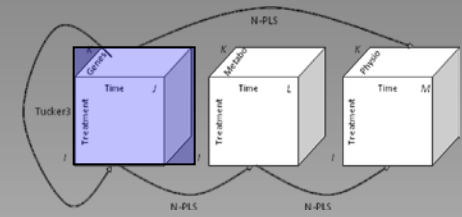
1	(1, 1, 1)	83.88774%	-43.83053	1921.11493
2	(2, 2, 1)	11.08756%	-15.93475	253.91636
3	(2, 2, 2)	2.58702%	-7.69710	59.24540
4	(1, 2, 2)	1.62071%	-6.09228	37.11591



Integration of high dimensional arrays in omics data

Other pre-processings

Centring across genes
Tucker 3 [2 2 2], 63.85%



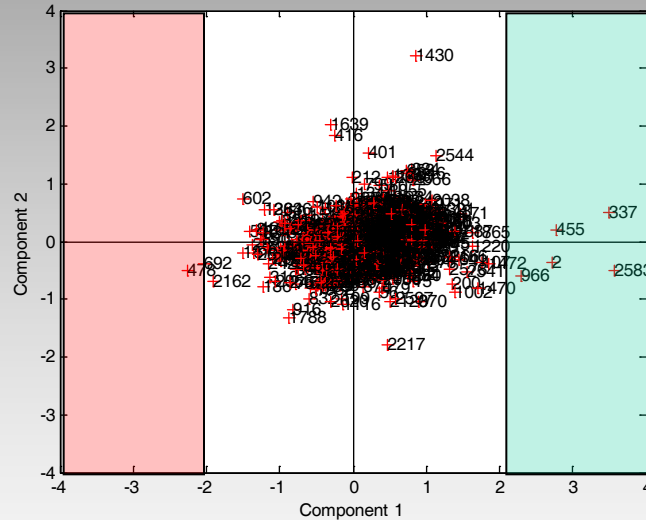
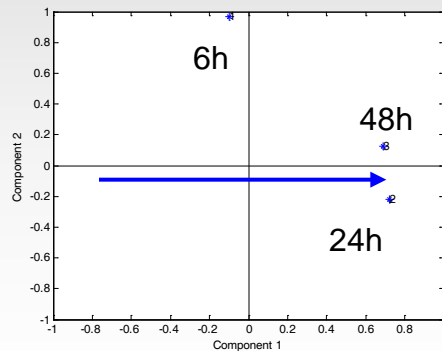
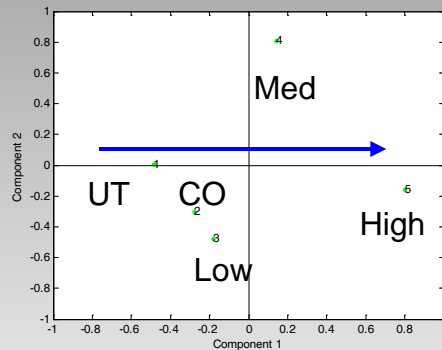
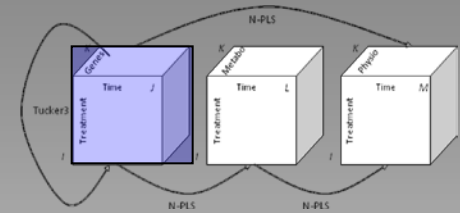
No biological meaning

1	(1, 1, 1)	83.32917%	-43.54066	1895.78938
2	(2, 2, 1)	11.76709%	-16.36180	267.70841
3	(2, 2, 2)	1.73950%	6.29083	39.57459
4	(1, 2, 2)	1.64884%	-6.12471	37.51207



Comparison with PARAFAC

2 components, for comparison
3 components yielded degeneracy



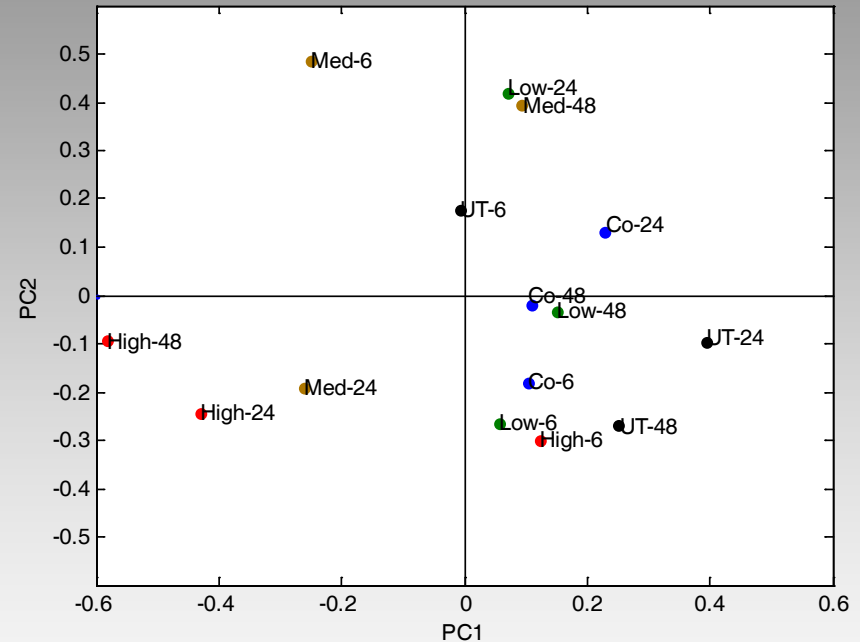
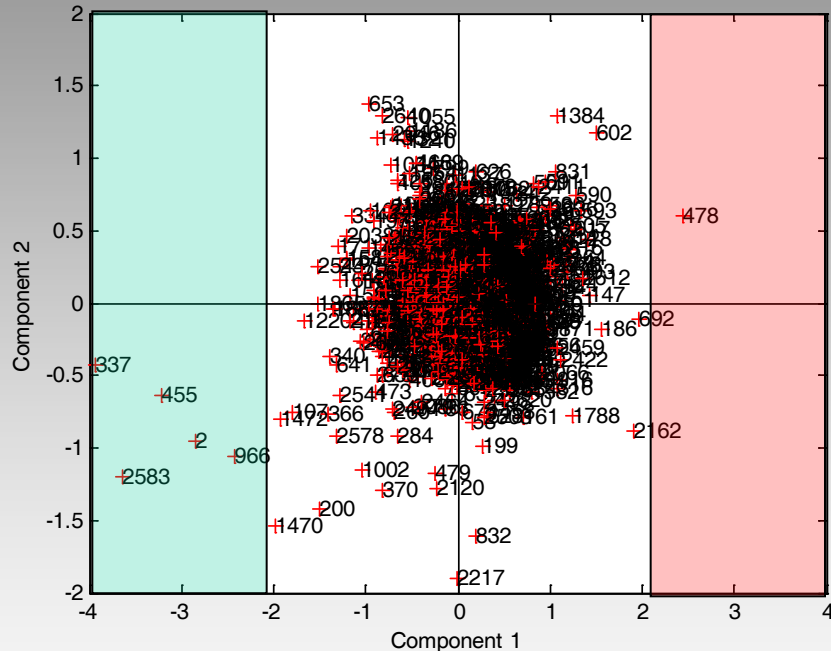
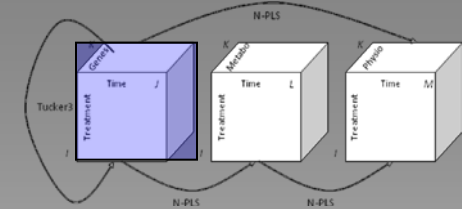
Gene Number	Gene name
2583	Gsta2
337	Ephx1
455	Akr7a3
2	Gsta2
966	Gstm1
1472	Ferritin
1470	Ferritin
2162	Phyh
692	Cyp4a22
478	Spin2b

PARAFAC explained 38.38% of the variation in the data structure. Tucker3 model main core elements explained 98.15% of 42.87% of variation = 42.08%



Describing transcriptomic data

Comparison with Unfold-PCA



The Score Plot provides the same results as the ones obtained from the Tucker3 model

For the loadings, evolutions of the doses with time is not so clear, making difficult the process understanding

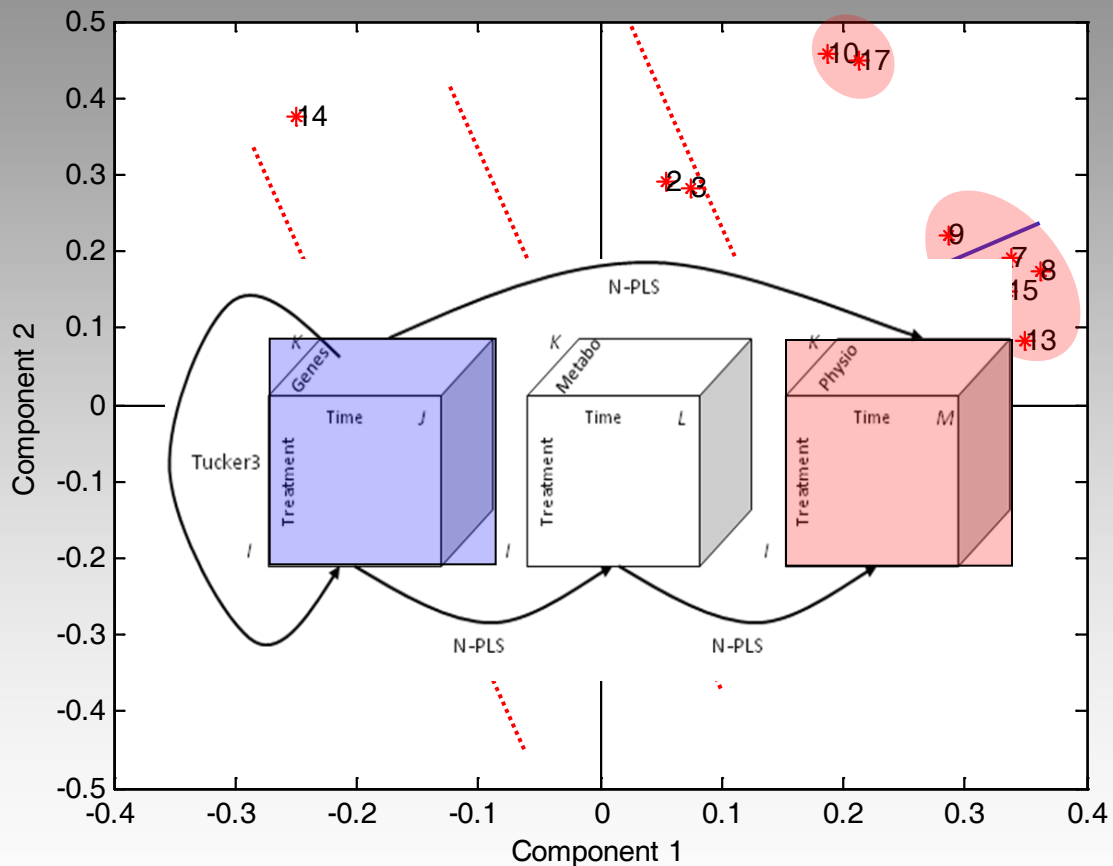


Finding relationships between transcriptomic and physiological data

N-PLS model
2 components
51% of variation

Physiological data

Code	Name
19	A/G Ratio
4	Liver
5	Liver/BW
6	Bilirrubin Tot
11	ALP
7	ASAT
8	ALAT
9	LDH
13	Cholesterol
15	Phospholipids
10	Albumn g/l
17	Tot. Protein (g/l)

