Factor number selection in the tensor decomposition of EEG data:

Mission (im)possible?

Zuzana Rošťáková, Roman Rosipal

Institute of Measurement Science
Slovak Academy of Sciences

MEASUREMENT 2021
Parallel factor analysis (PARAFAC) and EEG

Factor number selection in EEG data

[Harshman, 1970]
Parallel factor analysis (PARAFAC) and EEG

number of factors $F = ?$

[Harshman, 1970]

Factor 1 $A^{(3)}_1$: freq. $+$ Factor 2 $A^{(3)}_2$: freq. $+$ Factor 3 $A^{(3)}_3$: freq. $+$ Factor 4 $A^{(3)}_4$: freq. $+$ ...
Factor number selection methods

- **scree-plot**
  - Variance Explained (VarExpl)
  - Core Consistency Diagnostics (CCD)
    - Core consistency diagnostics aided by reconstruction error (CCDaRE)
    - ... 

- **redundant factors**
  - Non-redundant model order selection (NORMO)
  - ... 

- **model fit**
  - Difference in fit (DIFFIT)
  - Fast difference in fit (fastDIFFIT)
  - Numerical convex hull (NCH)
  - ... 

- **eigenvalues**
  - Minimal description length (MDL)
  - Modified eigenvalue estimation for Tucker rank determination (MEET)
  - Sparse core method (SCORE)
  - ... 

- **Bayesian statistics**
  - Automatic relevance determination (ARD)
  - ... 

Zuzana Roštáková, Roman Rosipal
Factor number selection in EEG data
MEASUREMENT 2021 2 / 17
Core consistency diagnostics (CCD)

- [Bro and Kiers, 2003]
- Parallel factor analysis (PARAFAC) models with $f = 1, \ldots, F_{\text{max}}$ factors

$$CCD(f) \in (-\infty, 100]$$

- an "elbow"/rapid change in the CCD decreasing profile
Non-redundant model order selection (NORMO)

- [Fernandes et al., 2020]
- PARAFAC models with $f = 1, \ldots, F_{\text{max}}$ factors

\[
NORMO(f) = \text{number of redundant factors} \rightarrow \text{correlation} > 0.7
\]

- $F \rightarrow$ the largest $f$ with $NORMO(f) = 0$ and $NORMO(f + 1) > 0$
- $NORMO_E/NORMO_B$ - all/selected PARAFAC models are considered
Numerical convex hull (NCH)

- [Ceulemans and Kiers, 2006]
- PARAFAC models with $f = 1, \ldots, F_{\text{max}}$ factors
- maximal change in fit between models with a consecutive number of factors
  → only models, which fit belongs to the fit convex hull boundary, are considered
Minimal description length (MDL)

- [Liu et al., 2016]
- Eigenvalues of matricized versions of the tensor
  \[ \Rightarrow \] "indirect" assumption of factors orthogonality
- \( F \in \arg \min_{f \in \{2, \ldots, F_{\text{max}}\}} \text{MDL}(f) \)
Automatic relevance determination (ARD)

- [Mørup and Hansen, 2009]
- factor elements are assumed to follow
  - normal distribution → ridge version of ARD (ARD$_R$)
  - Laplacian distribution → sparse version of ARD (ARD$_S$)

iterative algorithm:

- $k^{th}$ step: PARAFAC model with $F_k$ factors is estimated
  → factors with lowest weights are pruned out
  ⇒ number of factors decreases to $F_{k+1} \leq F_k$

- modified selection criterion
Simulated EEG data

\[ S_{EEG} = C \times S_{source\ signal} \]

64 × T
EEG signal

64 × 2004
forward model

2004 × T (number of time points)
oscillations + broadband brain activity

fractional Brownian motion with the Hurt exponent H = 0.6
Simulated EEG data

- $N_0$ - only 4 oscillations
  5 Hz, 8 Hz, 11 Hz, 14 Hz

- $NG_{low}$ - $N_0$ + gaussian noise $\mathcal{N}(0, 200)$

- $NG_{high}$ - $N_0$ + gaussian noise $\mathcal{N}(0, 1000)$

- $NBBA_{low}$ - 4 oscillations and "low" BBA

- $NBBA_{high}$ - 4 oscillations + "high" BBA

→ 20 sets for each type of data
Results
Mission 1 - $N_0$ data

| CCD | NORMO$_E$, NORMO$_B$ | NCH | MDL | ARD$_R$, ARD$_S$ |

![Chart showing the number of datasets and factors for different methods.](chart.png)

