Practical Problems in Tensor Modeling
(In chemometrics)

Rasmus Bro
What we work with

Dioxin, Environment, Dose-response

Food quality, Raw material influence, Production optimization

Genomics, Systems biology, Cancer, Diabetes, Pharma...
Fluorescence
High resolution NMR
Mass spectrometry
Near-infrared
Raman
Ultrasound
Hyperspectral Imaging
Imaging
...

Data
Fluorescence

Excitation-emission matrix – a chemical fingerprint
fluorescence

Excitation-emission matrix fingerprint

Food Technology - LMT - KVL - http://models.kvl.dk
Basic

Plotting
Uncertainty estimates
Automated analysis
problems
**Interpretation**
How to interpret a scatter plot

<table>
<thead>
<tr>
<th></th>
<th>Workload</th>
<th>Distance to work</th>
<th>Salary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smith</td>
<td>1.0</td>
<td>0.2</td>
<td>1.2</td>
</tr>
<tr>
<td>Johnson</td>
<td>2.0</td>
<td>0.0</td>
<td>0.3</td>
</tr>
<tr>
<td>Williams</td>
<td>-1.0</td>
<td>0.1</td>
<td>-1.0</td>
</tr>
<tr>
<td>Jones</td>
<td>-2.0</td>
<td>-0.4</td>
<td>-0.1</td>
</tr>
<tr>
<td>Davis</td>
<td>0.0</td>
<td>-0.4</td>
<td>-0.4</td>
</tr>
</tbody>
</table>

[Source](www.models.kvl.dk)
Plotting

• Two-way PCA - orthonormal basis (loadings)
• Hence distances in scores reflect both manifest and latent distances

• PARAFAC/Tucker - Oblique bases
• Distances reflect only latent distances not manifest
Plotting
Plotting the first mode component in scatter plots reflecting *latent* variation.
Plotting on an orthogonal basis reflecting raw data distances
Uncertainty of parameters

\[ S^2 = \frac{\sum_{i=1}^{N} (X_i - \bar{X})^2}{N - 1} \]

No degrees of freedom in PARAFAC (and probably not in other multilinear models)
PARAFAC on fluorescence

A: Concentration
B: Emission
C: Excitation

\[ X = a_1 b_1 c_1 + a_2 b_2 c_2 + a_3 b_3 c_3 + a_4 b_4 c_4 \]
Jack-knifing the model

Leave out sample 1

Sample 2

Sample I

1st jack-knife segment

2nd jack-knife segment

\( i\)th jack-knife segment

I PARAFAC sub-models:
- Standard error
- Outlier detection

Jack-knifing the model

Emission spectral profiles
Removing low excitation and sample #2,3,5,10

Emission spectral profiles
Automatic

Meta-parameters

Goodness

Result
Scatter
Long story ...
Must be approximately valid

- Sufficient number of adequate samples
- Sufficient spectral resolution
- Beers law valid

Then decide

- Low excitation wavelengths to exclude
- How to handle Rayleigh scattering
- Number of components to use
- Outliers to exclude
Goodness criterion

Goodness = Fit*CoreConsistency*Splithalf

\[
FIT = 1 - \frac{\sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} e_{ijk}^2}{\sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} x_{ijk}^2}
\]

\[
COREC = 100 \left( 1 - \frac{\sum_{d=1}^{F} \sum_{e=1}^{F} \sum_{f=1}^{F} (g_{def} - t_{def})^2}{F} \right)
\]
Before EEMizer

After EEMizer

EEMizer result
Traditional approach for cancer diagnostics and monitoring: Biomarkers
Conc. Comp. 4

Cancer
Non-cancer

It works!
Conclusion

Still needed
Better algorithms
Better statistics
Better software
m-files, e-courses, data sets, etc.

www.models.life.ku.dk

If you want lots of papers on applied tensor analysis, come by with a USB