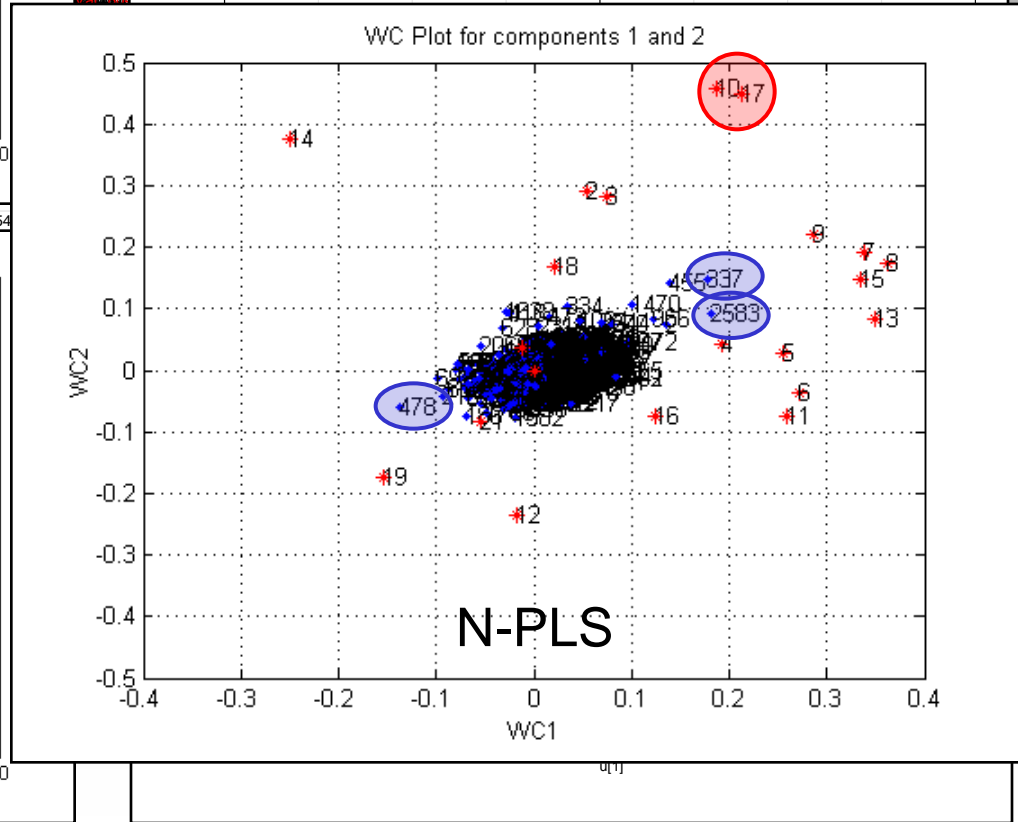
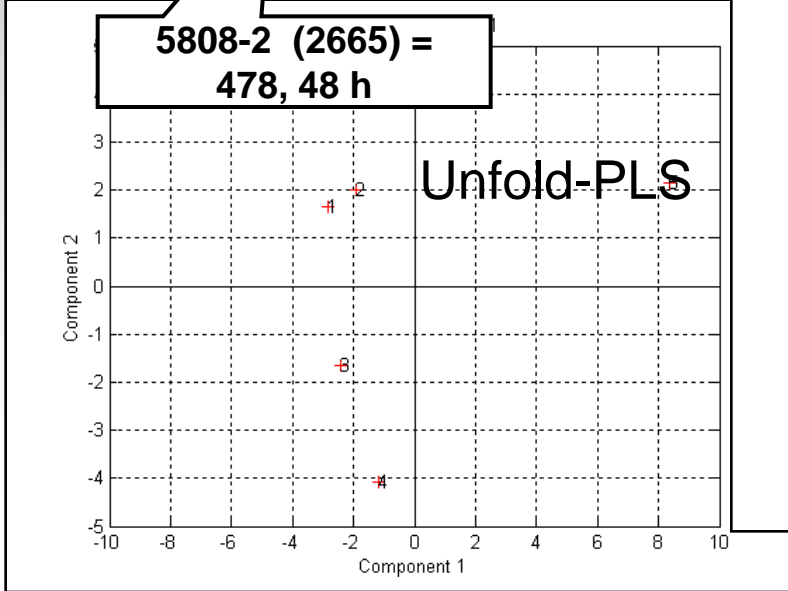
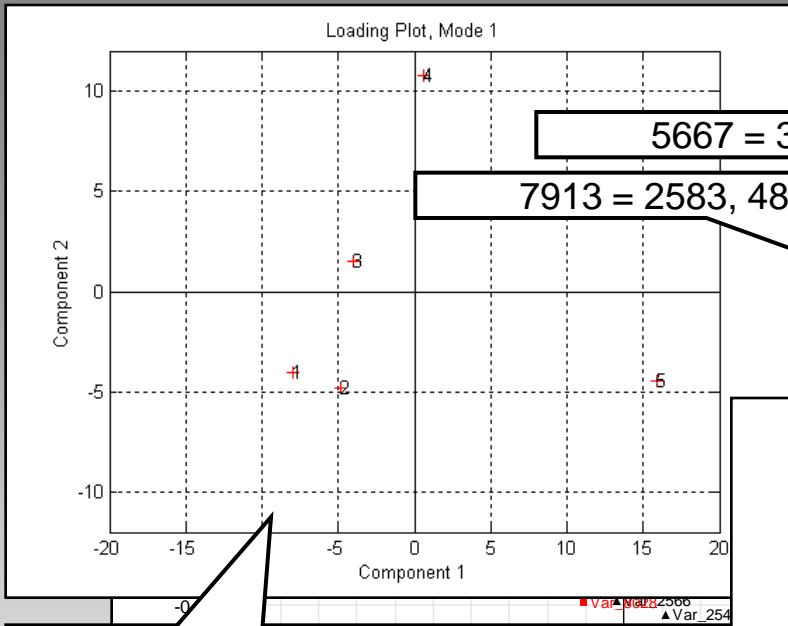


The same genes appearing in Tucker3 model show up as the most relevant



Integration of high dimensional arrays in omics data

N-PLS



Unfold-PLS



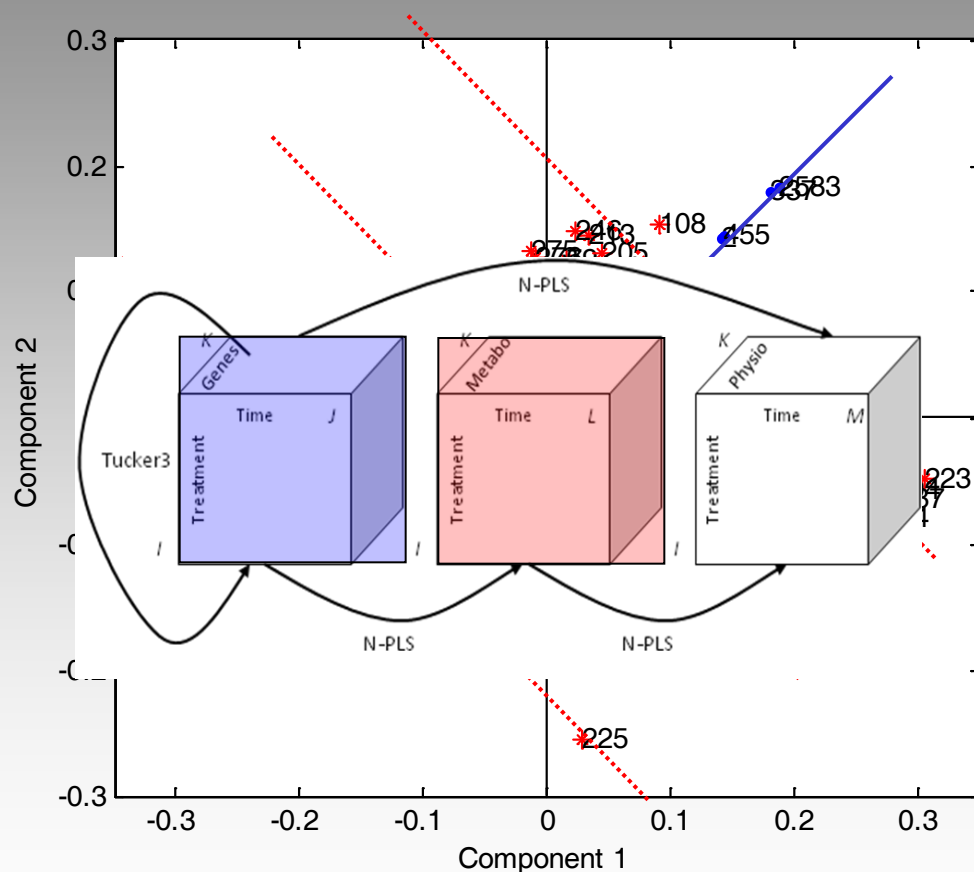
Integration of high dimensional arrays in omics data

