Bilinear and trilinear data structures and models for the resolution and interpretation of complex chemical data systems

Romà Tauler[1] and Anna de Juan[2]

[1]IIQAB-CID-CSIC

[2]University of Barcelona

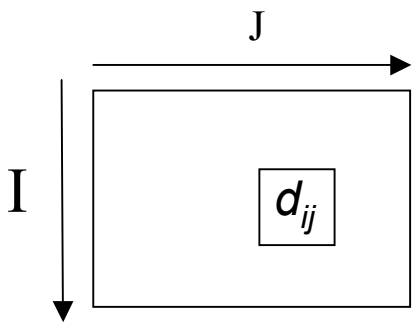
Outline:

- Introduction and motivations of this work
- Models and structures for complex (three-way) chemical data
- Results of comparison of models and methods to analyze complex (three-way) chemical data
- Conclusions

Introduction and motivations of this work

Basically, two type of models have been proposed for the interpretation and resolution of complex chemical data systems:

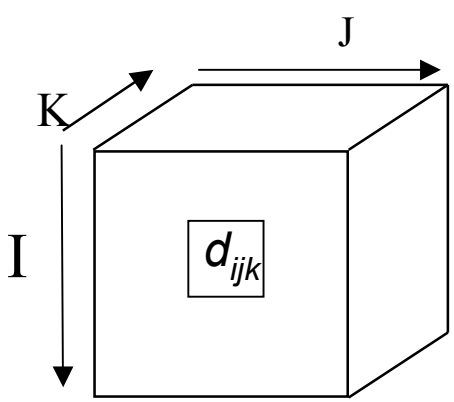
- Bilinear models for two-way and three-way data
- Trilinear models for three-way data



Two-way
Bilinear Model

$$d_{ij} = \sum_{n=1}^{N} c_{in} s_{nj} + e_{ij}$$

$$\mathbf{D} = \mathbf{C}\mathbf{S}^{\mathrm{T}}$$



Three-way
Trilinear model

$$d_{ijk} = \sum_{n=1}^{N} c_{in} s_{nj} z_{nk} + e_{ijk}$$

$$\mathbf{D}_r^{\mathrm{T}} = [\mathbf{D}_1^{\mathrm{T}}, \mathbf{D}_2^{\mathrm{T}}, \ldots, \mathbf{D}_K^{\mathrm{T}}] = \mathbf{S}\mathbf{I}_r^{\mathrm{T}} (\mathbf{Z}^{\mathrm{T}} \otimes \mathbf{C}^{\mathrm{T}})$$

Introduction and motivations of this work

• In this presentation, we will review this situation in a chemical context, considering results obtained by application of different three-way methods based or not on trilinear models.

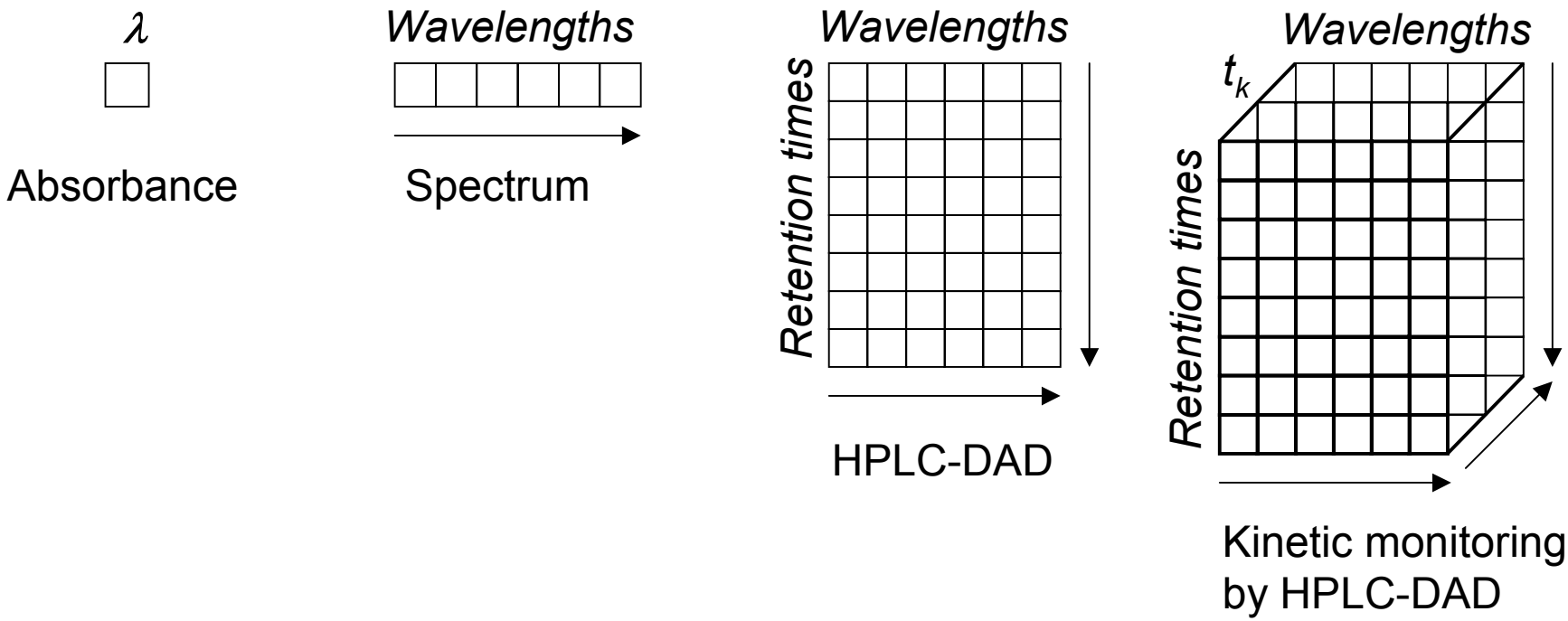
• All this has been analyzed in a previos paper: 'Comparison of three-way resolution methods for non-trilinear chemical data sets'. A. de Juan and R. Tauler. **J.of Chemometrics, 2001, 15, 749-772**

Outline:

- Introduction and motivations of this work
- Models and structures for complex (three-way) chemical measurements data
- Results of comparison of models and methods to analyze complex (three-way) chemical measurements data
- Conclusions

Chemical measurements

0-way data sets **1-way data sets** **2-way data sets** **3-way data sets**



Information quality

(-) ⟶ (+)

Data analysis complexity

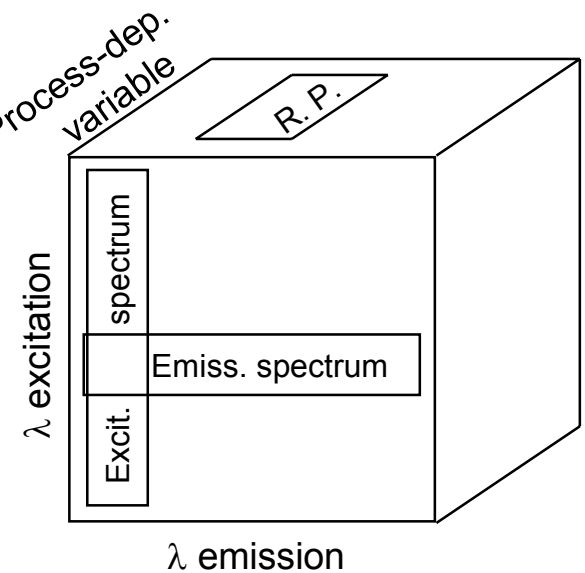
Chemical measurements
Three-way data sets

Two modulated spectral modes
- *excitation mode*
- *emission mode*

One chemical mode
sample, reaction/process evolution, pH, time, temperature,.

This is the 'archetypical' three-way data set fulfilling a trilinear model

3-way data sets



Is this always true?

Baseline problems, instrumental reproducibility, scattering, missing, outliers…

Chemical measurements
Three-way data sets

Two chemical modes

- chromatographic (elution), kinetic, equilibrium, temperature..., mode

- sample, run, reaction/process number

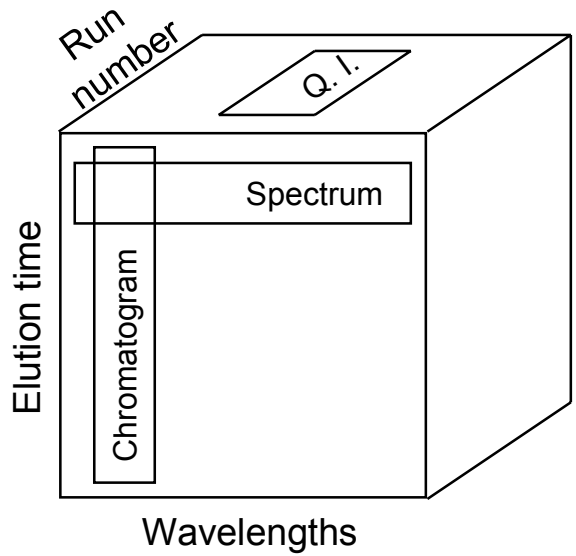
One spectral mode

UV-VIS, NIR, FT-IR, NMR, CD, spectra ...

These are the more common three-way data sets in Chemistry!!!

Do these data fulfill a trilinear model?

3-way data sets



Models to describe chemical measurements

Models for what?

Models for:

1. exploratory data analysis?
2. data interpretation?
3. **data resolution?**
4. .....

Models for data resolution ➔ resolution of the 'true' underlying 'physical/chemical' sources of data variation

- hard-modeling (physico-chemical model)
- **soft-modeling** (no physico-chemical model, soft constraints)

Chemometric soft-models to describe chemical measurements

One way data ➔ Linear and non-linear models

Two way data ➔ **Bilinear** and non-bilinear models
Non-bilinear data can still be linear in one of the two modes

Three-way data ➔ **Trilinear and non-trilinear models**
Non-trilinear data can still be linear in two of the modes (bilinear) ➔
This is the more common situation in Chemistry!!!

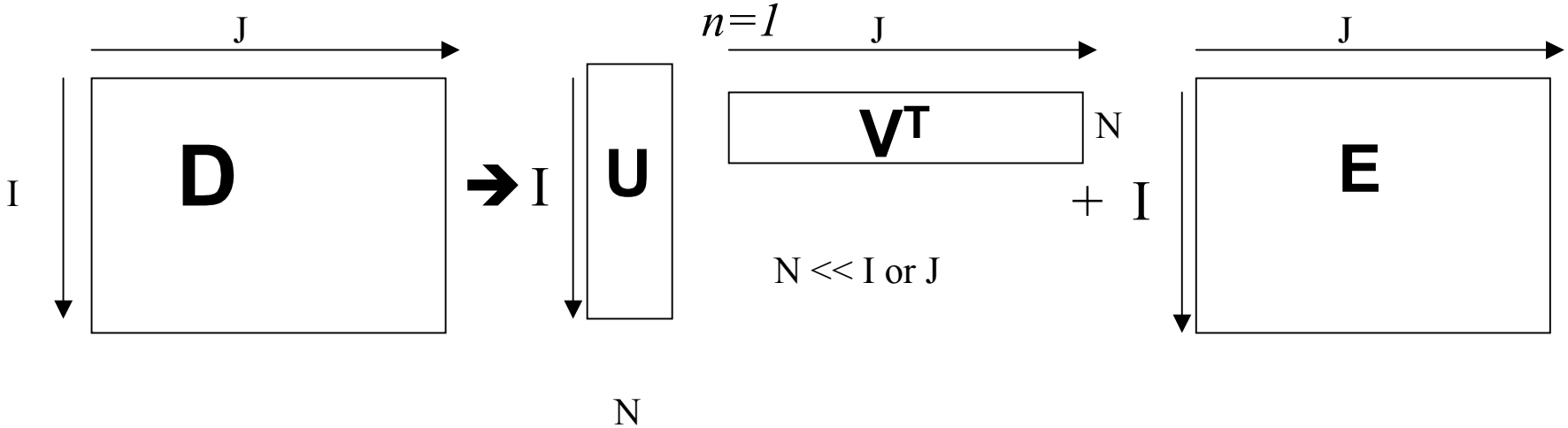
Factor Analysis/Principal Component Analysis
Bilinear Model

$$\mathbf{D} = \mathbf{U}\,\mathbf{V}^{\mathbf{T}} + \mathbf{E}$$

Unique solutions but without physical meaning

Constraints: **U** orthogonal, $\mathbf{V}^{\mathbf{T}}$ orthonormal
$\mathbf{V}^{\mathbf{T}}$ in the direction of maximum variance

$$d_{ij} = \sum_{n=1}^{N} u_{in} v_{nj} + e_{ij}$$



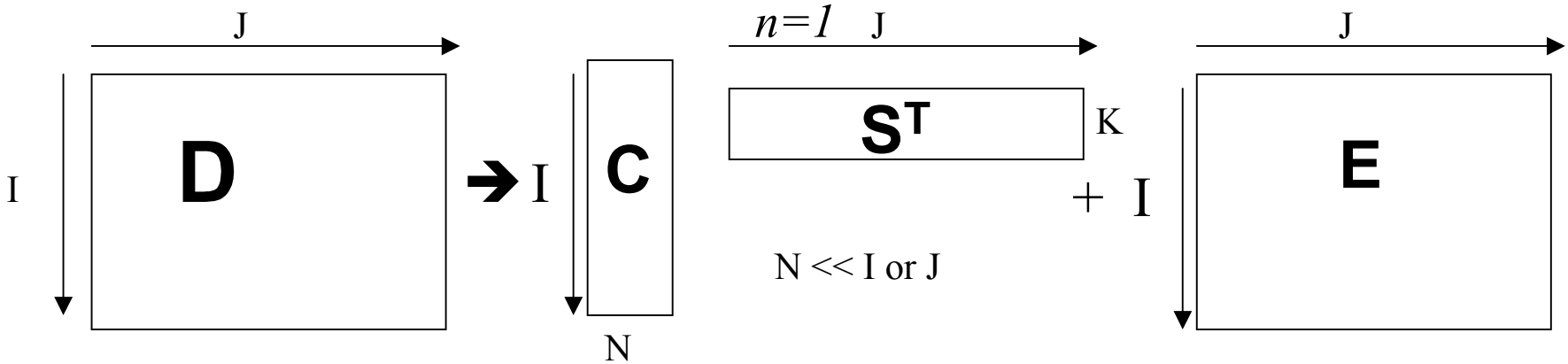
Multivariate Curve Resolution Bilinear Model

$$\mathbf{D} = \mathbf{C}\,\mathbf{S}^{\mathbf{T}} + \mathbf{E}$$

Non-unique solutions but with physical meaning (rotational/ intensity ambiguities are present)