Zuzana Rošťáková, Roman Rosipal  
Factor number selection in EEG data  
MEASUREMENT 2021  
11 / 17
Mission 1 - $N_0$ data

- CCD
- NORMO$_E$, NORMO$_B$
- NCH
- MDL
- ARD$_R$, ARD$_S$

![Bar chart showing number of datasets for different numbers of factors.](image)

- Number of datasets for each method (CCD, NORMO$_E$, NORMO$_B$, NCH, MDL, ARD$_R$, ARD$_S$) as a function of the number of factors.

Zuzana Rošťáková, Roman Rosipal

Factor number selection in EEG data

MEASUREMENT 2021
Mission 2 - $NG_{low}$ data

<table>
<thead>
<tr>
<th>Number of Factors</th>
<th>Number of Datasets</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>CCD</td>
</tr>
<tr>
<td>5</td>
<td>NORMO$_E$, NORMO$_B$</td>
</tr>
<tr>
<td>10</td>
<td>NCH</td>
</tr>
<tr>
<td>15</td>
<td>MDL</td>
</tr>
<tr>
<td>20</td>
<td>ARD$_R$, ARD$_S$</td>
</tr>
</tbody>
</table>

The graph above shows the number of datasets for different numbers of factors. The datasets are categorized as CCD, NORMO$_E$, NORMO$_B$, NCH, MDL, ARD$_R$, and ARD$_S$. The x-axis represents the number of factors, and the y-axis represents the number of datasets.
Mission 2 - $NG_{\text{low}}$ data

CCD  NORMO$_E$, NORMO$_B$  NCH  MDL  ARD$_R$, ARD$_S$
Mission 2 - $NG_{low}$ data

- CCD
- NORMO$_E$, NORMO$_B$
- NCH
- MDL
- ARD$_R$, ARD$_S$

The graph shows the MDL criterion as a function of the number of factors for various datasets: CCD, NORMO$_E$, NORMO$_B$, NCH, MDL, ARD$_R$, ARD$_S$. The x-axis represents the number of factors, and the y-axis represents the MDL criterion in units of $10^4$. The graph illustrates how the MDL criterion decreases as the number of factors increases for each dataset.
Mission 2 - $NG_{low}$ data

<table>
<thead>
<tr>
<th>Number of Factors</th>
<th>CCD</th>
<th>NORMO&lt;sub&gt;E&lt;/sub&gt;, NORMO&lt;sub&gt;B&lt;/sub&gt;</th>
<th>NCH</th>
<th>MDL</th>
<th>ARD&lt;sub&gt;R&lt;/sub&gt;, ARD&lt;sub&gt;S&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

![Bar chart showing the number of datasets for different number of factors and methods.](chart.png)

Zuzana Roštáková, Roman Rosipal | Factor number selection in EEG data | MEASUREMENT 2021
Mission 3 - $NG_{high}$ data

Number of datasets:
- CCD
- NORMO\(_E\), NORMO\(_B\)
- NCH
- MDL
- ARD\(_R\), ARD\(_S\)

Number of factors:
- 0
- 5
- 10
- 15
- 20

Graph showing the number of datasets for each method across different numbers of factors.
Mission 3 - $N_{G_{\text{high}}}$ data

- CCD
- NORMO$_E$, NORMO$_B$
- NCH
- MDL
- ARD$_R$, ARD$_S$

### Graph

- X-axis: Number of factors
- Y-axis: Number of redundant factors

#### Legend
- CCD
- NORMO$_E$
- NORMO$_B$
- NCH
- MDL
- ARD$_R$
- ARD$_S$
Mission 3 - $NG_{high}$ data

CCD  NORMO$_E$, NORMO$_B$  NCH  MDL  ARD$_R$, ARD$_S$

Number of datasets:
- CCD
- NORMO$_E$
- NORMO$_B$
- NCH
- MDL
- ARD$_R$
- ARD$_S$