Improvement of information provided by metabolomic data

N-PLS model
2 components
55.66 % of variation

Metabolomic data

Code	Name
100	Glutathione
101	Dimethylglycine
102	Glutathione
108	Cysteine



The same genes appearing in Tucker3 model show up as the most relevant

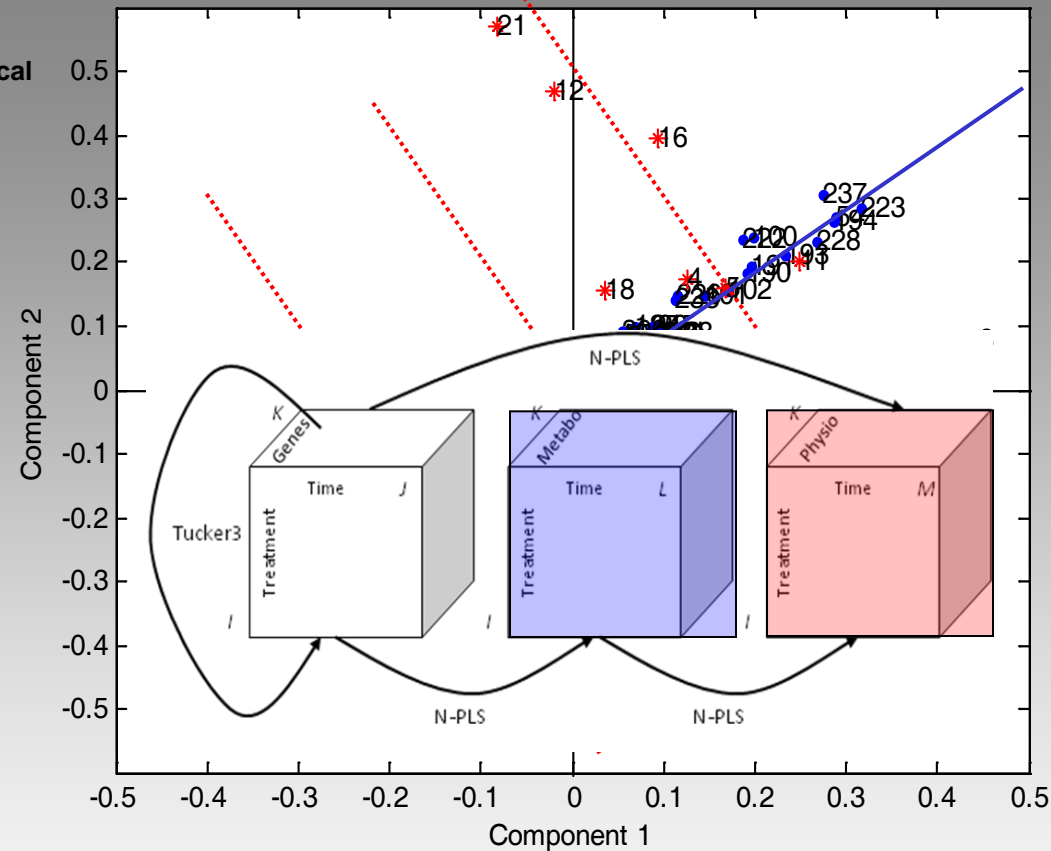


Integration of high dimensional arrays in omics data

Weight (1 st component) <i>M vs Ph</i>	Weight (1 st component) <i>G vs Ph</i>	Difference
-0,25	-0,25	-0,97%
-0,17	-0,15	8,68%
-0,07	-0,05	39,38%
-0,05	-0,01	287,02%
-0,02	-0,02	1,70%
0,00	0,00	
0,03	0,02	38,52%
0,03	0,06	-37,76%
0,09	0,07	15,91%
0,10	0,13	-18,85%
0,17	0,19	-11,73%
0,21	0,19	11,09%
0,23	0,21	10,28%
0,24	0,26	-7,30%
0,24	0,25	-4,05%
0,27	0,27	-1,18%
0,30	0,29	2,84%
0,33	0,34	-1,71%
0,34	0,35	-2,54%
0,35	0,34	2,11%
0,37	0,36	1,44%

Physiological variable

- 14
- 19
- 21
- 01**
- 12
- 20
- 18
- 02
- 03
- 16
- 04
- 10
- 17
- 11
- 05
- 06
- 09
- 15
- 13
- 07
- 08



N-PLS model
2 components
57.26 % of variation



Conclusions

Multiway models can help us to integrate and understand the behaviour of the huge amount of data related to omics structures.

The most important source of covariance with the physiological data is brought from the genes (transcriptomic data), although the metabolites are also adding more information, mainly gathered by the second component.

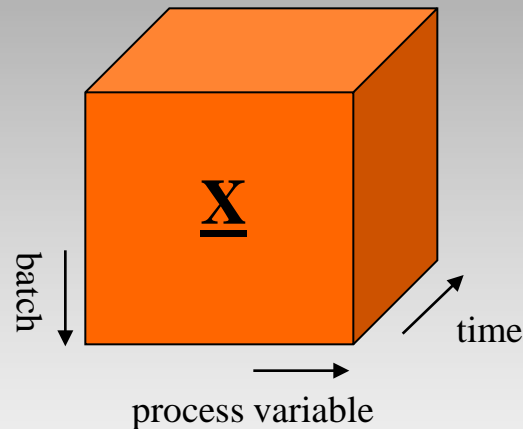
Physiological parameters are slightly better correlated to metabolomics measurements (57%) than to gene expression (51%) suggesting intermediate regulatory mechanisms from the transcriptomics level through the metabolome to the phenotype.

Again, using multiway models (Tucker3 and N-PLS) rather than unfolded ones (U-PCA and U-PLS) yield more interpretable results.



Several goals:

1. Goodness of fit
2. Goodness of prediction
3. Process understanding
4. Fault detection & diagnosis



Several methods:

- Unfold-PCA
- Tucker3
- PARAFAC



- On an “equal component basis” (maintaining the same number of components in the batch direction) and with the same preprocessing, Unfold-PCA fits better than Tucker-3, which in turn fits better than PARAFAC.
- The more complex (flexible), the better fit.
- The question is how complex the model should be.
- Looking for parsimonia: Cross-validated models.

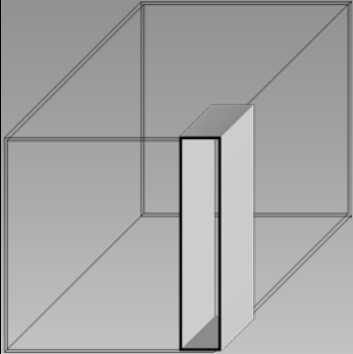
Westerhuis *et al.*, *J. Chemometrics* 13, (1999)

Louwerse & Smilde, *Chemical Engineering Science* 55, (2000)



Batch Process Monitoring: Key issue 2: Centering

- Centering across the batch mode removes the mean trajectories of the process variables (main non-linear and dynamic behaviour).
- Monitoring the deviation from the average trajectories is what matters for batch process monitoring.

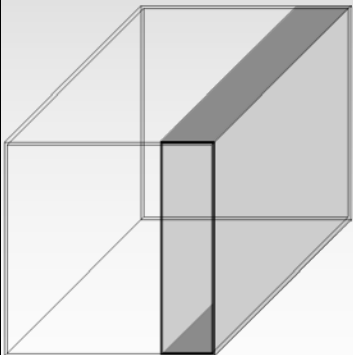


Westerhuis *et al. J. Chemometrics* 13 (1999)

- Slab (variable)-centering removes the grand-mean.

Wold *et al. Chemolab* 44 (1998)

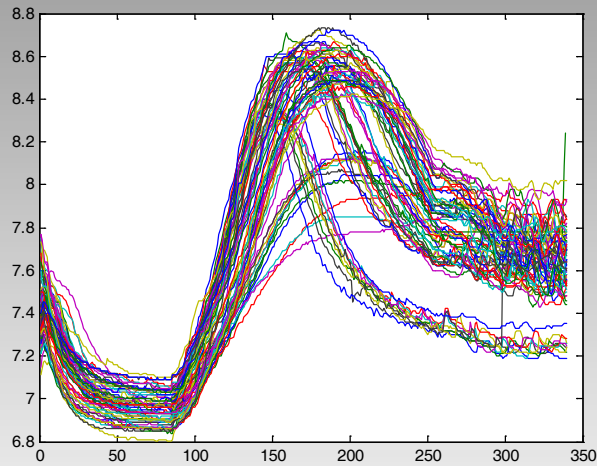
- If mean trajectories are not removed, a large number of components are necessary to describe it before systematic variation from average is modelled.



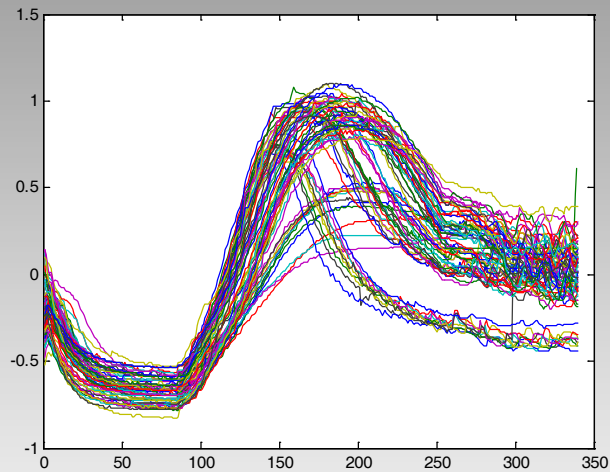
Westerhuis *et al. J. Chemometrics* 13 (1999)



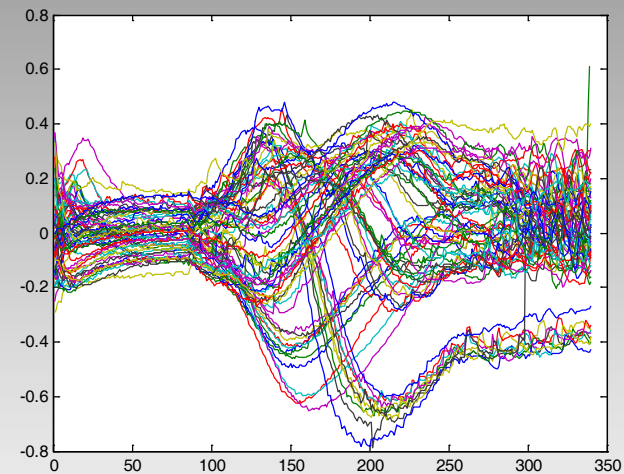
Batch Process Monitoring: Key issue 2: Centering