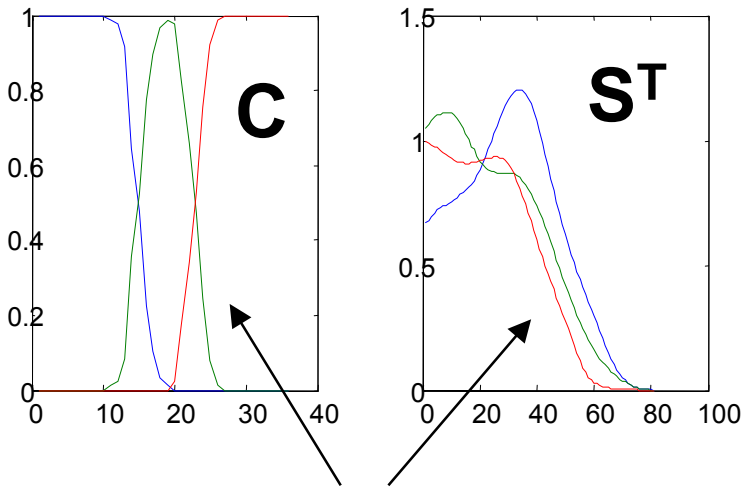
Constraints: **C** and $\mathbf{S}^{\mathbf{T}}$ non-negative
C or $\mathbf{S}^{\mathbf{T}}$ scaled (normalization, closure)
Other constraints (unimodality, local rank, selectivity...)

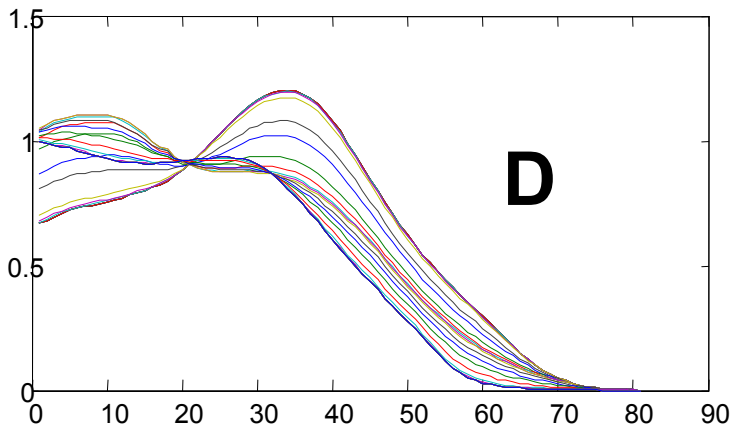
$$d_{ij} = \sum_{n=1}^{N} c_{ik} s_{kj} + e_{ij}$$

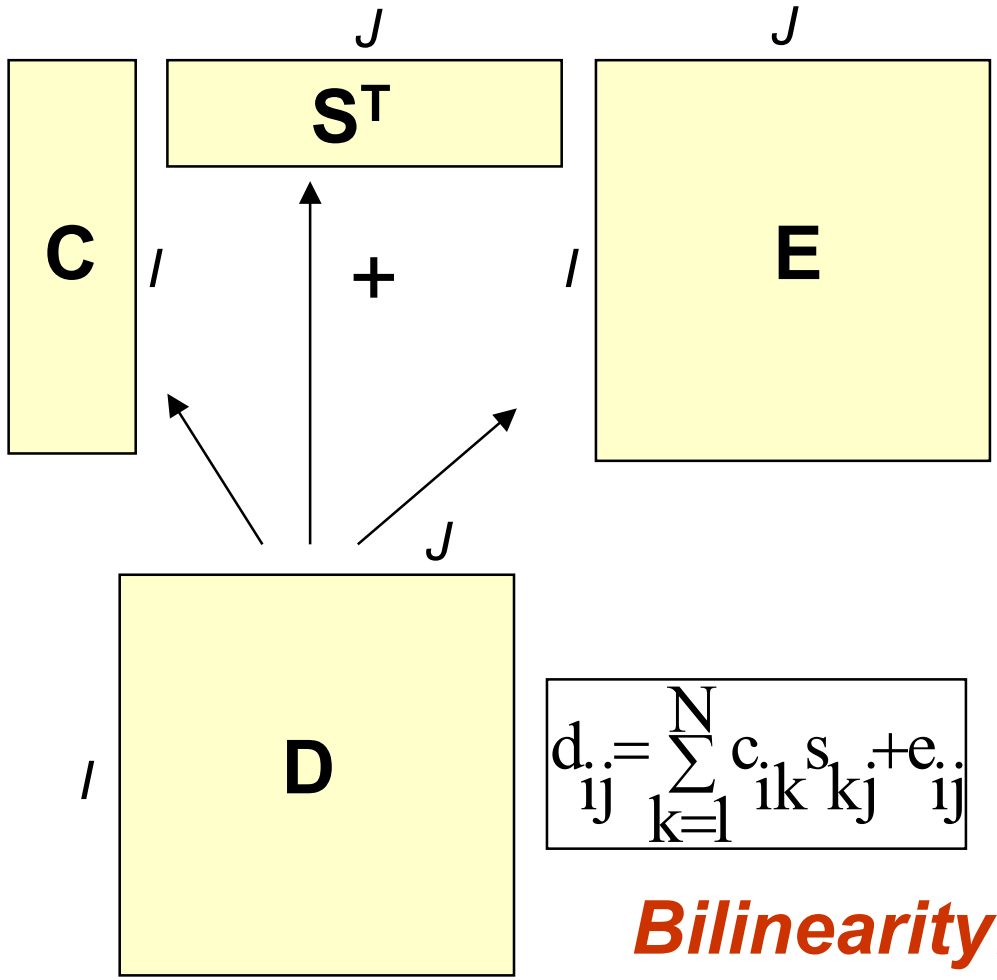


Bilinear models for two-way resolution: Multivariate Curve Resolution

(reaction/process data modeling)







$$d_{ij} = \sum_{k=1}^{N} c_{ik} s_{kj} + e_{ij}$$

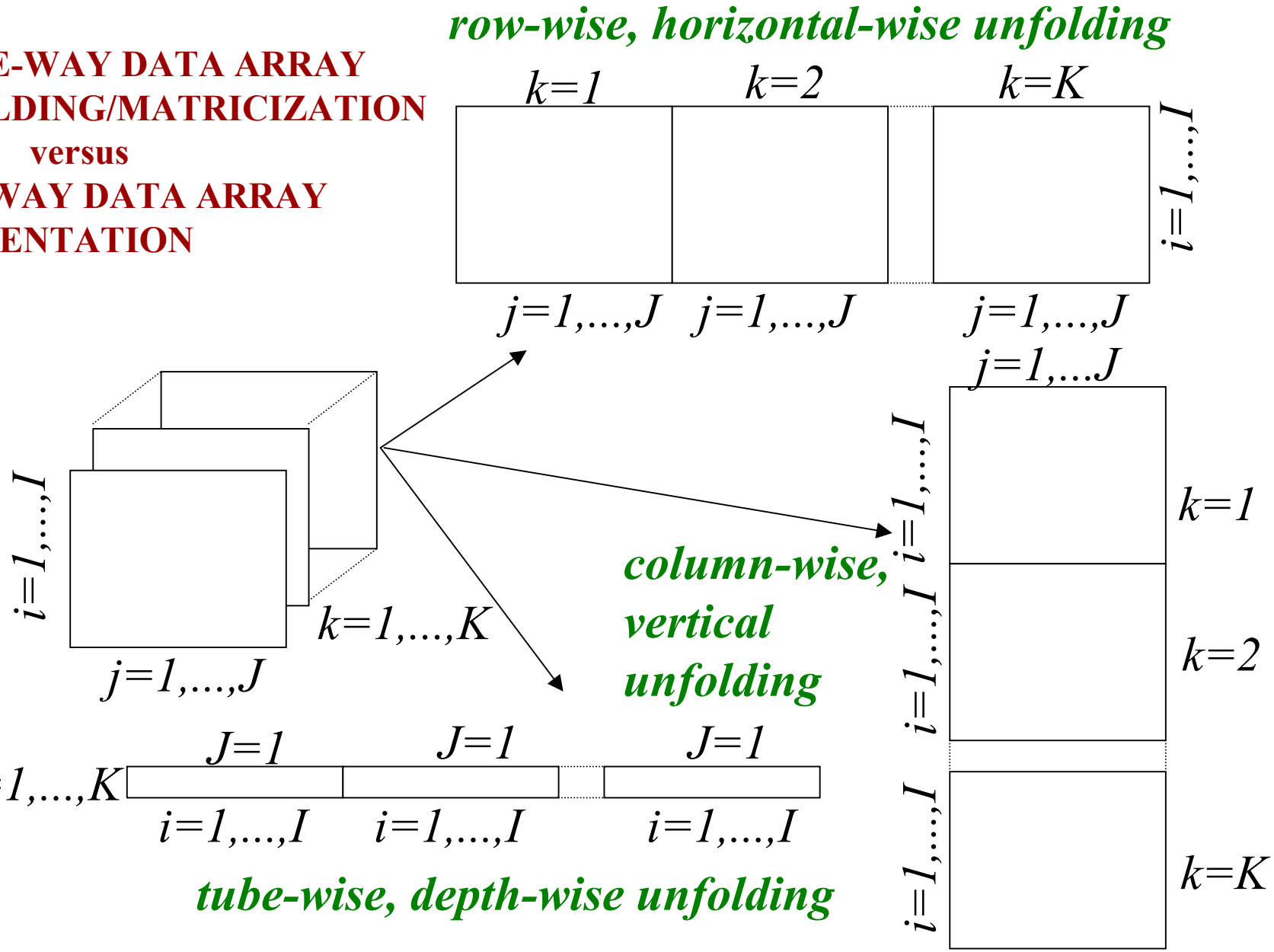
Bilinearity!

Extension of bilinear models to three-way data

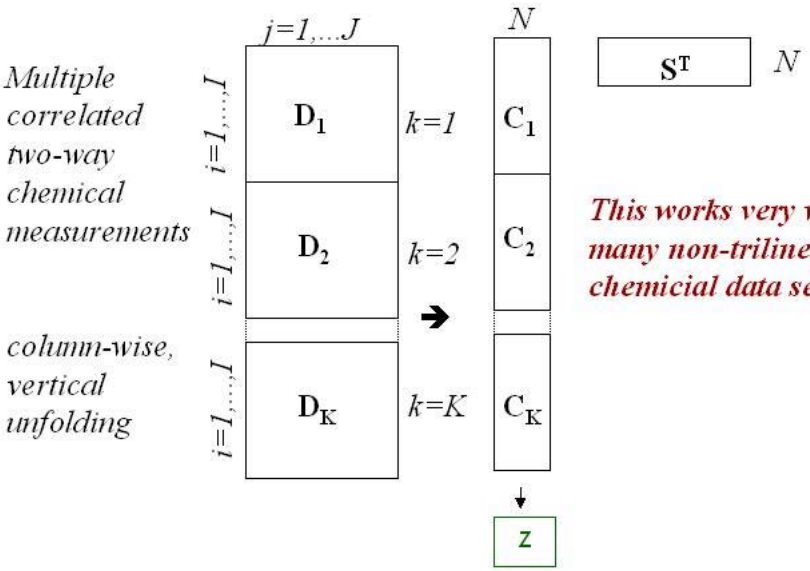
THREE-WAY DATA ARRAY UNFOLDING/MATRICIZATION

versus

TWO-WAY DATA ARRAY AUGMENTATION

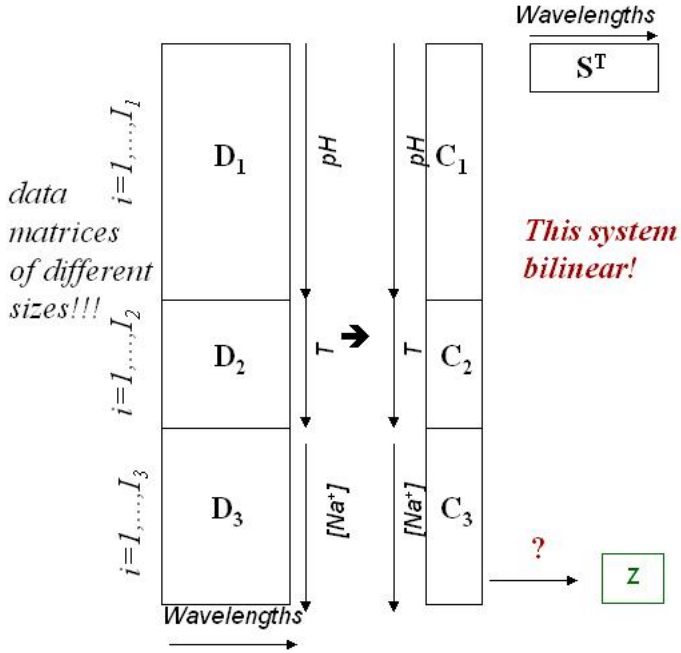


Multivariate Curve resolution for Three Way data



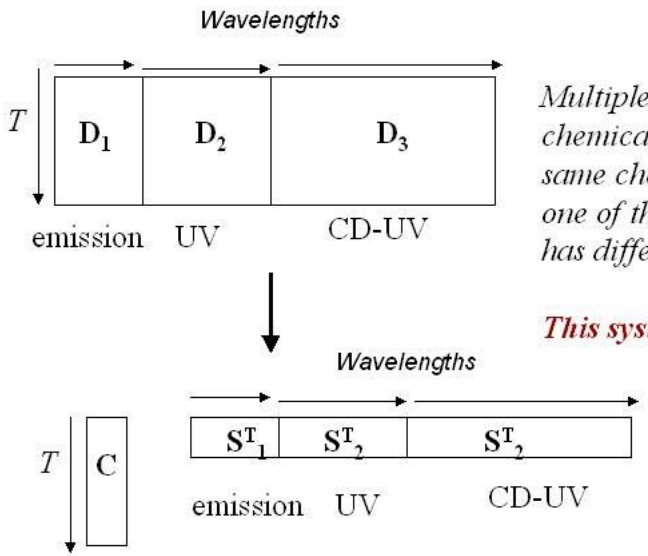
This works very well for many non-trilinear chemicial data sets

Multivariate Curve resolution for Three Way data



This system is still bilinear!