Number of factors:
- NaN
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10

Graph showing the number of datasets for different methods at various numbers of factors.
Mission 3 - $NG_{\text{high}}$ data

CCD, NORMO_E, NORMO_B, NCH, MDL, ARD_R, ARD_S

Zuzana Roštáková, Roman Rosipal

Factor number selection in EEG data
Mission 4 - $NBBA_{low}$ data

- CCD
- $NORMO_E$, $NORMO_B$
- NCH
- MDL
- $ARD_R$, $ARD_S$

The diagram shows the number of datasets for different methods as a function of the number of factors. The methods include CCD, $NORMO_E$, $NORMO_B$, NCH, MDL, $ARD_R$, and $ARD_S$. The y-axis represents the number of datasets, and the x-axis represents the number of factors.

The methods are color-coded as follows:
- CCD (gray)
- $NORMO_E$ (blue)
- $NORMO_B$ (teal)
- NCH (green)
- MDL (black)
- $ARD_R$ (yellow)
- $ARD_S$ (orange)

The graph indicates the distribution of datasets across different numbers of factors for each method.
Mission 4 - $NBBA_{low}$ data

- CCD
- NORMO$_E$, NORMO$_B$
- NCH
- MDL
- ARD$_R$, ARD$_S$
Mission 4 - $NBBA_{low}$ data

- CCD
- $\text{NORMO}_E$, $\text{NORMO}_B$
- NCH
- MDL
- ARD$_R$, ARD$_S$

The diagram shows the number of datasets for different numbers of factors. The number of datasets decreases as the number of factors increases, with specific colors for each dataset type.
Mission 4 - $NBBA_{low}$ data

<table>
<thead>
<tr>
<th>Number of Factors</th>
<th>CCD</th>
<th>NORMO$_E$, NORMO$_B$</th>
<th>NCH</th>
<th>MDL</th>
<th>ARD$_R$, ARD$_S$</th>
</tr>
</thead>
<tbody>
<tr>
<td>NaN</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>0</td>
<td>5</td>
<td>10</td>
<td>15</td>
<td>20</td>
<td></td>
</tr>
</tbody>
</table>

Number of datasets:
- CCD
- NORMO$_E$
- NORMO$_B$
- NCH
- MDL
- ARD$_R$
- ARD$_S$

Zuzana Roštáková, Roman Rosipal
Factor number selection in EEG data
MEASUREMENT 2021
Mission 5 - \( NBBA_{\text{high}} \) data

<table>
<thead>
<tr>
<th>Number of Factors</th>
<th>Number of Datasets</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>NaN</td>
</tr>
<tr>
<td>5</td>
<td>CCD</td>
</tr>
<tr>
<td>10</td>
<td>NORMO_E, NORMO_B</td>
</tr>
<tr>
<td>15</td>
<td>NCH</td>
</tr>
<tr>
<td>20</td>
<td>MDL</td>
</tr>
<tr>
<td></td>
<td>ARD_R, ARDS</td>
</tr>
</tbody>
</table>

Zuzana Roštáková, Roman Rosipal

Factor number selection in EEG data

MEASUREMENT 2021 16 / 17
Conclusion: Mission impossible?

⇒ none method produced satisfactory results for PARAFAC decomposition of simulated EEG data

- NORMO, MDL - problems already for noiseless data

- CCD - no visible change-point in CCD due to the gaussian noise/BBA

- NCH - data with BBA $\rightarrow F \approx$ minimal allowed value 2

- ARD - the most promising method $\rightarrow$ modification in the future

$\rightarrow$ new approaches are needed
A new efficient method for determining the number of components in PARAFAC models.

Selecting among three-mode principal component models of different types and complexities: A numerical convex hull based method.
*British Journal of Mathematical and Statistical Psychology, 59*(1):133–150.

NORMO: A new method for estimating the number of components in CP tensor decomposition.
*Engineering Applications of Artificial Intelligence, 96.*

Foundations of the PARAFAC procedure: Models and conditions for an “explanatory” multimodal factor analysis.

Detection of number of components in CANDECOMP/PARAFAC models via minimum description length.

Automatic relevance determination for multi-way models.

A comparison of non-negative Tucker decomposition and Parallel Factor Analysis for identification and measurement of human EEG rhythms.