Original



Removing the
Grand Mean

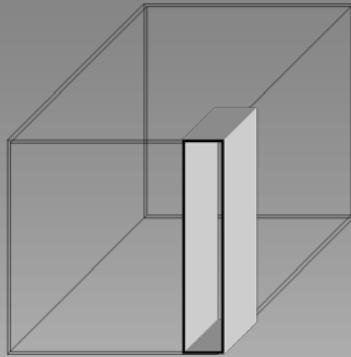


Removing the
mean trajectory

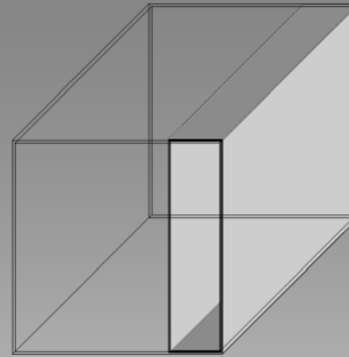


Batch Process Monitoring: Key issue 3: Scaling

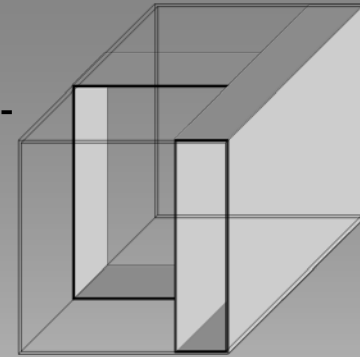
Column



Slab



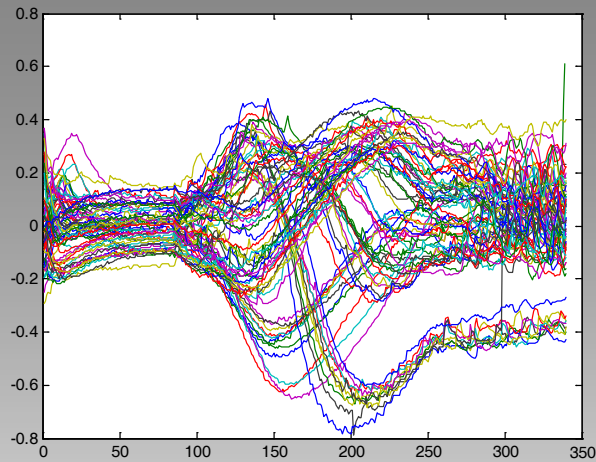
Double-slab



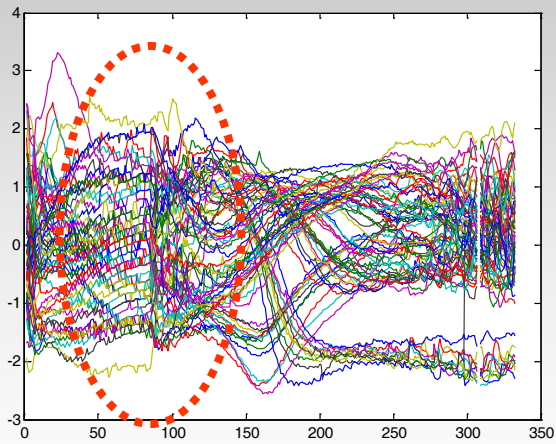
- **Column (auto)-scaling** may increase the magnitude of noise by blowing up those variables with low information content. If multilinear structure does exist, column-scaling may distort it. **Slab (variable)-scaling** and **Double-slab scaling** are a good choice (Gurden *et al. Chemolab* 59, 2001).
- Scaling is problem dependent. It should be chosen such that it improves the detection of faults. With **slab (variable)-scaling** periods with more variability will be weighted more and will have a greater influence on the model. Periods with a lot noise in a specific variable are weighted more than periods with consistent but highly structured variability (Westerhuis *et al. J. Chemometrics* 13, 1999).
- If there is no prior knowledge of the process behaviour and the type of faults **autoscaling** is preferred. If there is prior knowledge, more weight should be done to the critical period of the process (e.g. sampling more frequently) (Kourti & MacGregor, 1999).



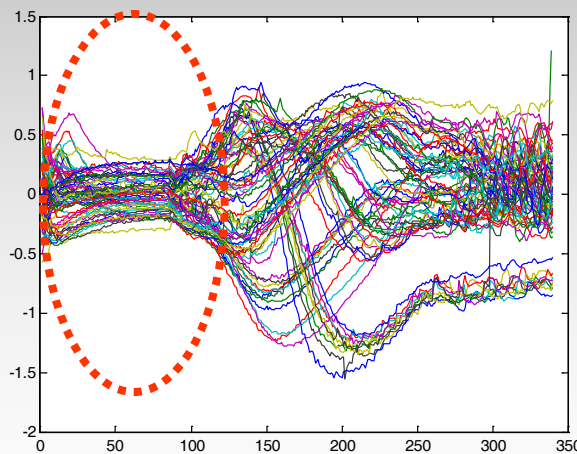
Batch Process Monitoring: Key issue 3: Scaling



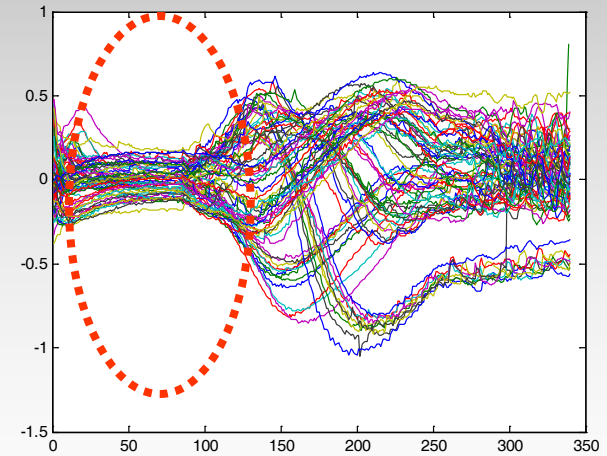
Centred across 1st mode



Column scaling



Slab scaling



Double slab scaling



Batch Process Monitoring: Key issue 4: Interpretation

- In Unfold-PCA loading matrices are very difficult to interpret as they convolute time and variable information.

Wise *et al.* *J. Chemometrics* 13 (1999)

- If there is a strong reason to believe that a multilinear structure does not exist, bilinear models should be used. Otherwise, trilinear models are preferred for their interpretative properties.

Gurden *et al.* *Chemolab* 59 (2001)

- The nature of three-way batch process data is often quite different from the trilinear PARAFAC decomposition

Westerhuis *et al.* *J. Chemometrics* 13 (1999)



Batch Process Monitoring: Key issue 5: Process Analysis & Monitoring

- For overall fault detection, there is no king method. All depends on what type of faults are to be expected.
Louwerse & Smilde Chemical Engineering Science 55 (2000)
- Unfold-PCA is the preferred three-way modelling approach of batch processes
Westerhuis et al. J. Chemometrics 13 (1999)
- The methods are complementary and a well-trained practitioner will find all to be useful
Chiang et al. Chemolab 81 (2006)

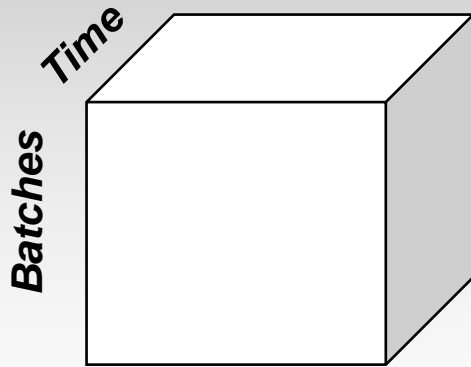


Comparison of four monitoring approaches

Nomikos and MacGregor's approach (BW-Unfold)

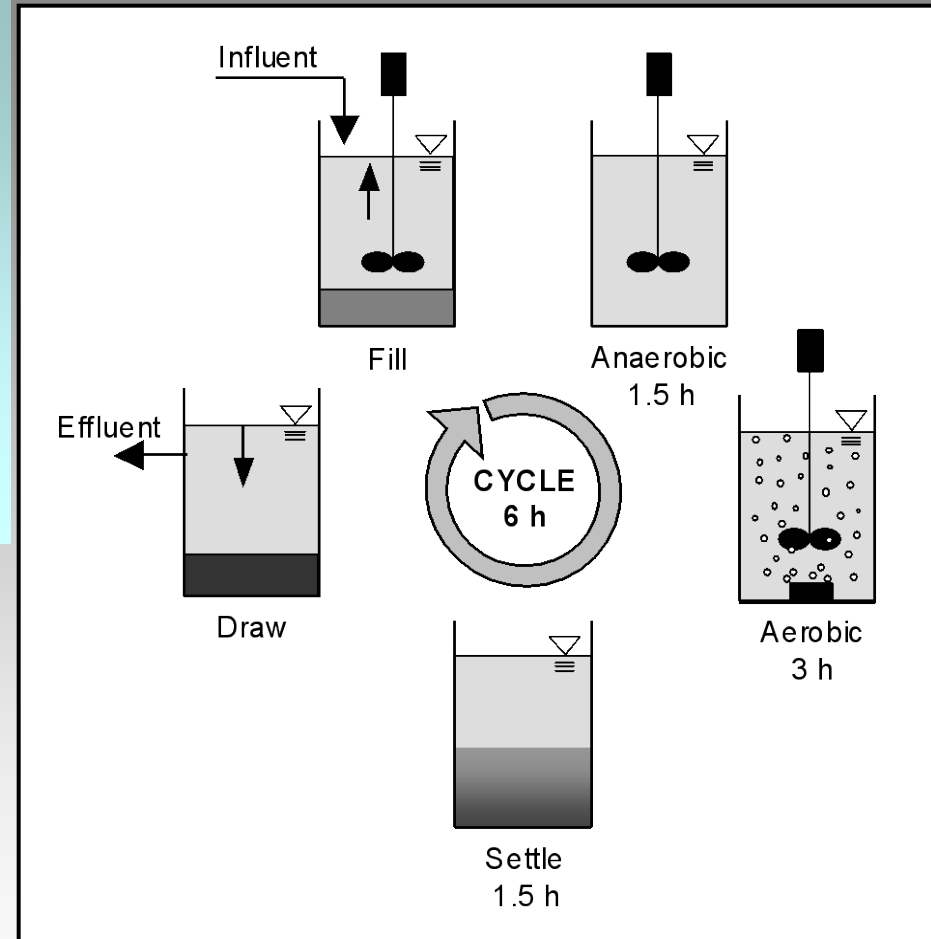
Wold et al's approach and W2 approach (VW-Unfold & BW-Unfold)

Tucker3 model



Variables

Goal: to achieve a good level of enhanced biological phosphorous removal (EBPR) by efficient fault detection and diagnosis