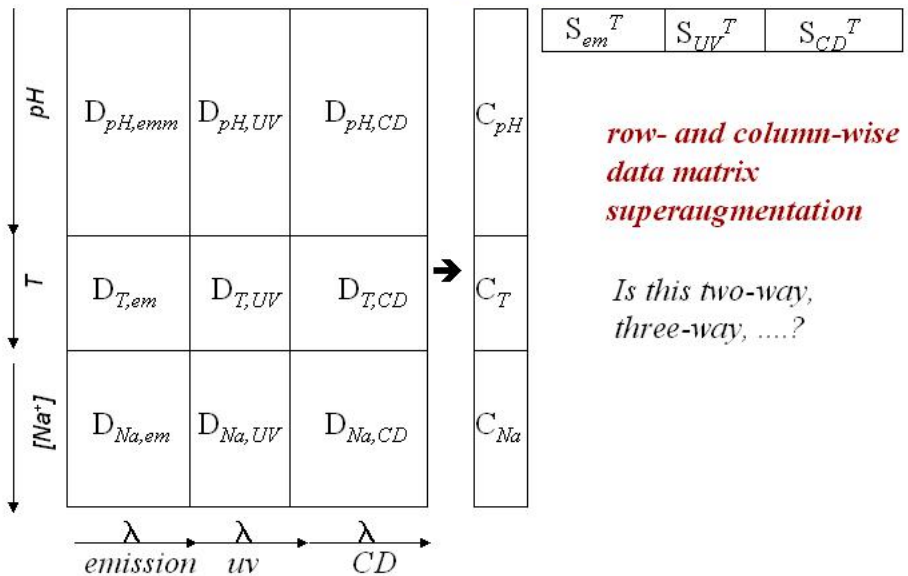
Multivariate Curve Resolution for Three Way data



Multiple correlated two-way chemical measurements (the same chemical system where one of the two orders has different size)

This system is bilinear!

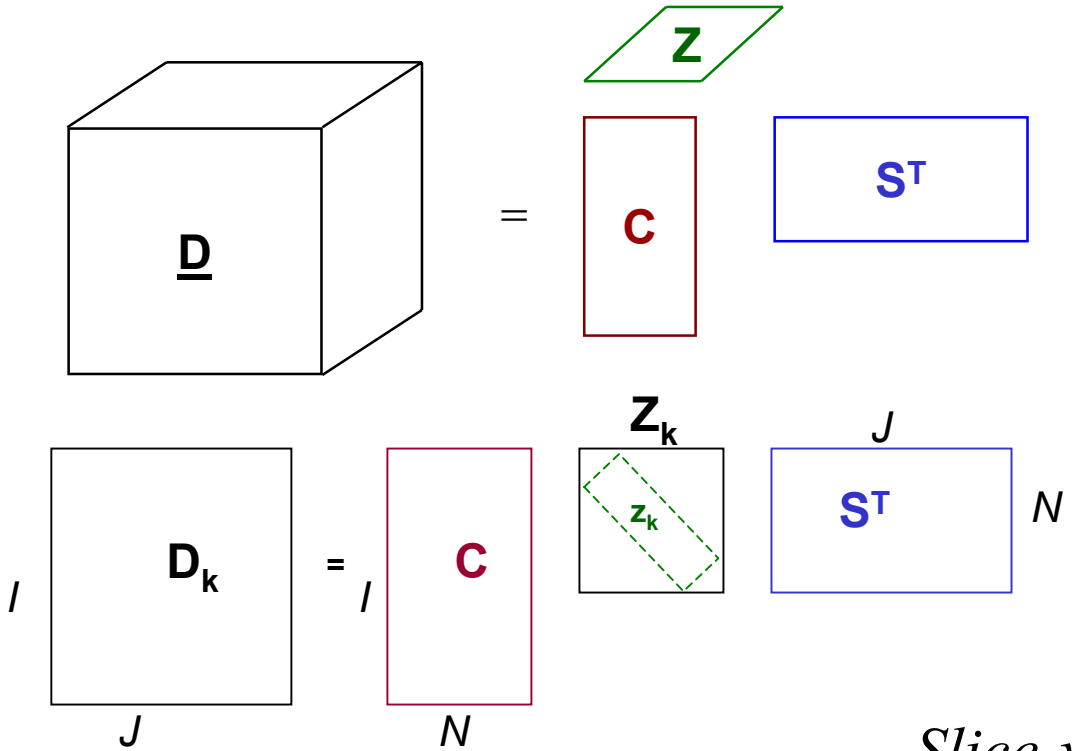
Multivariate Curve Resolution for Three Way data
Data Matrix Superaugmentation



row- and column-wise data matrix superaugmentation

Is this two-way, three-way, ...?

Trilinear models for three-way data: PARAFAC



- Same number of components in the different modes
- No interaction between components in different modes

$$D_k = C Z_k S^T$$ ← *Slice-wise representation*

$$d_{ijk} = \sum_{n=1}^{N} c_{in} s_{jn} z_{kn} + e_{ijk}$$ ← *Element-wise representation*

Stretched/unfolded representation ←

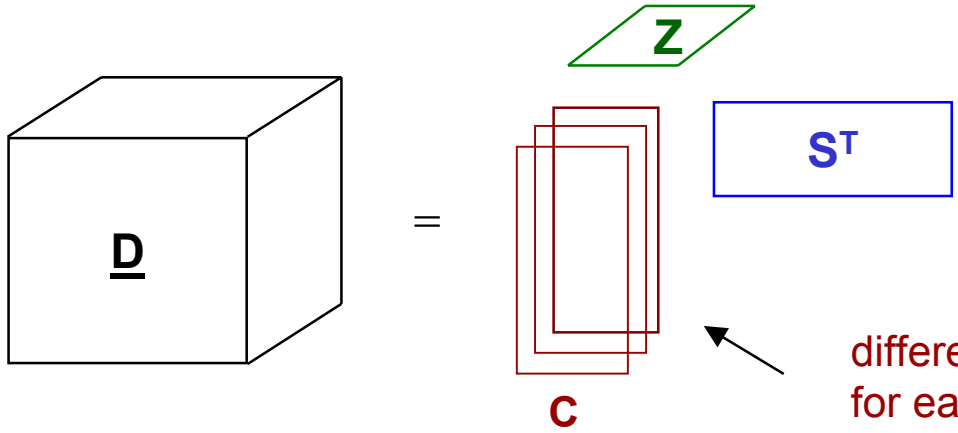
$$D_r^T = [D_1^T, D_2^T, \ldots, D_K^T] = SI_r^T (Z^T \otimes C^T)$$

Trilinear models advantages:

- **Very efficient** in the investigation of complex three-way data structures.
- **They provide unique solutions** avoiding the presence of factor analysis rotation ambiguities, frequently present when bilinear models are applied to two-way data.

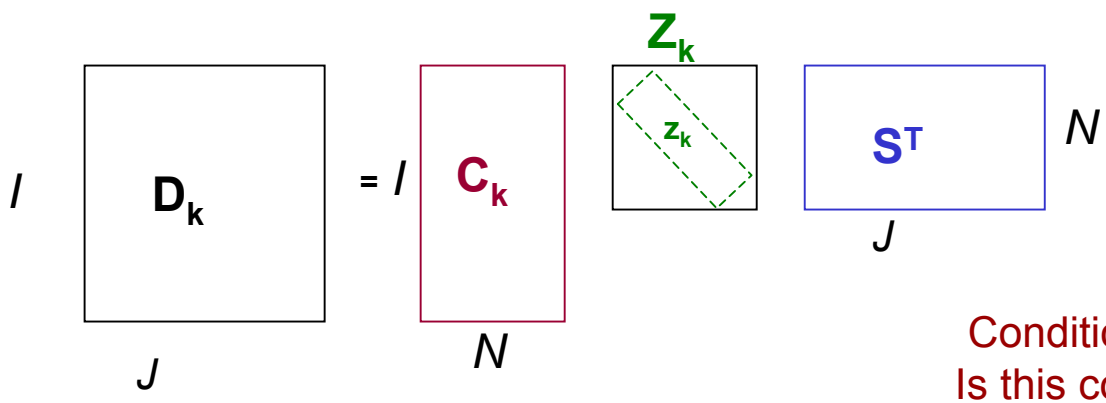
Trilinear models disadvantages

- **Very (or too!) rigid/constrained** in practice
- Many times, strictly trilinear models are not appropriate for the **resolution** of underlying physic-chemical models nor **for the estimation of the 'true' vector profiles** causing the observed data variance



different C_k matrices
for each D_k slab
C profiles are not required
to be identical in all slabs!

PARAFAC2

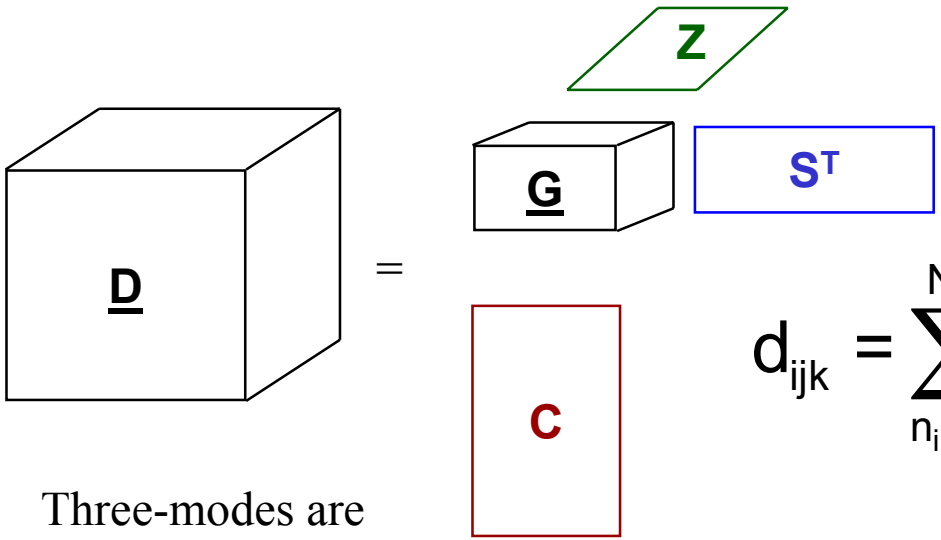


$$D_k = C_k Z_k S^T$$

$$C_1 C_1^T = C_2 C_2^T = \ldots = C_k C_k^T$$

Condition to be fulfilled.
Is this condition too rigid?
What really means?

Non-trilinear models for three-way data: Tucker3 models



Three-modes are reduced!

- Different number of components in the different modes $Ni \neq Nj \neq Nk$
- Interaction between components in different modes is possible

Element-wise representation

$$d_{ijk} = \sum_{n_i=1}^{N_I} \sum_{n_j=1}^{N_J} \sum_{n_k=1}^{N_K} c_{in_i} s_{jn_j} z_{kn_k} g_{n_i n_j n_k} + e_{ijk}$$

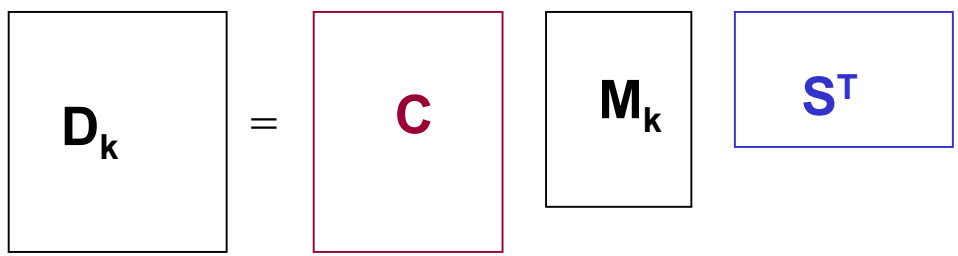
Stretched/unfolded representation

$$D_r^T = [D_1^T, D_2^T, \ldots, D_K^T] = SG_r^T (Z^T \otimes C^T)$$

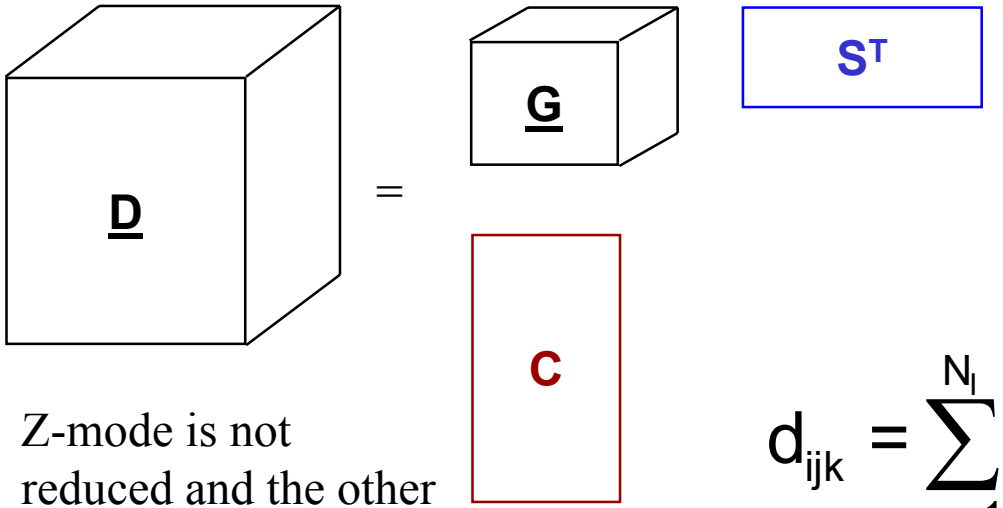
Slice-wise representation

$$D_k = CM_k S^T$$

$$M_k = \left(\sum_{nk=1}^{NK} z_{k,nk} G_k \right)$$



Non-trilinear models for three-way data: Tucker2 models



Z-mode is not reduced and the other two (C- and S^T-modes are reduced

- Different number of components in the two modes $Ni \neq Nj$
- Interaction between components in the two modes is possible

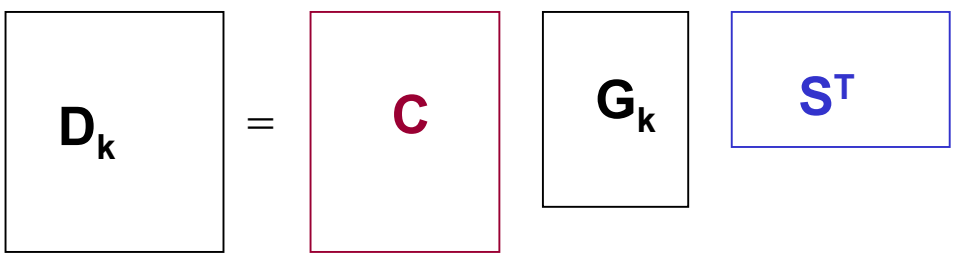
Element-wise representation

$$d_{ijk} = \sum_{n_i=1}^{N_I} \sum_{n_j=1}^{N_J} c_{in_i} s_{jn_j} g_{n_i n_j k} + e_{ijk}$$

Stretched/unfolded representation

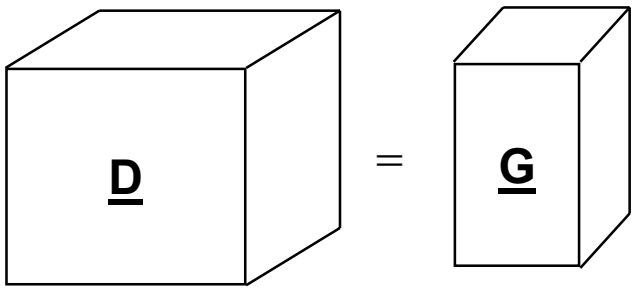
$$D_r^T = [D_1^T, D_2^T, \ldots, D_K^T] = S G_r^T (I_K \otimes C^T)$$

Slice-wise representation



$$D_k = C G_K S^T$$

Non-trilinear models for three-way data: Tucker1 models



S^T

Interaction between components in different modes is not possible

Element-wise representation

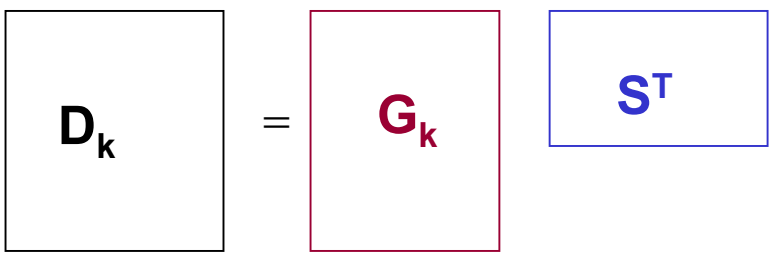
$$d_{ijk} = \sum_{n_j=1}^{N_J} g_{n_j i\, k} s_{j\, n_j} + e_{ijk}$$

Only S^T mode is reduced!
C and Z modes are in G

Stretched/unfolded representation

$$D_r^T = [D_1^T, D_2^T, \ldots, D_K^T] = SG_r^T (I_K \otimes I_J)$$

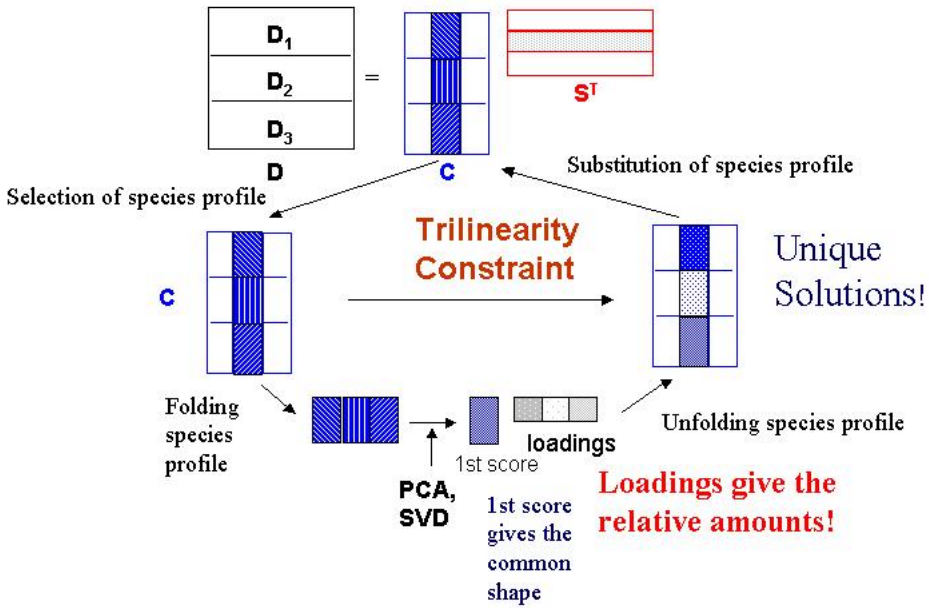
Slice-wise representation



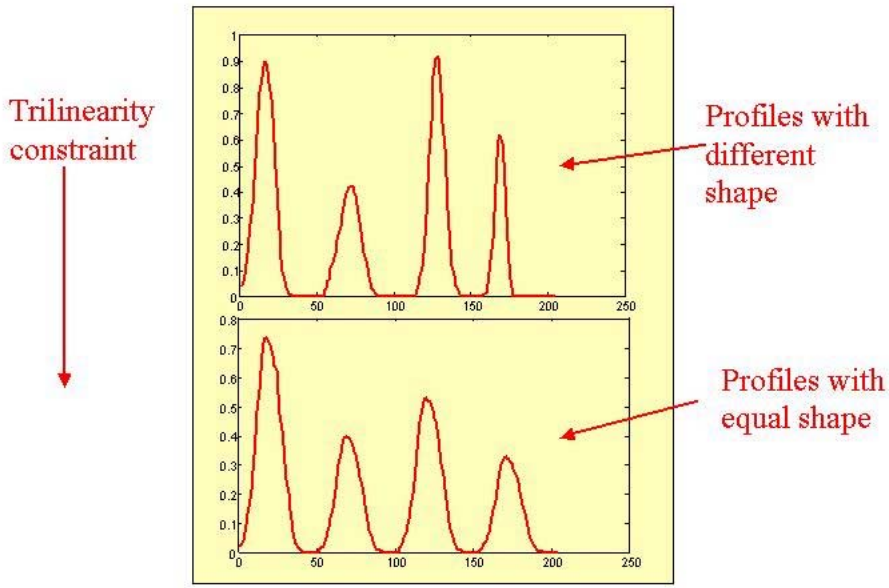
$$D_k = G_k S^T$$

Tucker1 model is equivalent to unfolded bilinear model!!

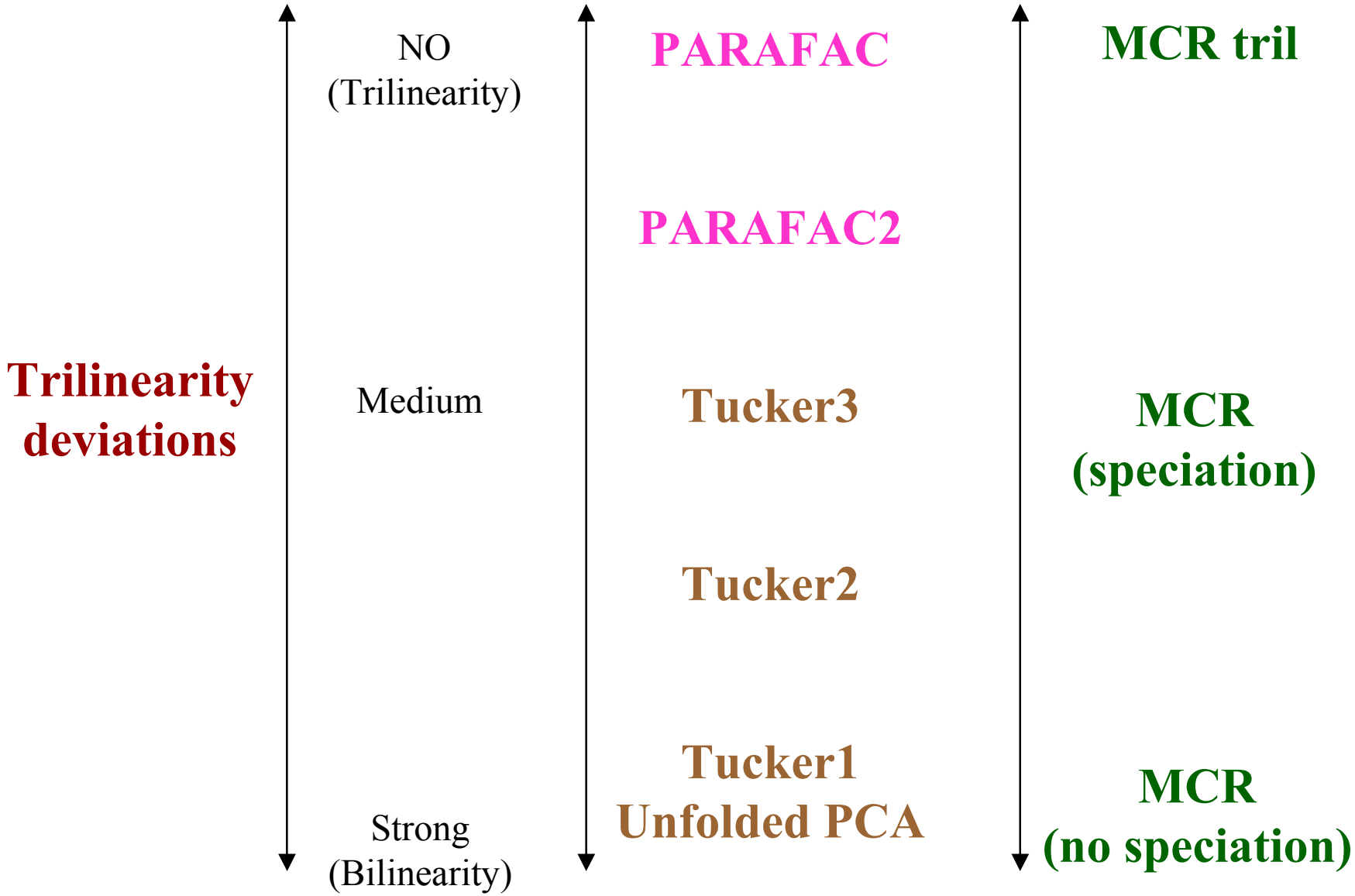
Trilinearity can be implemented independently for each component (chemical species) in MCR-ALS!