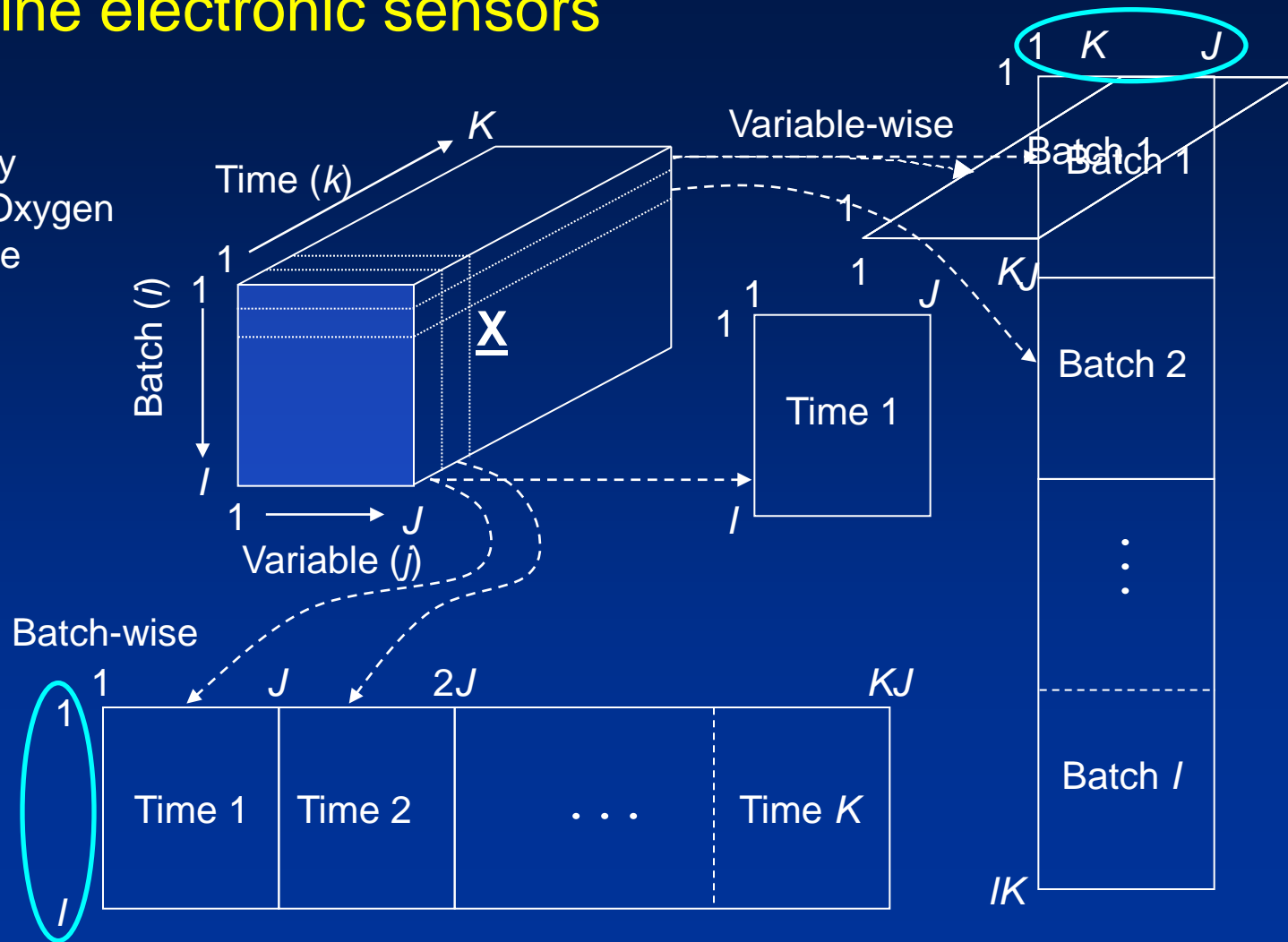
Batch classification

Class	Number of batches	Class description
C1	70	Normal cycles (used for model building).
C2	16	Fault in conductivity measurement.
C3	6	Start-up cycles (system still evolving to stationary state).
C4	6	Noise in the measurements due to analogue filter breakage.
C5	6	Fault in DO measurement due to the presence of air bubble beneath the DO sensor.
C6	2	Aeration system fault in the aerobic stage.
C7	4	Mixing fault at the beginning of the cycle .
C8	35	Normal cycles (used for validation).



- On-line electronic sensors

- $x_1 = \text{pH}$
- $x_2 = \text{ORP}$
- $x_3 = \text{Conductivity}$
- $x_4 = \text{Dissolved Oxygen}$
- $x_5 = \text{Temperature}$

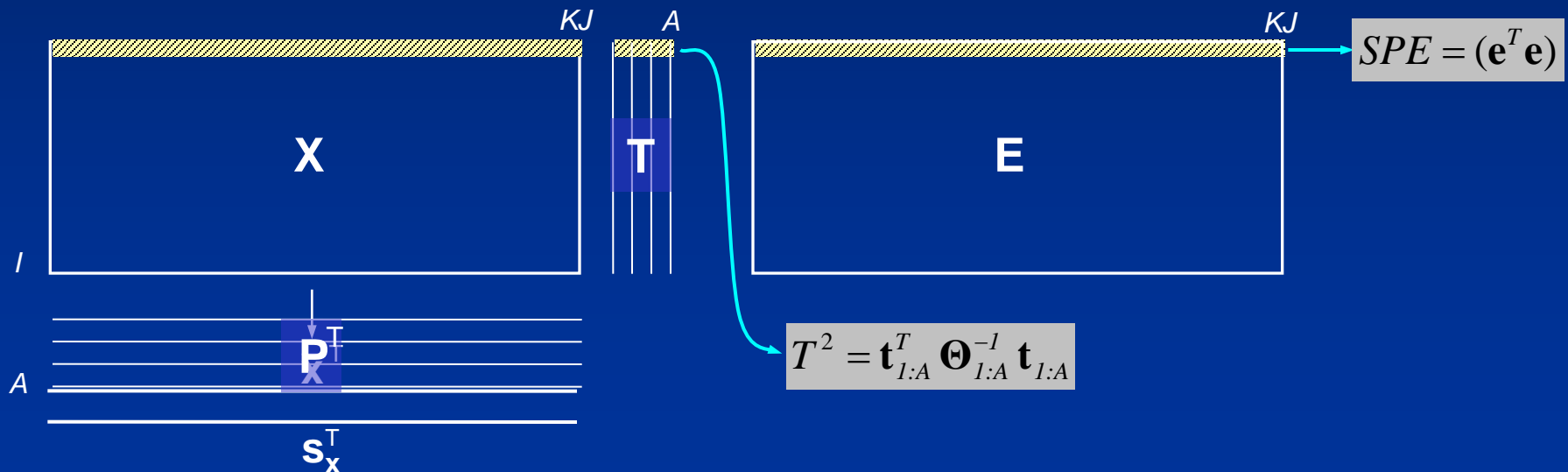


- Nomikos and MacGregor (1995) approach (NM):

Batch-wise unfolding

Pre-processing | centred
| scaled to unit variance

A PCA model is fitted



Main characteristics of the fitted models

Approach	# Comp	R ² (cum)	Q ² (cum)
NM: Nomikos and MacGregor (1995)	7	0,863	0,820
WKFH: Wold <i>et al.</i> (1998)			
Observation level	4	0,837	0,649
Batch level	7	0,960*	0,952*
W2: batch-wise pre-processing followed by variable-wise PCA analysis			
Observation level	4	0.898	---
Batch level	7	0,841*	0.784*
T3: Tucker3, slab scaling	[5,5,5]	0.954	0.952
T3: Tucker3, double slab scaling	[5,5,5]	0.969	0.968
T3: Tucker3, unit variance	[5,5,5]	0.789	0.780

Note: --- components were forced

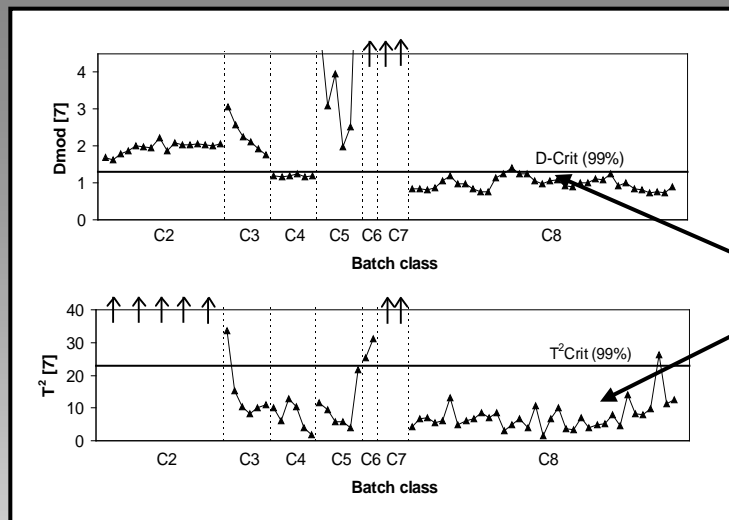
* explained variance refer to the scores from the OL not to the process data