Effect of application of a trilinearity constraint



Three-way models options



Outline:

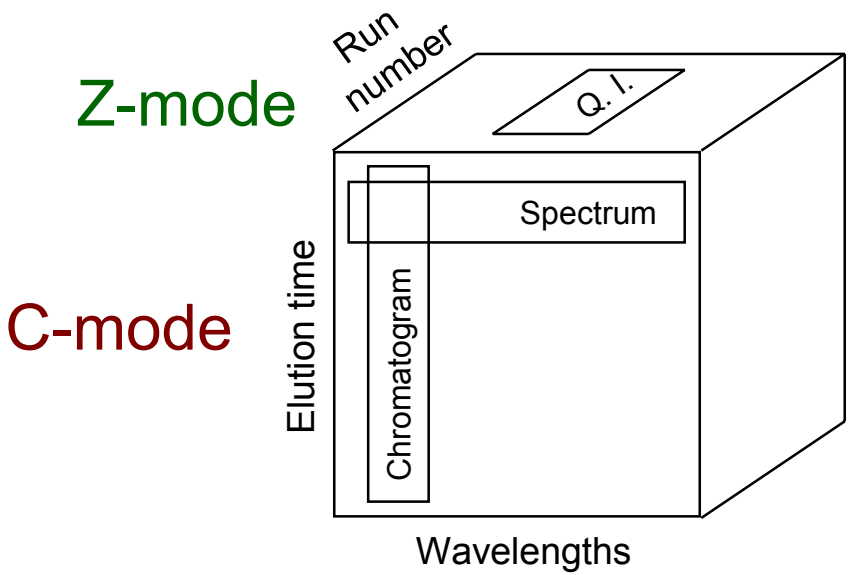
- Introduction and motivations of this work
- Models and structures for complex (three-way) chemical measurements data
- Results of comparison of models and methods to analyze complex (three-way) chemical measurements data
- Conclusions

HPLC-DAD DATA SETS

C-mode: chromatographic profiles.

S-mode: spectra profiles.

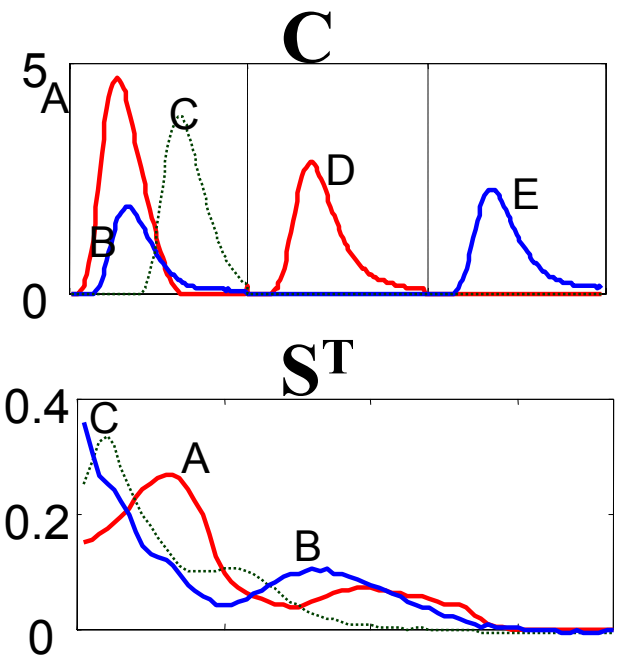
Z-mode: quantitative profiles



DATA SET 1 (real data): LC-DAD determination of organophosphorous pesticides in natural waters

Total nr. of chemical compounds: 3. (A,B known, C unknown)
Nr. of pure spectra: 3
Nr. of chromatographic profiles: 5

Nr. of slabs (data matrices): 3
D_1 (A,B,C)
D_2 (A standard)
D_3 (B standard)

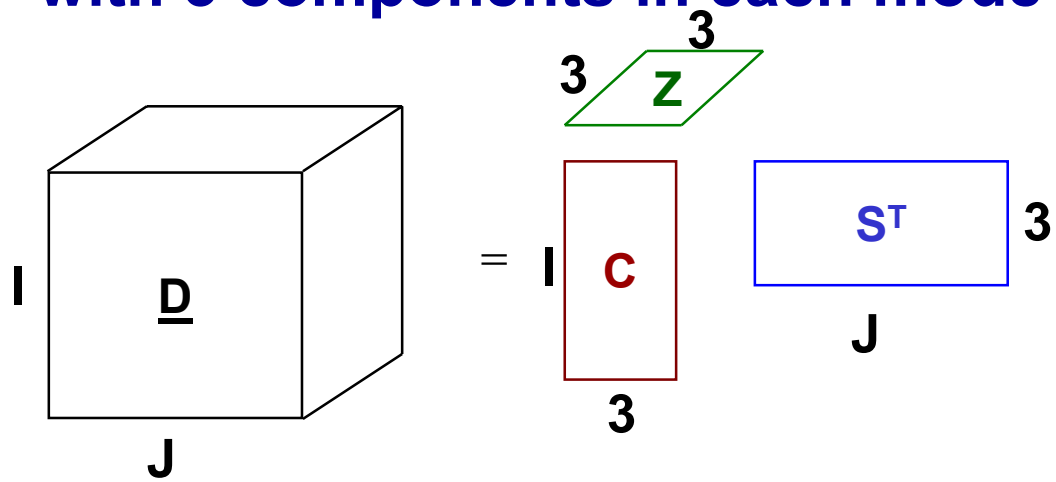


Every slab (data matrix) is bilinear!

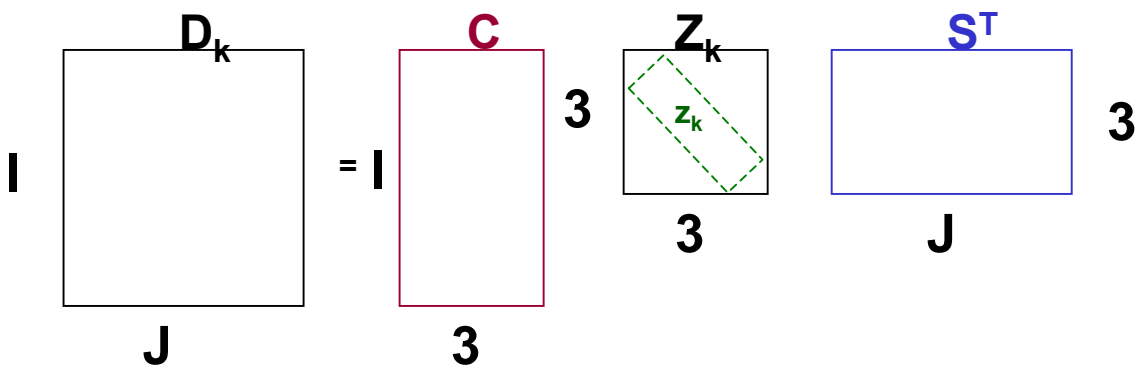
$$D_1 = c_A s^T_A + c_B s^T_B + c_C s^T_C + E_1$$
$$D_2 = c_D s^T_A + E_2$$
$$D_3 = c_E s^T_B + E_3$$

Data are not trilinear since c_A and c_D and c_B and c_E are different in shift and shape

Building three-way models: PARAFAC model is built with 3 components in each mode



All slabs are modeled with the same **C** and $\mathbf{S}^T$ considering only three profiles in each mode!

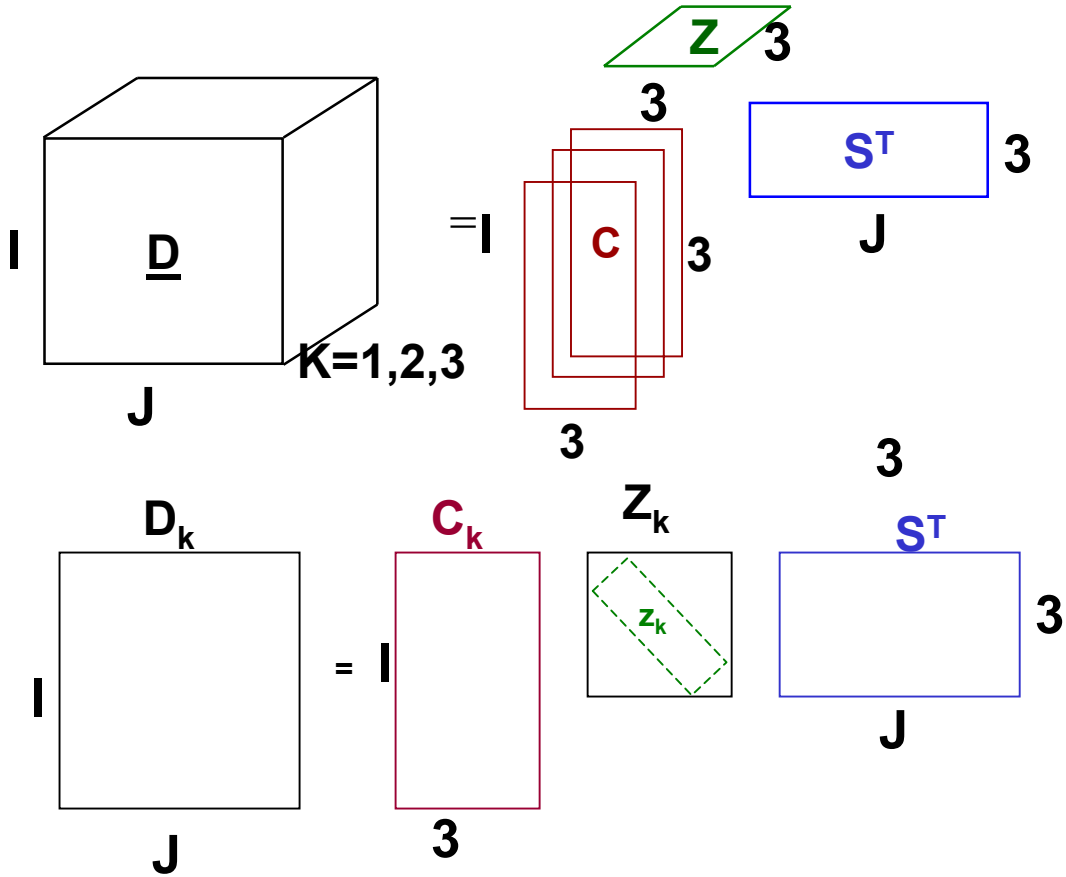


$$\mathbf{D}_k = \mathbf{C}\,\mathbf{Z}_k\,\mathbf{S}^{Tc}$$

PARAFAC stretched (unfolded) representation

$$\left[\mathbf{D}_1^T \mathbf{D}_2^T \mathbf{D}_3^T\right] = \mathbf{S}\left[\mathbf{I}_1^T \mathbf{I}_2^T \mathbf{I}_3^T\right](\mathbf{Z}^T \otimes \mathbf{C}^T)$$

Building three-way models: PARAFAC2 model is built with 3 components in each mode



All slabs are modeled with the same $\mathbf{S}^T$ but different $\mathbf{C}_k$, considering three profiles in each mode! Components in C-mode can be slightly different!

$$\mathbf{D}_k = \mathbf{C}_k \mathbf{Z}_k \mathbf{S}^T$$

However this condition should be obeyed:

$$\mathbf{C}_1\mathbf{C}_1^T = \mathbf{C}_2\mathbf{C}_2^T = \ldots = \mathbf{C}_k\mathbf{C}_k^T$$

What this really means?

Building three-way models: MCR model is built using 3 components in the S mode and 3 components in the (unfolded) augmented C mode

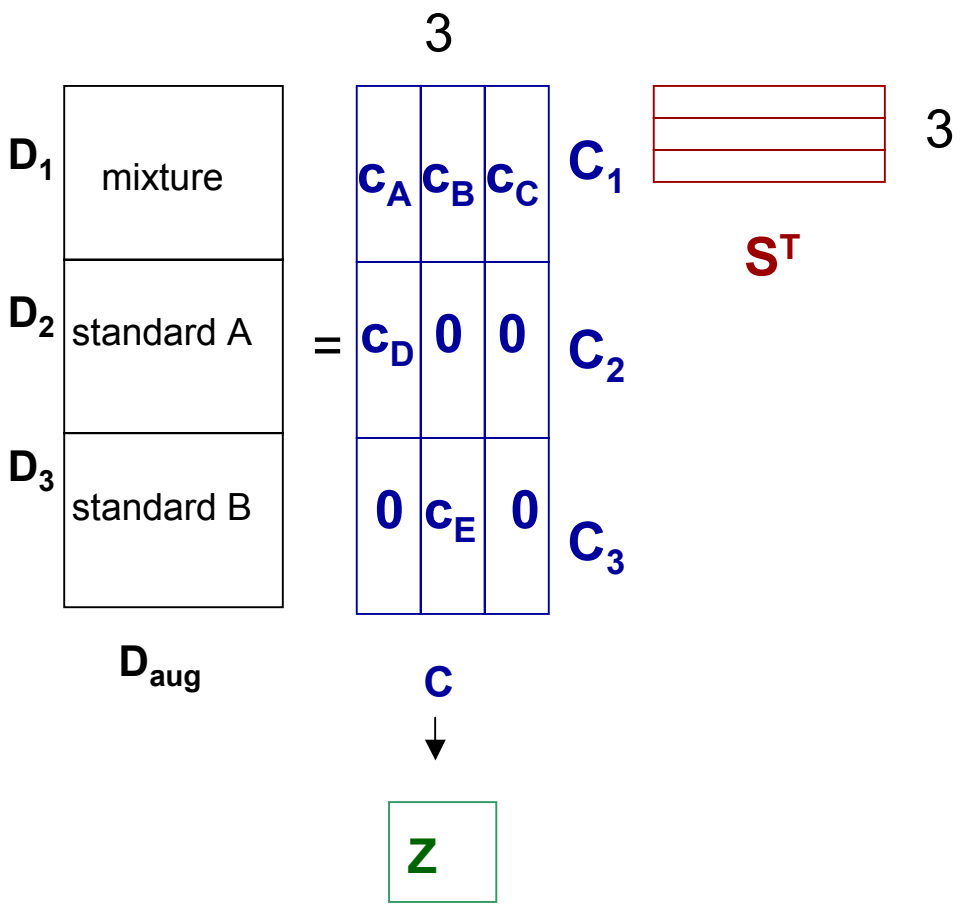
Data set 1

$$D_c = [D_1;D_2;D_3] = [C_1;C_2;C_3]\, S^T$$

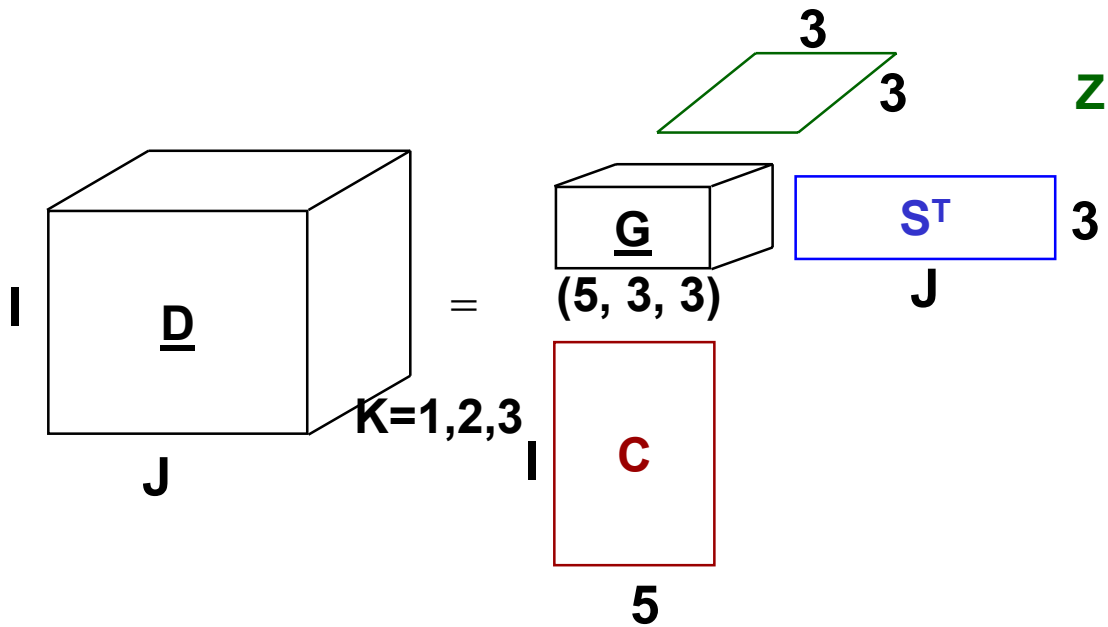
$$D^T_r = [D^T_1\ D^T_2\ D^T_3] = S\,[C^T_1\ C^T_2\ C^T_3]$$

Every slabs is modeled with three different profiles in C_1, C_2, C_3, but with the same three profiles in S^T

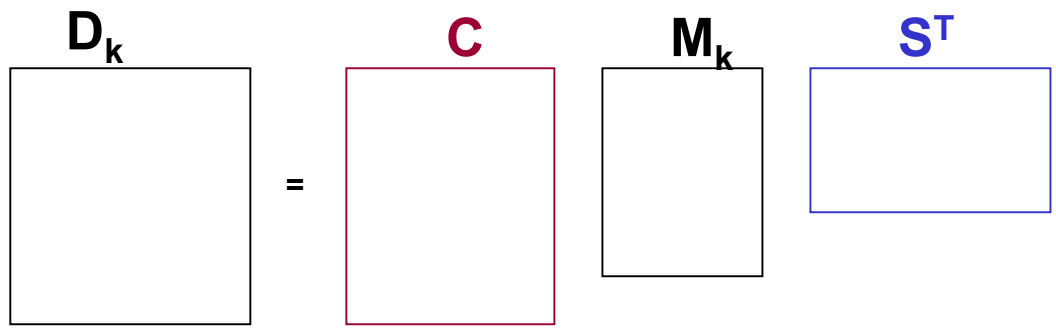
$$D_k = C_k\, S^T$$



Building three-way models: Restricted Tucker3 model is built using 5 components in the C-mode and 3 components in the S- and Z-modes



$$D_k = C\left(\sum_{nk=1}^{3} z_{k,nk} G_k\right) S^T$$



Fit values for data set 1

Method	Fit %
PARAFAC	93.0
PARAFAC2	98.7
MCR	98.0
Tucker3	97.8

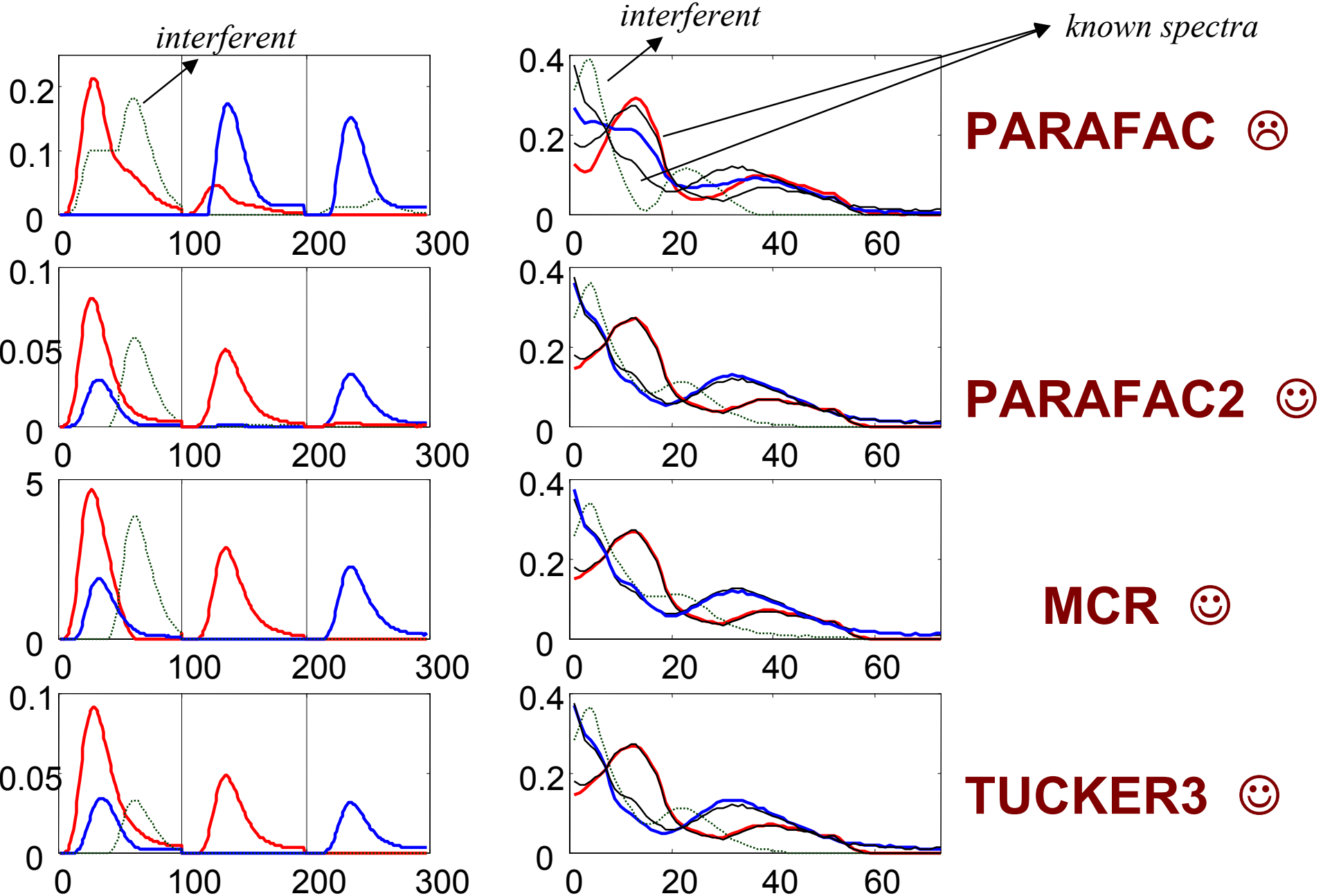
Applied constraints:
Non-negativity
Unimodality

Different type of initial Estimates

Maximum number of Iterations: 100

$$Fit\% = 100\left(1 - \sqrt{\frac{\sum_{i,j,k} e_{ijk}^2}{\sum_{i,j,k} d_{ijk}^2}}\right)$$

COMPARISON OF RESOLVED PROFILES

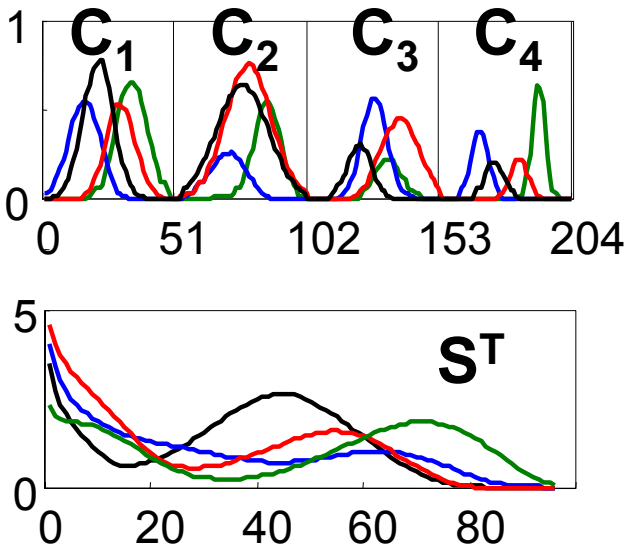


DATA SET 2

Total nr. of compounds: 4. (A,B, C and D)
Nr. of pure spectra: 4
Nr. of chromatographic profiles: 16

Nr. of slabs (data matrices): 4
D_1, D_2, D_3, D_4 (A,B,C,D)

Two data sets, with and without noise

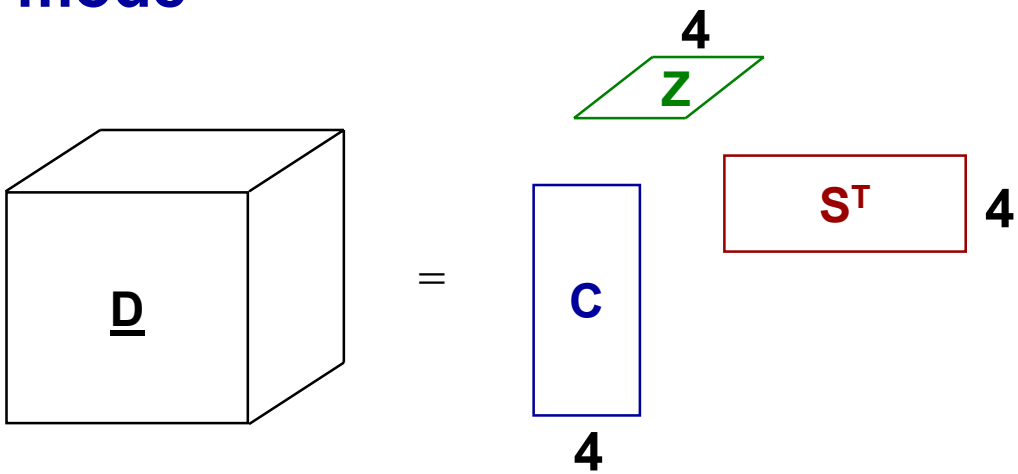


Every slab (data matrix) is bilinear!

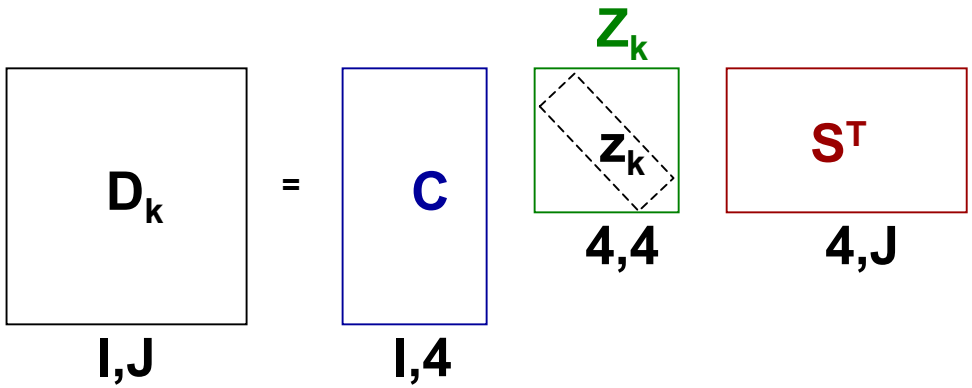
$$D_1 = c_A s^T_A + c_B s^T_B + c_C s^T_C + c_D s^T_D + E_1$$
$$D_2 = c_E s^T_A + c_F s^T_B + c_G s^T_C + c_H s^T_D + E_2$$
$$D_3 = c_I s^T_A + c_J s^T_B + c_K s^T_C + c_L s^T_D + E_3$$
$$D_1 = c_M s^T_A + c_N s^T_B + c_O s^T_C + c_P s^T_D + E_4$$

Data are not trilinear since concentration profiles of A, B, C and D are different in shift and shape!