Tucker3 by crossvalidation

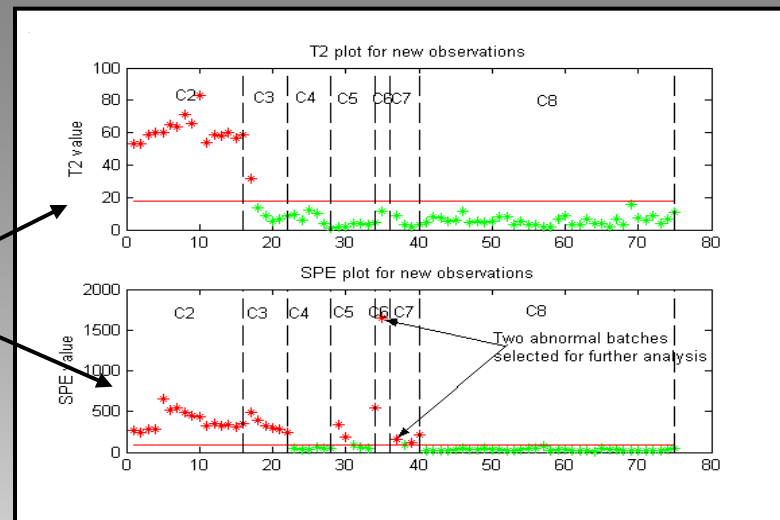


Off-line fault detection, diagnosis and monitoring system in a SBR

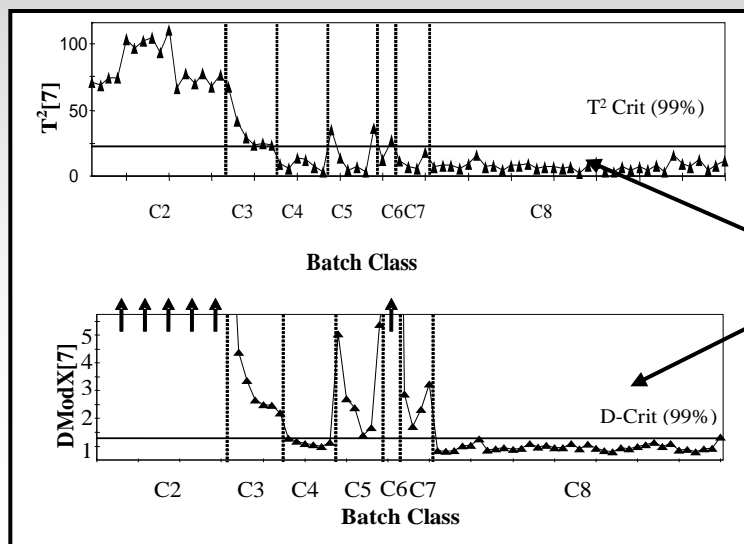
Nomikos and MacGregor's approach



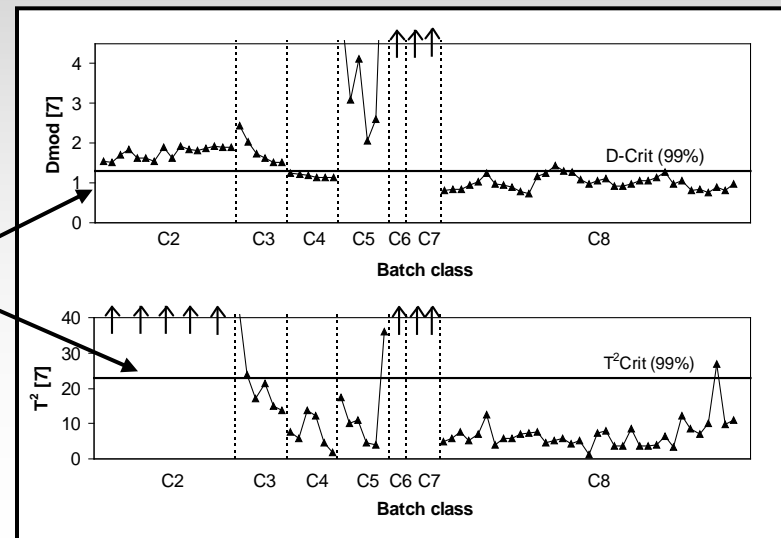
Tucker3 model



WKFH approach



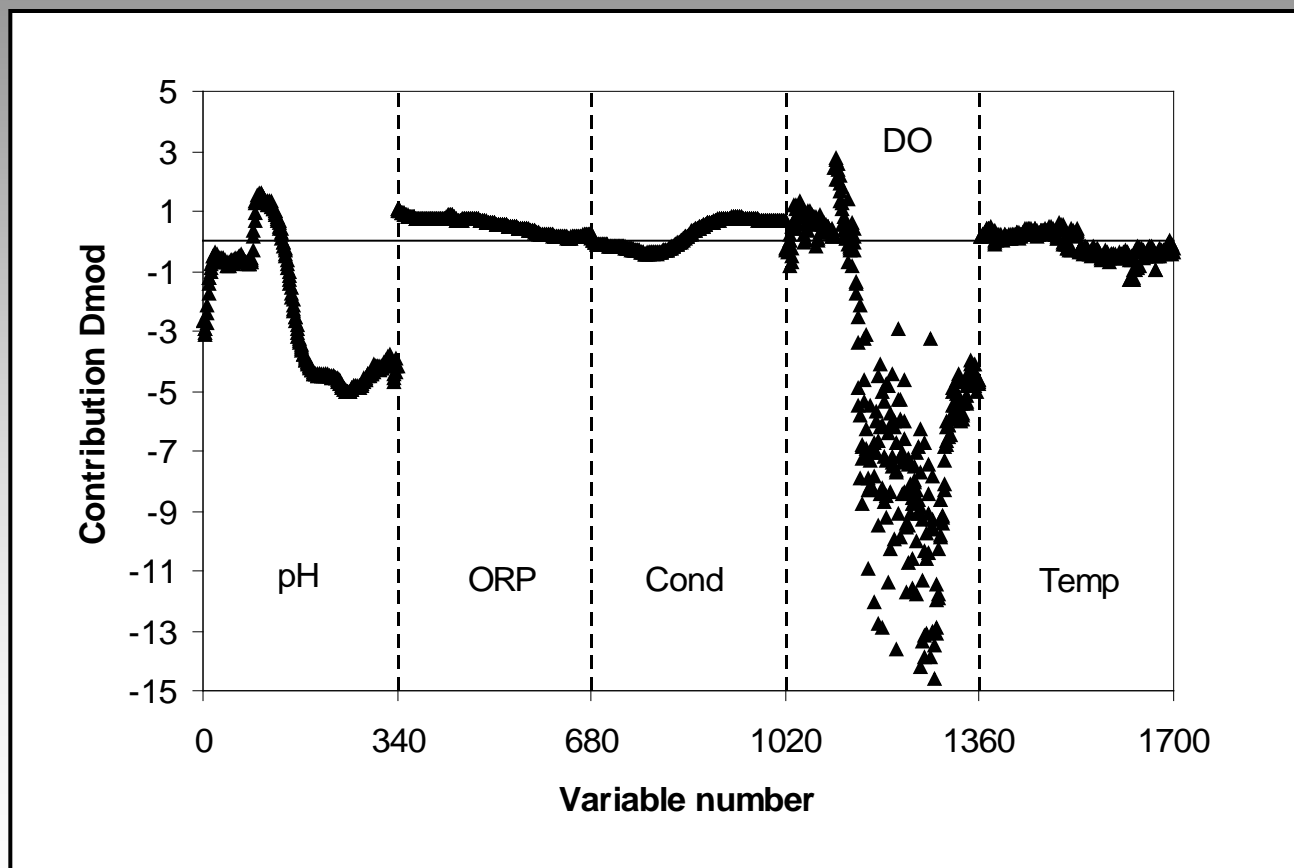
W2 approach



Off-line fault detection, diagnosis and monitoring system in a SBR

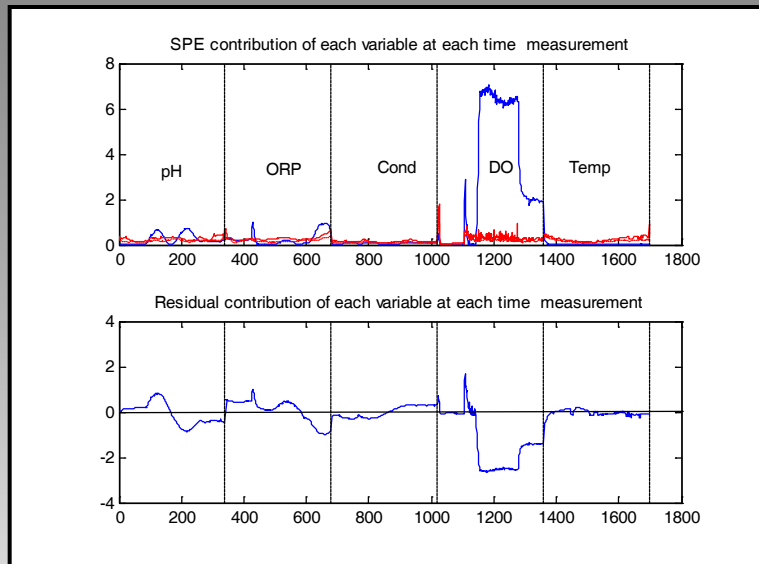
Contribution plot, batch 1, class C6 (aeration system fault in aerobic stage)

Nomikos & MacGregor's approach

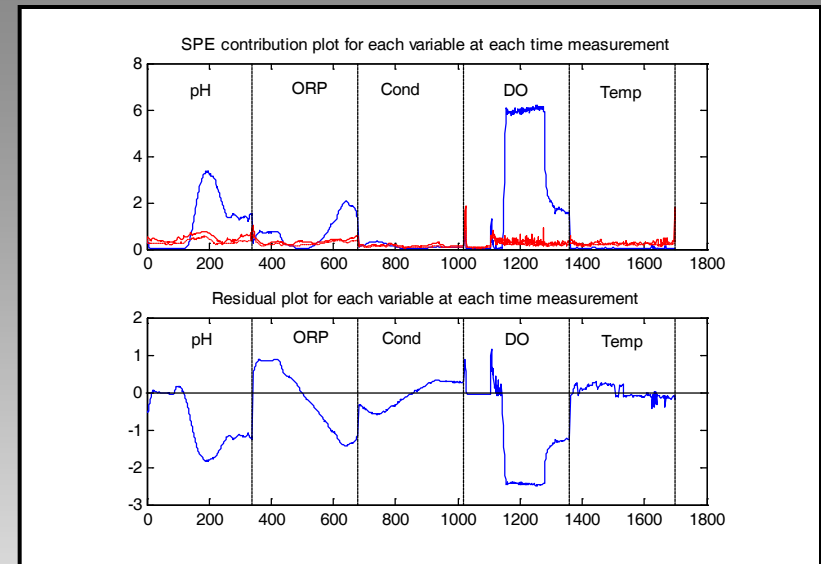


Off-line fault detection, diagnosis and monitoring system in a SBR

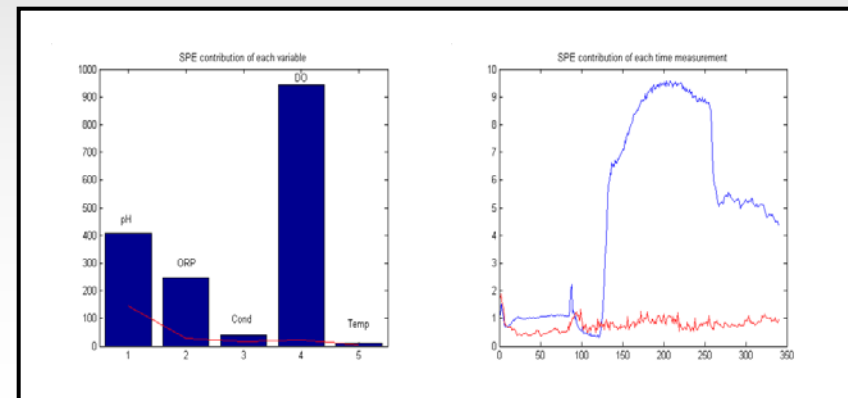
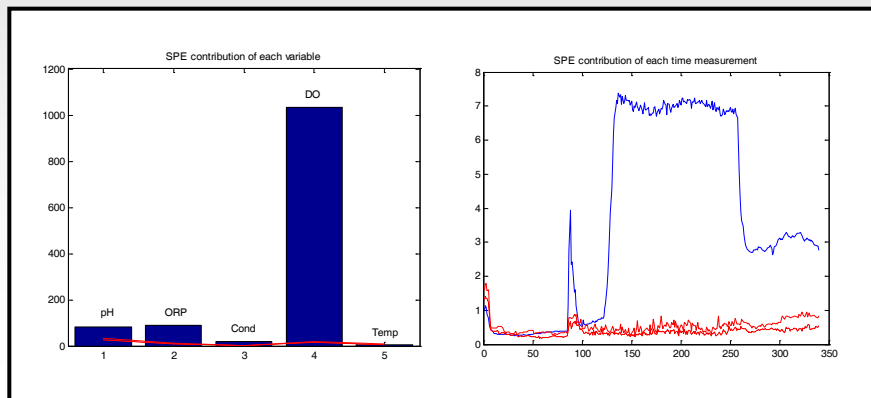
Contribution plot, batch 1, class C6 (aeration system fault in aerobic stage) (Tucker3)



Centring across batches mode
Scaling within variables and time modes



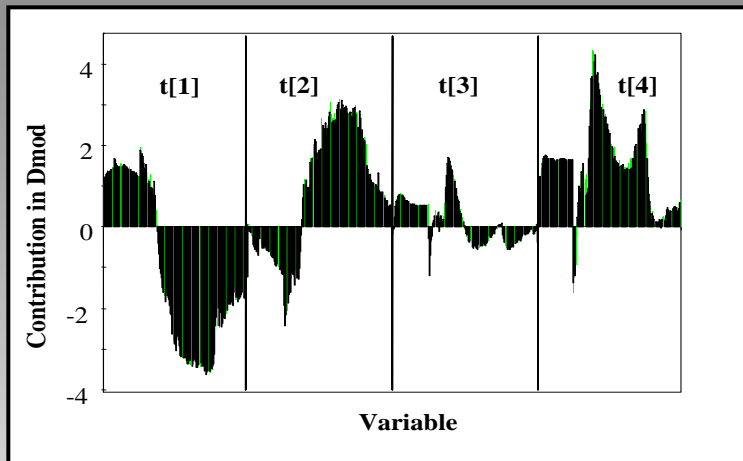
Centring across batches mode
Scaling within variables mode



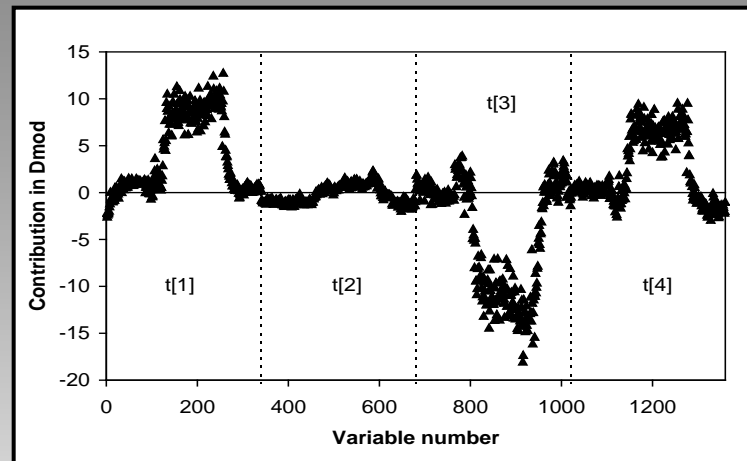
Off-line fault detection, diagnosis and monitoring system in a SBR