Building three-way models: PARAFAC model is built with 3 components in each mode



All slabs are modeled with the same **C** and $\mathbf{S}^T$ considering four profiles in all modes!



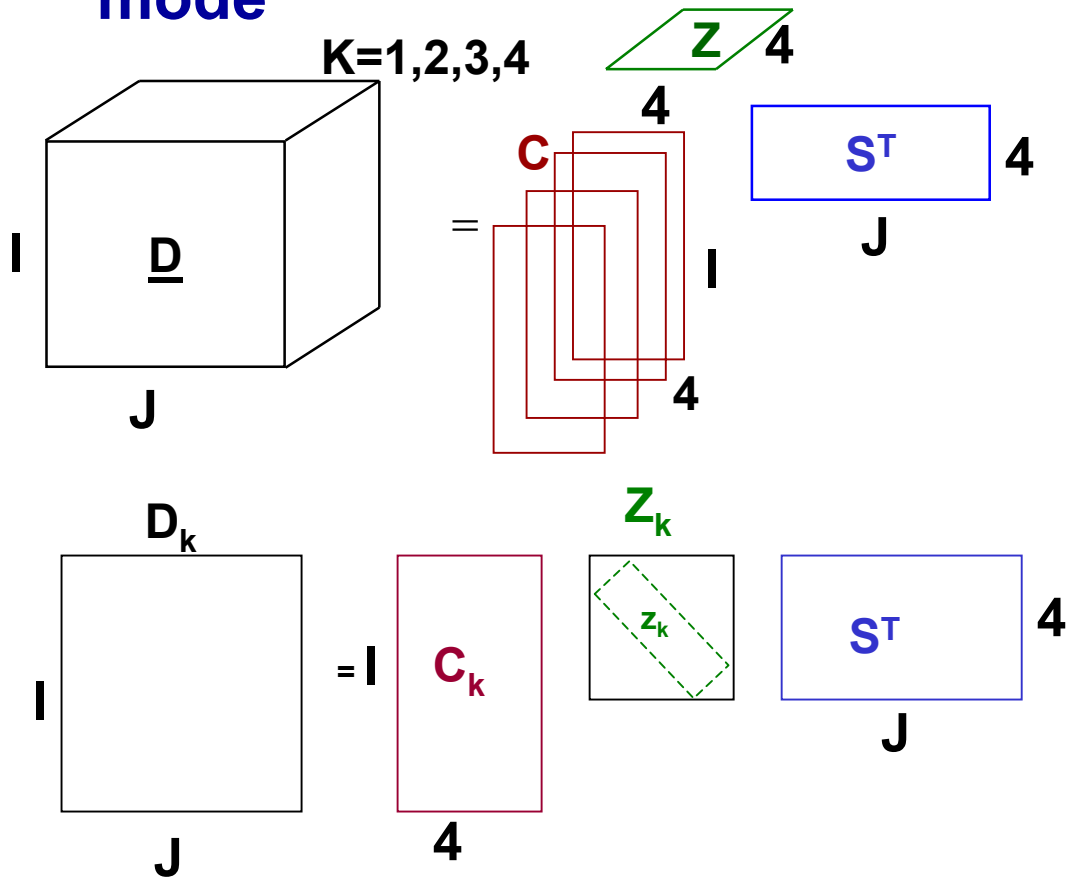
PARAFAC slice-wise representation

$$\mathbf{D}_k = \mathbf{C}\,\mathbf{Z}_k\,\mathbf{S}^T$$

PARAFAC stretched (unfolded) representation

$$\left[\mathbf{D}_1^T \mathbf{D}_2^T \mathbf{D}_3^T \mathbf{D}_4^T\right] = \mathbf{S}\left[\mathbf{I}_1^T \mathbf{I}_2^T \mathbf{I}_3^T \mathbf{I}_4^T\right](\mathbf{Z}^T \otimes \mathbf{C}^T)$$

Building three-way models: PARAFAC2 model is built with 3 components in each mode



All slabs are modeled with the same $\mathbf{S}^T$ but different $\mathbf{C}_k$ and considering only three profiles in each mode! Components in C-mode can be different!

$$\mathbf{D}_k = \mathbf{C}_k \mathbf{Z}_k \mathbf{S}^T$$

$$\mathbf{C}_1\mathbf{C}_1^T = \mathbf{C}_2\mathbf{C}_2^T = \ldots = \mathbf{C}_k\mathbf{C}_k^T$$

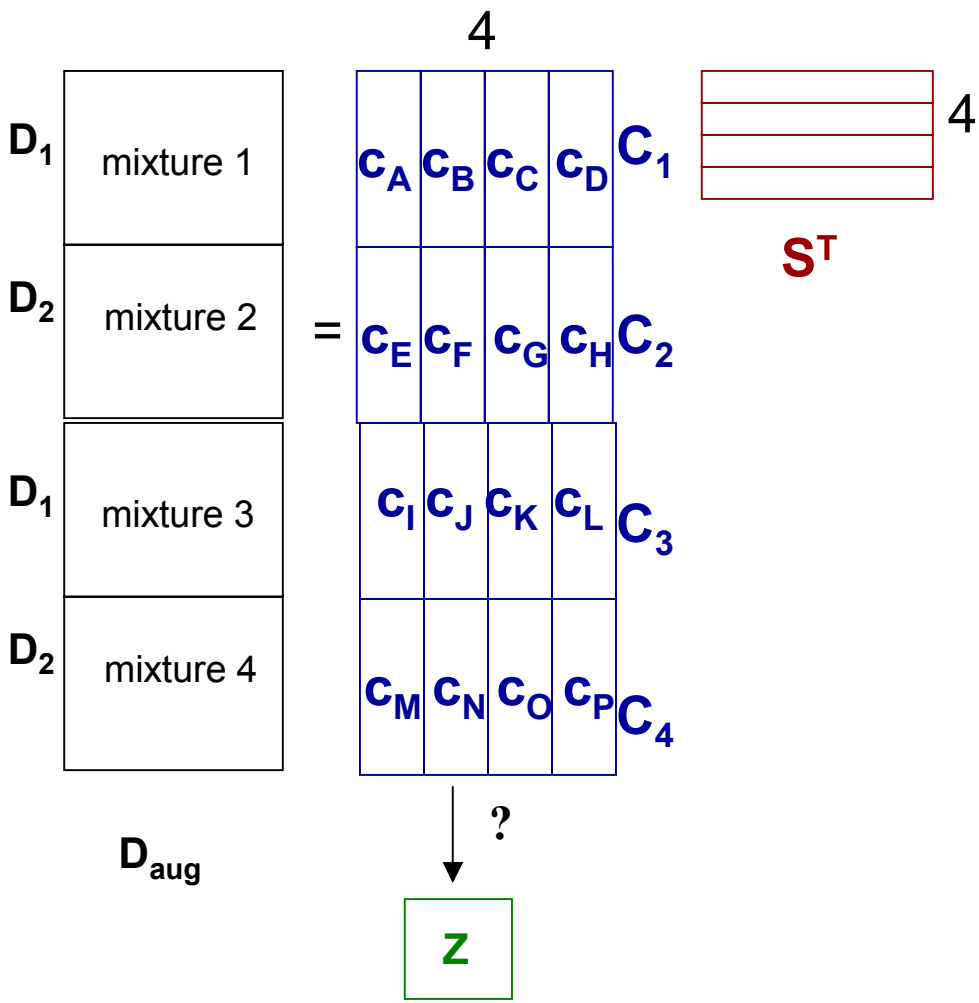
Building three-way models:
MCR model is built using four components in the S mode and four components in the (unfolded) augmented C mode

$$D_{aug} = [D_1;D_2;D_3;D_4] = [C_1;C_2;C_3;C_4]\, S^T$$

$$D^T_{aug} = [D^T_1\, D^T_2\, D^T_3\, D^T_4] = S\, [C^T_1\, C^T_2\, C^T_3\, C^T_4]$$

Every slab is modeled with four different profiles in C_1, C_2, C_3, C_4 but with the same four profiles in S^T

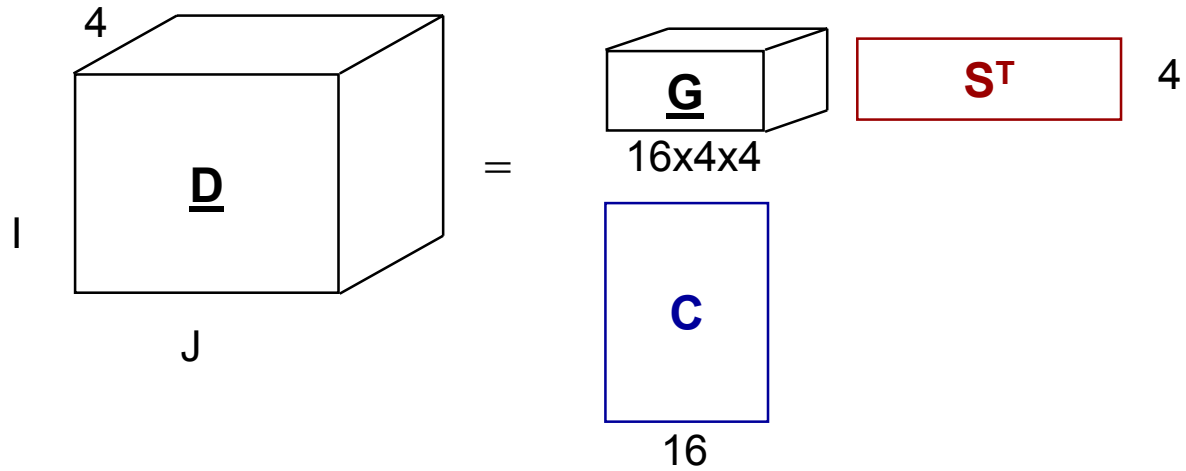
$$D_k = C_k\, S^T$$



Building three-way models:Tucker2 model is built using 16 components in the C-mode and 4 components in the S- mode

It is not possible to build a Tucker3 model, with the 3 modes reduced

Z-mode is confounded in C-mode!

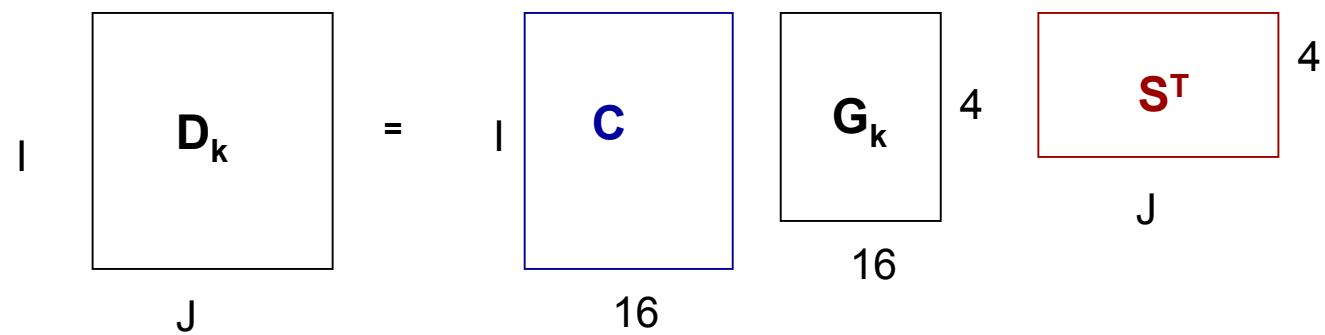


Stretched/unfolded representation

$$\left[D_1^T D_2^T D_3^T D_4^T\right] = S\left[G_1^T G_2^T G_3^T G_4^T\right]\left(\left[I_1\ I_2\ I_3\ I_4\right] \otimes C^T\right)$$

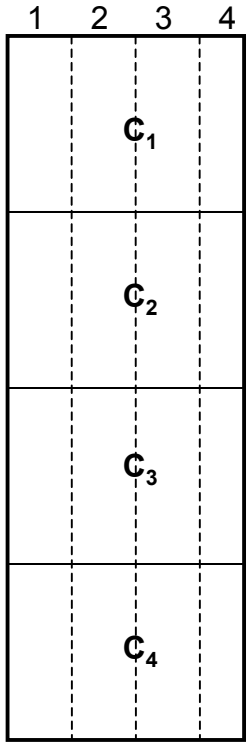
Slice-wise representation

$$D_k = C\ G_k\ S^T$$



Building three-way models Tucker2 model. How is G ?