Contribution plot, batch 1, class C6 (aeration system fault in aerobic stage)

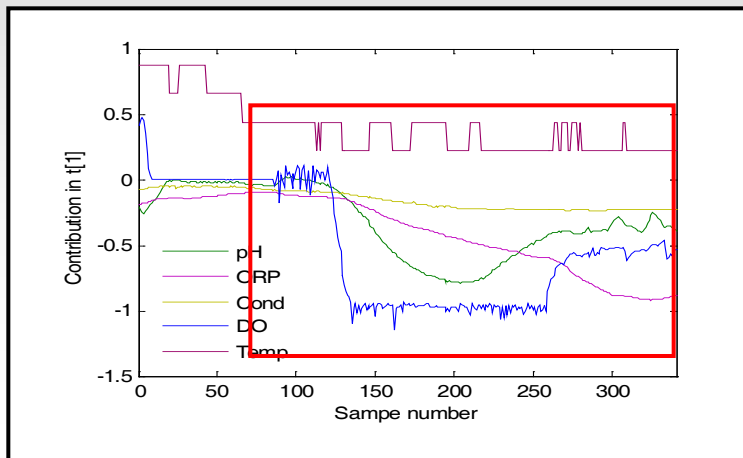
WKFH approach (batch level)



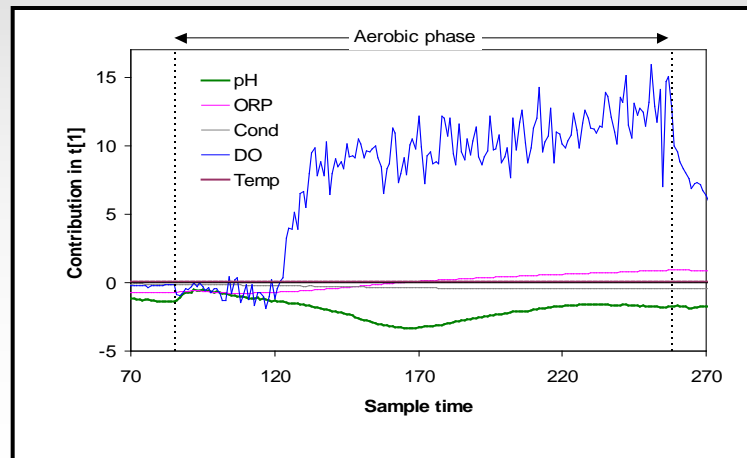
W2 approach (batch level)



Contribution plot of t[1] from the observation level, WKFH approach



Contribution plot of t[1] from the observation level, W2 approach



Comparative performance

Comparison between the developed models based on the projection of the finished batches from classes C2 to C8. Results obtained with 99% control limits.

	NM approach	WKFH approach (batch level)	W2 approach (batch level)	Tucker3 model
C1	0%	0%	0%	0%
C2	100%, T ² , Dmod, Cond	100%, T ² , Dmod, Cond	100%, T ² , Dmod, Cond	100%, T ² , SPE, pH, ORP, Cond
C3	100%, Dmod, pH	100%, T ² , Dmod, Cond, Temp	100%, Dmod, pH	100%, T ² , SPE, pH, ORP
C4	0%	0%	0%	0%
C5	100%, Dmod, DO	100%, T ² , Dmod, DO	100%, Dmod, DO	50%, SPE, DO, pH
C6	100%, T ² , Dmod, DO	100%, T ² , Dmod, DO	100%, T ² , Dmod, DO	100%, SPE, DO, pH, ORP
C7	100%, T ² , Dmod, DO	100%, Dmod, DO	100%, T ² , Dmod, DO	75%, SPE, DO
C8	5.7% (2/35), Dmod, T ²	0%(0/35)	5.7%(2/35), Dmod, T ²	0%(0/35)



Conclusions

- CV Tucker3 fit slightly better than bilinear models
- Tucker3 shows slightly lower detection ability in some faults
- Tucker3 and WKFH fail to correctly diagnose the responsible variables for some faults
- NM (Batch-wise unfolding) and Tucker3 provide more simple procedures for fault diagnosis
- The most consistent fault diagnosis was achieved by NM

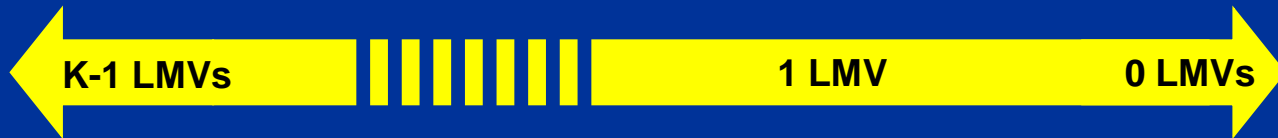
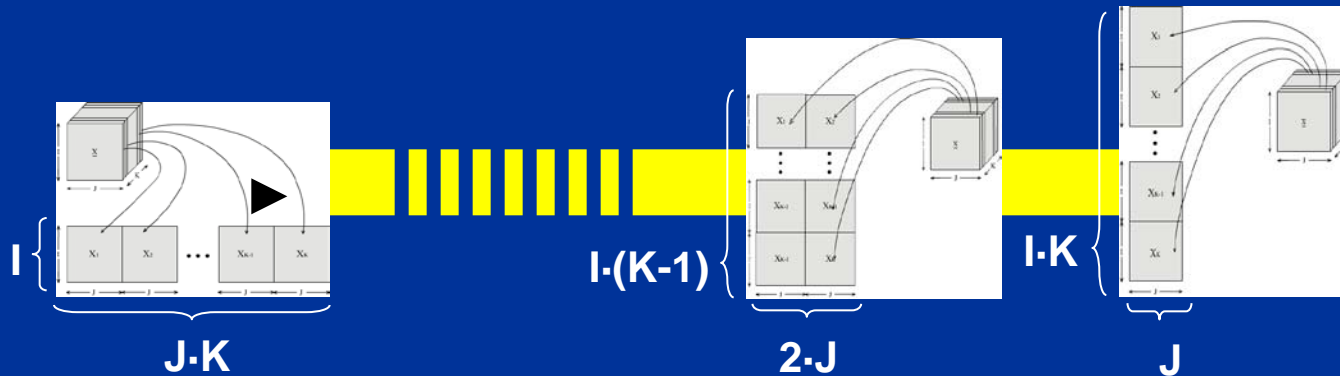


Final remarks

- *Multiway data is usually well modelled by multiway methods*
- *Preprocessing is a critical step*
- *Usually multiway methods yield easily interpretable models*
- *For batch process multivariate analysis the advantages of MW over Unfold methods is not so clear, due to complexity of changing dynamics over time*



One question: unfolding implications on Tucker 3 model



+ LMVs

- Parsimonious

+ Dynamics captured

- Constrained Dynamics

+ Auto-corr Stats.

- LMVs

+ Parsimonious

- Dynamics captured

+ Constrained Dynamics

- Auto-corr. Stats.



J. Camacho, J. Picó and A. Ferrer. *Bilinear modelling of batch processes. Part I: Theoretical discussion*, Journal of Chemometrics 22 (5), 299-308 (2008).





UNIVERSIDAD
POLITECNICA
DE VALENCIA

Multiway analysis: A practitioner's perspective



***José Manuel Prats-Montalbán
&
Alberto Ferrer***

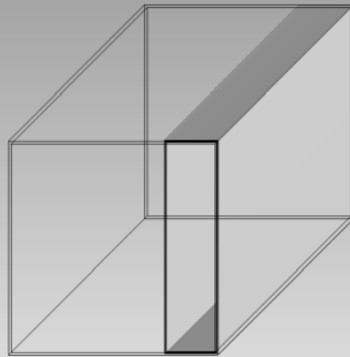
Multivariate Statistical Engineering Group
<http://mseg.webs.upv.es/>

Dp. of Applied Statistics, Operations Research & Quality

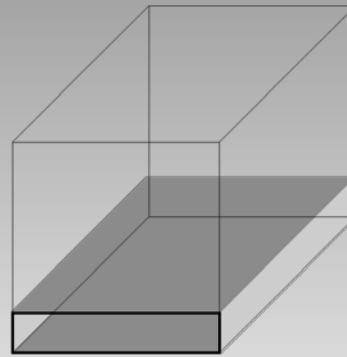
Universidad Politécnica de Valencia, Spain



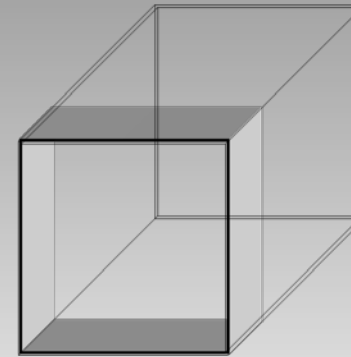
Three-way batch data



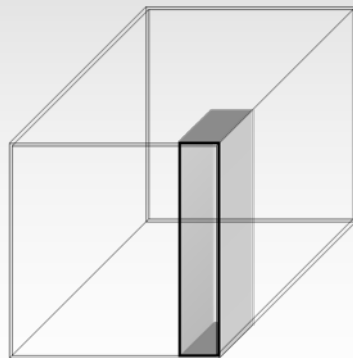
Variable slab



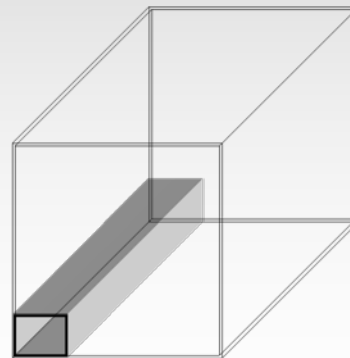
Batch slab



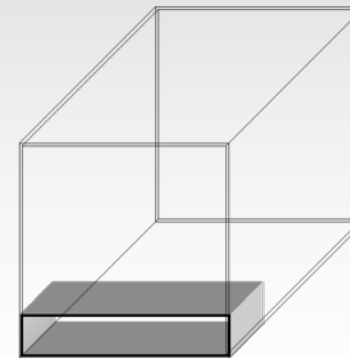
Time slab



Batch fiber



Time fiber



Variable fiber



Scaling and centering of batch process data

- Centering

- Column centering: $\mathbf{X}_{ijk}^c = \mathbf{X}_{ijk} - \bar{\mathbf{X}}_{jk}$

- Scaling

- Column scaling: $\mathbf{X}_{ijk}^s = \frac{\mathbf{X}_{ijk}}{\|\mathbf{X}_{jk}\|^2}$

- Slab scaling: $\mathbf{X}_{ijk}^s = \frac{\mathbf{X}_{ijk}}{\|\mathbf{X}_j\|^2}$



$$\underline{\mathbf{X}} = a_1 \underline{\mathbf{P}}_1 + a_2 \underline{\mathbf{P}}_2 + \underline{\mathbf{E}}$$

$$\underline{\mathbf{X}} = \begin{matrix} \underline{\mathbf{G}} \\ \underline{\mathbf{A}} \end{matrix} \begin{matrix} \underline{\mathbf{C}} \\ \underline{\mathbf{B}} \end{matrix} + \underline{\mathbf{E}}$$

r_3 (above $\underline{\mathbf{C}}$)
 r_2 (below $\underline{\mathbf{B}}$)

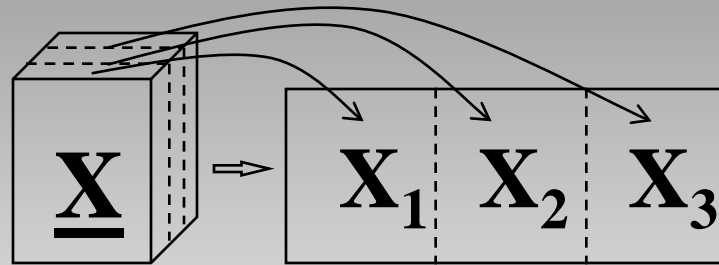
$$\underline{\mathbf{X}} = \begin{matrix} \underline{\mathbf{I}} \\ \underline{\mathbf{A}} \end{matrix} \begin{matrix} \underline{\mathbf{C}} \\ \underline{\mathbf{B}} \end{matrix} + \underline{\mathbf{E}}$$

r_1 (above $\underline{\mathbf{C}}$)



Multiway Component model

Unfold (matricized) PCA

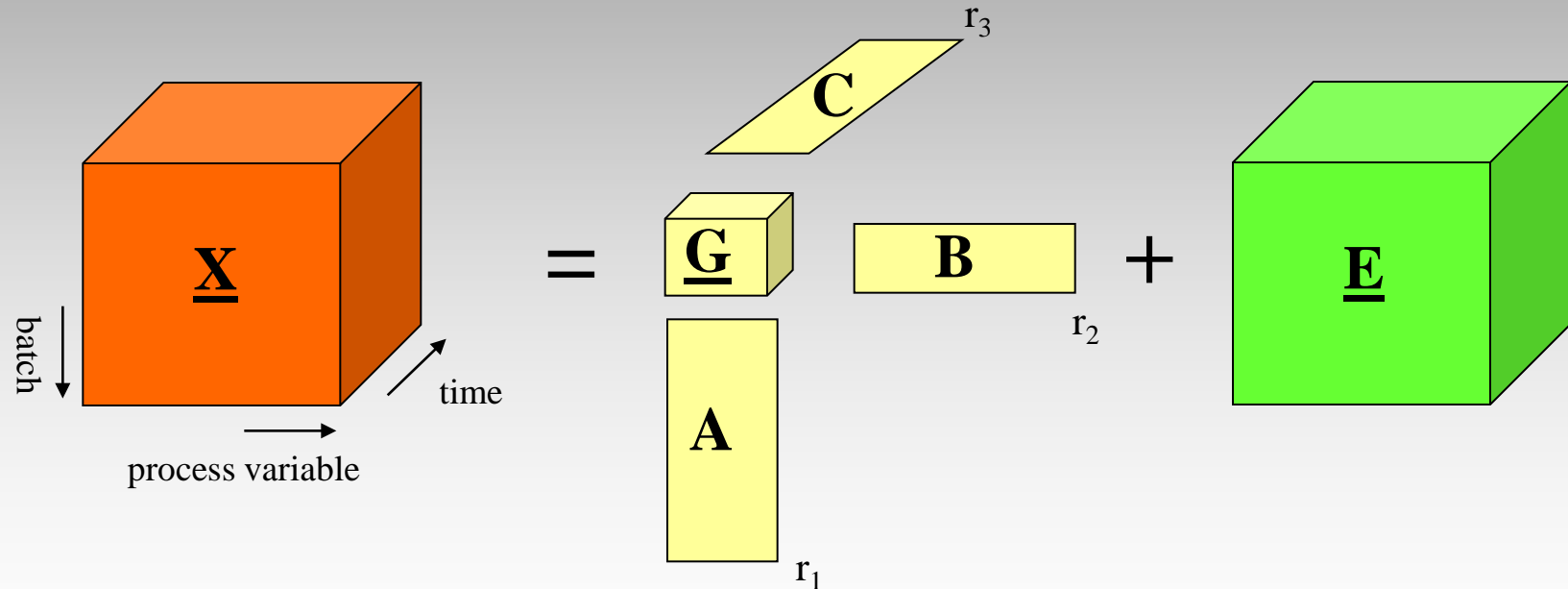


$$\begin{array}{|c|c|c|} \hline \mathbf{X}_1 & \mathbf{X}_2 & \mathbf{X}_3 \\ \hline \end{array} = \begin{array}{|c|} \hline \mathbf{p}_1 \\ \hline \mathbf{a}_1 \\ \hline \end{array} + \begin{array}{|c|} \hline \mathbf{p}_2 \\ \hline \mathbf{a}_2 \\ \hline \end{array} + \begin{array}{|c|c|c|} \hline \mathbf{E}_1 & \mathbf{E}_2 & \mathbf{E}_3 \\ \hline \end{array}$$



Multiway Component model

Tucker3

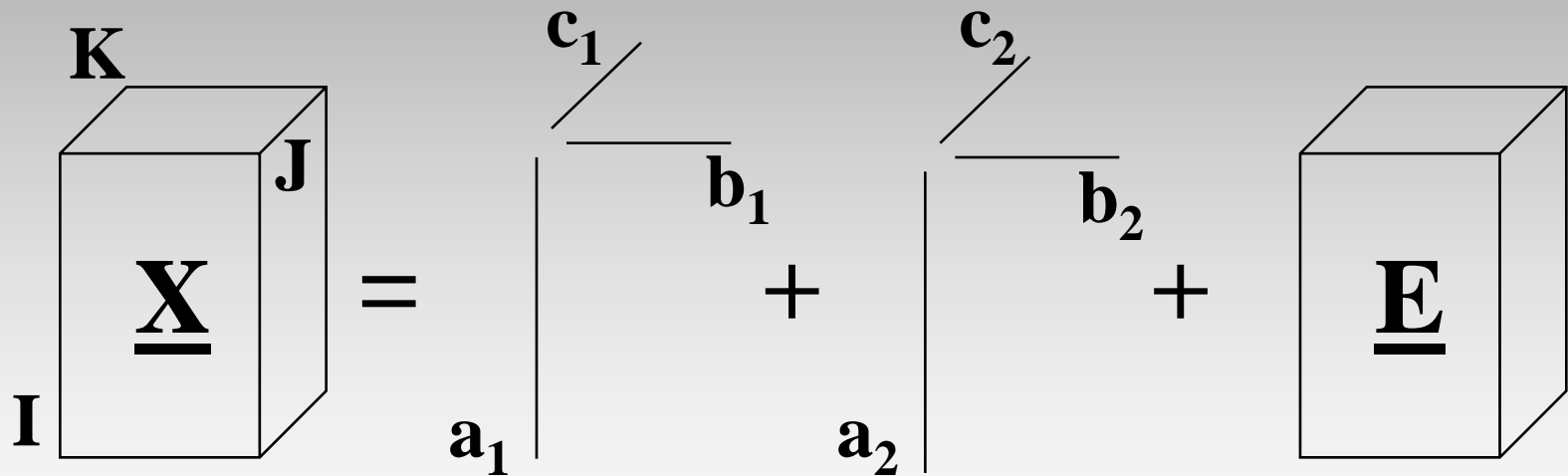


$\underline{\mathbf{G}}$ ($r_1 \times r_2 \times r_3$) general core-matrix

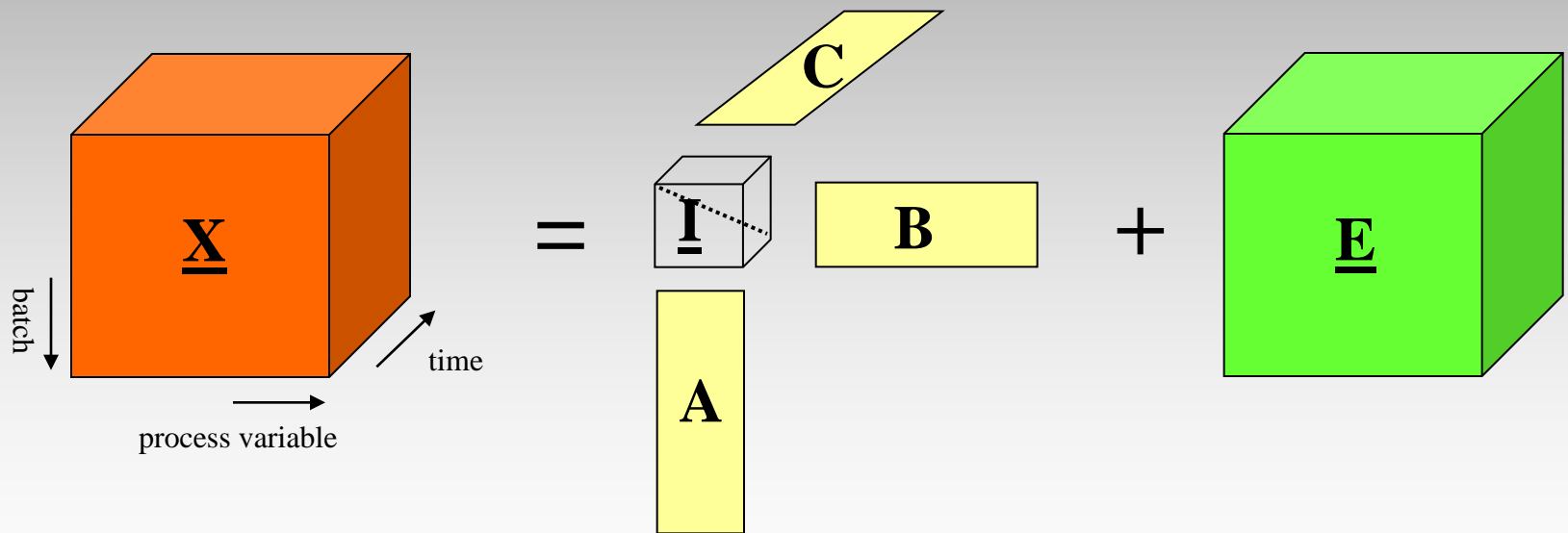


Multiway Component model

PARAFAC



Multiway Component model PARAFAC



\underline{I} (rxrxr) has superdiagonal and all other elements are zero.



Multiway component models

- Unfold PCA
 - Easy to calculate, largest number of parameters,
 - Interpretation of loadings is difficult
- Tucker3
 - Different number of components in each direction
 - Flexible, better interpretation
- PARAFAC
 - Parsimonious, mainly used for spectroscopic data
 - Scores may not be orthogonal