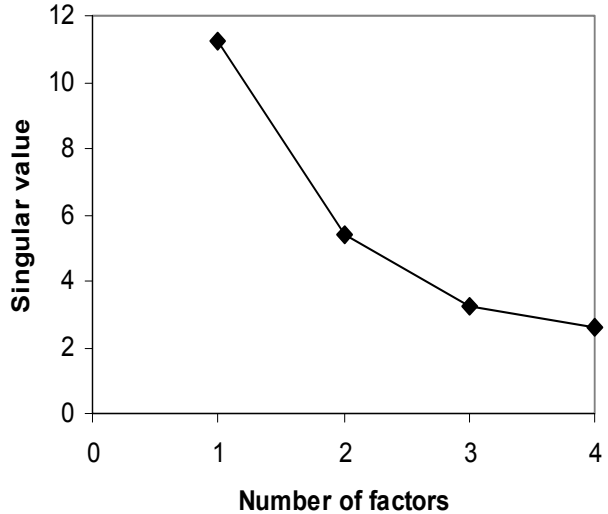
$$G = \left(\begin{array}{cccc|cccc|cccc|cccc}
 & G_1 & & & & G_2 & & & & G_3 & & & & G_4 & & \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{array}\right)$$



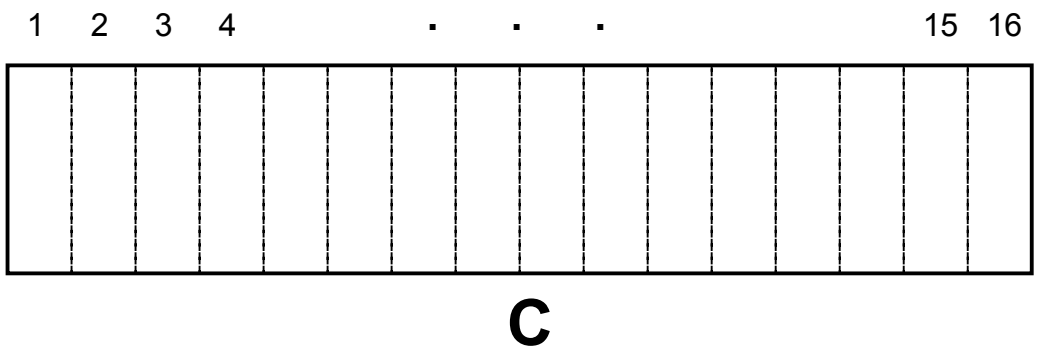
Concentration profiles used in the simulation

C in MCR is full rank 4!

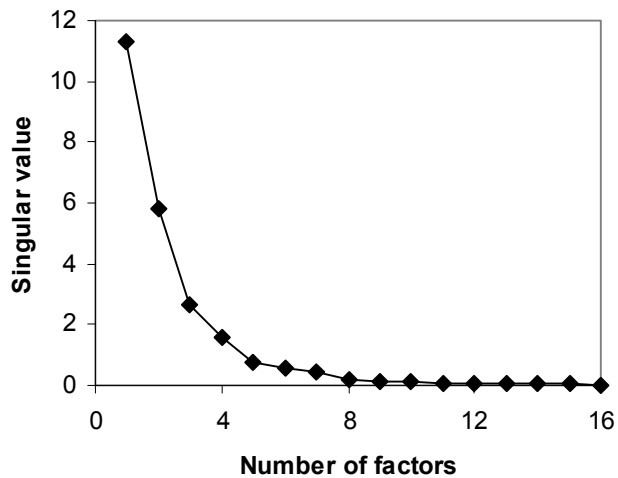
Condition Number is 4.3



Calculation of C^+ ?



C in Tucker2 is close to rank deficient!
Condition number is 405.8



Fit values for data set 2 (noise free)

Method	Fit %
PARAFAC	91.6
PARAFAC2	93.6
MCR	99.9
Tucker3	99.9

Applied constraints:
Non-negativity
Unimodality

Different type of initial Estimates

Maximum number of Iterations: 100

$$Fit\% = 100\left(1 - \sqrt{\frac{\sum_{i,j,k} e_{ijk}^2}{\sum_{i,j,k} d_{ijk}^2}}\right)$$

Fit values for data set 2

(heterocedastic proportional added noise 6.71%)

Method	Fit %
PARAFAC	89.3
PARAFAC2	93.4
MCR	93.3
Tucker3	93.5

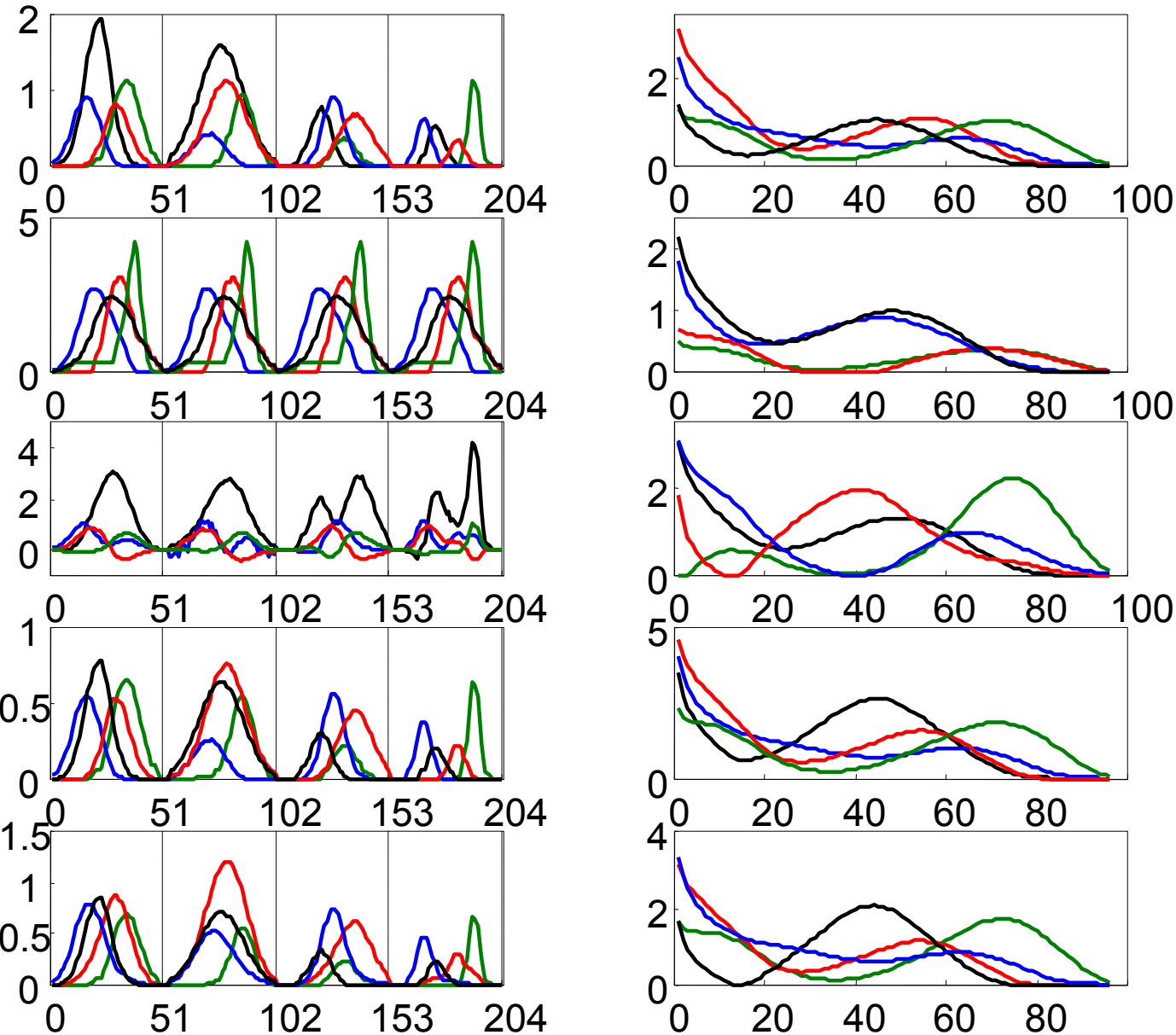
Applied constraints:
Non-negativity
Unimodality

Different type of initial Estimates

Maximum number of Iterations: 100

$$Fit\% = 100\left(1 - \sqrt{\frac{\sum_{i,j,k} e_{ijk}^2}{\sum_{i,j,k} d_{ijk}^2}}\right)$$

COMPARISON OF RESOLVED PROFILES
(noise free case)



TRUE PROFILES

PARAFAC ☹

PARAFAC2 ☹

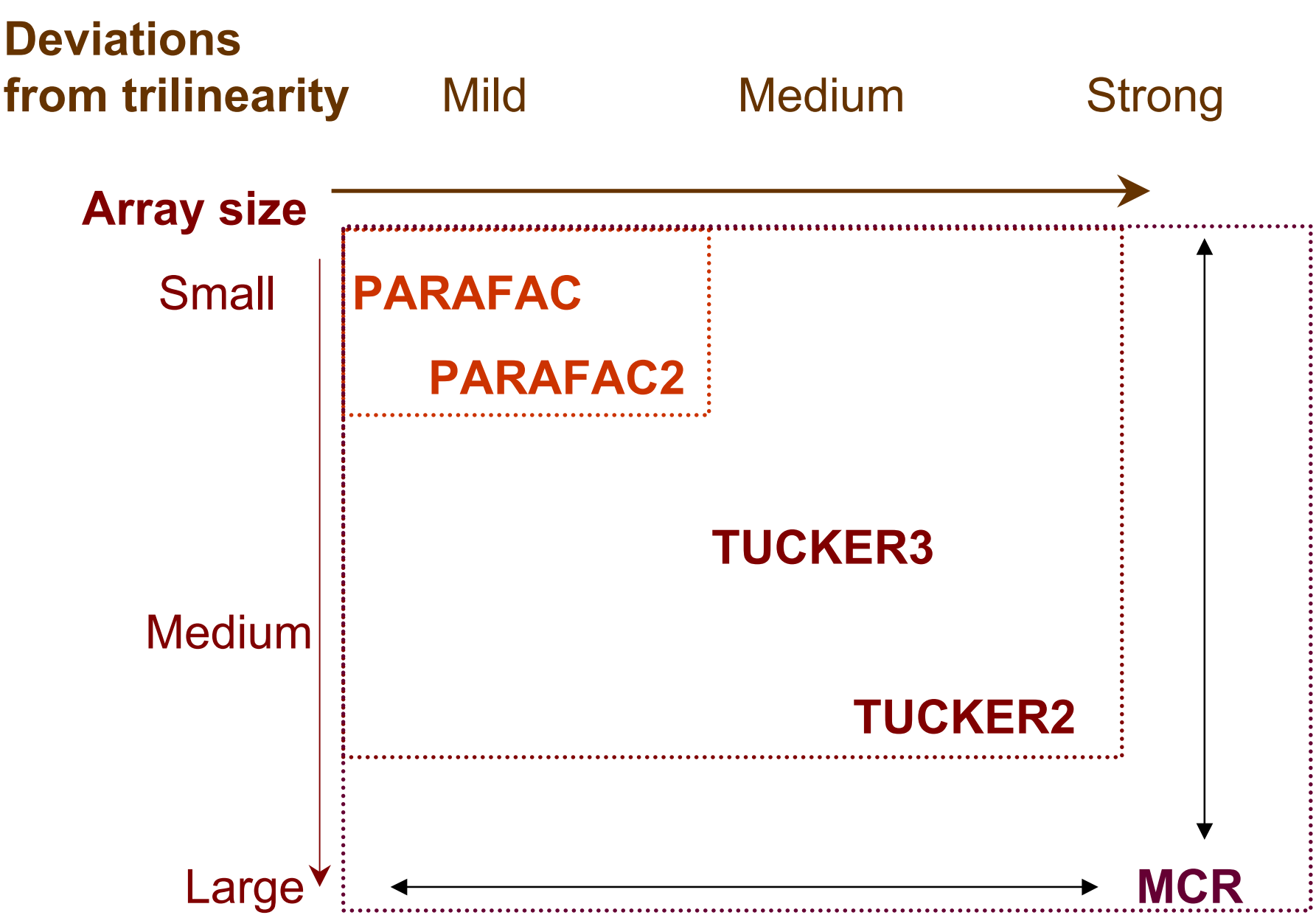
MCR ☺

TUCKER2 😐

Outline:

- Introduction and motivations of this work
- Models and structures for complex (three-way) chemical measurements data
- Results of comparison of models and methods to analyze complex (three-way) chemical measurements data
- Conclusions

Guidelines for method selection



CONCLUSIONS

✓PARAFAC performance is extremely vulnerable to deviations from trilinearity.

Performance diagnostic: comparison of lack of fit between PARAFAC and any other non-trilinear model-based method.

Similar lacks of fit

$\Rightarrow$ trilinear system

$\Rightarrow$ use recommended

Higher lack of fit for PARAFAC

$\Rightarrow$ non-trilinear system

$\Rightarrow$ avoid use

CONCLUSIONS

- ✓ PARAFAC2 requires the presence of strongly patterned deviations from trilinearity
  ($\mathbf{C_1C_1^T = C_2C_2^T = ... = C_kC_k^T}$).

C-mode (e.g., elution profiles) is unconstrained.

Performance diagnostic: examination of profile shape in C-mode.

Chemically meaningful shapes

$\Rightarrow$ PARAFAC2 pattern

$\Rightarrow$ use recommended

Chemically meaningless shapes

$\Rightarrow$ no PARAFAC2 pattern

$\Rightarrow$ avoid use

CONCLUSIONS

✓Restricted TUCKER and MCR perform similarly while not working with large data arrays.

✓Pseudoinversion of matrix and distinction of profiles related to the elution mode is more stable and gives better results for the MCR **C** matrix (with augmented **C** profiles) than for the TUCKER **C** matrix.

General Conclusions

✓Chemical measurements provide in many circumstances two-, three- and multi-way data

✓Chemical data usually **do fulfill a bilinear model**

✓Chemical data **do not fulfill a full trilinear model** in many cases

✓**Mixed bilinear and trilinear data models** can be optimal in many circumstances and they can be solved using constrained bilinear models of matricized/unfolded cubes or augmented matrices like in MCR

Software

1. N-way toolbox by C. Andersson and R. Bro.
http://www.models.kvl.dk/source/nwaytoolbox

2. MCR-ALS by R. Tauler and A. de Juan.
http://www.ub.es/gesq/mcr/mcr.htm