INSTITUTE OF MEASUREMENT SCIENCE - SAV ÚM

2021

13th International Conference on Measurement

Smolenice, Slovakia
May 17 - 19, 2021

Proceedings

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, Bratislava, Slovakia
Email: zuzana.rostakova@savba.sk

Abstract. *A number of factors is an essential parameter in the tensor decomposition methods. It significantly influences not only the decomposition quality but also its interpretation. Many approaches and heuristics were proposed for this purpose. However, their performance is usually demonstrated on data with a simplified structure, and therefore they can produce inferior results when applied to more complex real data. In this study, on a generated dataset closely mimicking the nature of a human multichannel electroencephalogram (EEG), we compared the performance of five methods for selecting the number of factors. We identified the best performing method, but not even this method led to sufficiently acceptable results.*

Keywords: Number of Factors, Electroencephalogram, Tensor Decomposition, PARAFAC

Introduction

Tensor decomposition is a powerful tool for detecting hidden structure in higher-order arrays (tensors), for example, in chemometrics, psychometrics, or neurophysiology [1]. An important parameter, which influences the decomposition quality, is the number of hidden factors F. To determine F, we have to apply a suitable method. Several approaches for selecting the number of factors or heuristics with different assumptions and computational complexity were proposed, but none of them became a state-of-the-art approach. Then, which one to use?

This study aims to help answer this question by comparing five methods for selecting F in the parallel factor analysis (PARAFAC) tensor decomposition [2] on artificial data that mimics the multichannel EEG signal character. We have two main reasons for choosing this particular type of data. First, our long-term research focuses on EEG tensor decomposition by PARAFAC. We often face the problem of correctly determining F. Second, existing methods' performance is usually compared on artificially generated data with a simplified structure that does not follow the real data character. For example, the factors are often generated as mutually orthogonal or independent and identically distributed random sample from a normal or uniform distribution [1, 3, 4]. However, EEG data are more complex, and both orthogonality and normality properties miss their neurophysiological interpretation. Consequently, methods used to determine F can produce inferior results when applied to such data.

Data and Methods

Data

We applied an anatomical forward model consisting of 2,004 dipoles placed in gray matter to generate one minute of scalp 64-channel EEG data [5]. The generated signal comprises a broadband brain activity (BBA) and four narrow-band oscillations – 5 Hz oscillation located in the frontal region, 8 Hz and 14 Hz oscillations in the central region and 11 Hz oscillation in the occipital region. The activity (presence) of each oscillation was generated as non-overlapping with any other in time. To simulate BBA, we used a realization of the fractional Brownian motion with the Hurst exponent $H = 0.6$. According to the amplitude of BBA, we define noiseless data (N_0), data with low ($NBBA_{low}$) and high ($NBBA_{high}$) levels of BBA [5]. Moreover, we added

Gaussian noise to each scalp EEG channel of the N_0 data. To mimic the signal-to-noise ratio at the occipital EEG electrodes of the $NBBA_{low}$ and $NBBA_{high}$ data, we considered two levels of the noise variance denoted as NG_{low} and NG_{high}. These data are less in line with the character of a real EEG signal (missing BBA), but follow the theoretical assumptions of methods described below. For each type of data, we generated 20 datasets/realisations. Before applying PARAFAC, the simulated EEG signal was segmented into two-second time windows with 1900 ms of overlap. The oscillatory part of the amplitude spectrum for each time window and each electrode was transformed into a three-way tensor, representing the time-space-frequency modes [5].

Parallel Factor Analysis

The PARAFAC model [2] decomposes a three-way tensor $\mathbb{X} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ into three matrices $A^{(n)} \in \mathbb{R}^{I_n \times F}, n = 1,2,3$, where F represents the number of factors, and follows the equation

$$x_{i_1 i_2 i_3} = \sum_{f=1}^{F} a_{i_1 f}^{(1)} a_{i_2 f}^{(2)} a_{i_3 f}^{(3)} + e_{i_1 i_2 i_3}, \qquad i_n = 1, \ldots, I_n, \quad n = 1,2,3.$$

The tensor $\mathbb{E} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ represents the error term. Similar to our previous studies [5, 6], we assume nonnegativity of $A^{(1)}$ and $A^{(2)}$ together with unimodality of $A^{(3)}$ columns to follow neurophysiology of data and simplify the interpretation of decomposition.

Factor Number Selection Methods

The approaches for factor number selection can be divided into five sets according to their character. By selecting the best performing candidate from each set, we focus on:

i) **core consistency diagnostics (CCD)** [7]: The CCD values are plotted against the number of factors F and a rapid change in the graph is visually detected.
ii) **non-redundant model order estimator (NORMO)** [8]: The method searches for the lowest F with any redundant factor in PARAFAC. But at least one redundancy occurs for $F+1$. The algorithm considers either all F (NORMO_E) or a subset of them (NORMO_B).
iii) **numerical convex hull (NCH)** [3]: The method focuses on the maximal change in fit between models belonging to the fit convex hull boundary.
iv) **minimal description length (MDL)** [4]: The approach analyses eigenvalues of matricized forms of a tensor.
v) **automatic relevance determination (ARD)** [1]: ARD begins with a large F and continuously prune out factors with a small weight in PARAFAC by using Bayesian statistics. The algorithm follows either the sparse (ARD_S) or ridge (ARD_R) version.

To avoid convergence to a local optimum, the PARAFAC algorithm was run five times for each F and the model with the lowest error was chosen.

We set the maximal number of factors F to 10. A method was considered successful if it selected $F = 4$. Moreover, the estimated factors as a by-product of CCD, NORMO or ARD were checked for consistency with the generated oscillations. Free parameters of each method were carefully tuned to achieve the best possible output. We set the threshold $\alpha > 0.9$ for F selection in NCH and MDL. For NORMO, we set $\alpha = 0.7$.

The original ARD does not allow us to apply the unimodality constraint, and therefore we had to modify the algorithm. Moreover, the tolerance tol for pruning out the unnecessary factors, as suggested by authors, could not decrease the maximal allowed $F = 10$. Different tol values led to different F, and without a priori knowledge about the true F (which is always unknown in practice), we could not set optimal tol value. Therefore, we propose the following modification. In each iteration step, we ordered the factors according to their increased norms, and we skipped the first few factors with cumulative normalized norms under 0.1.

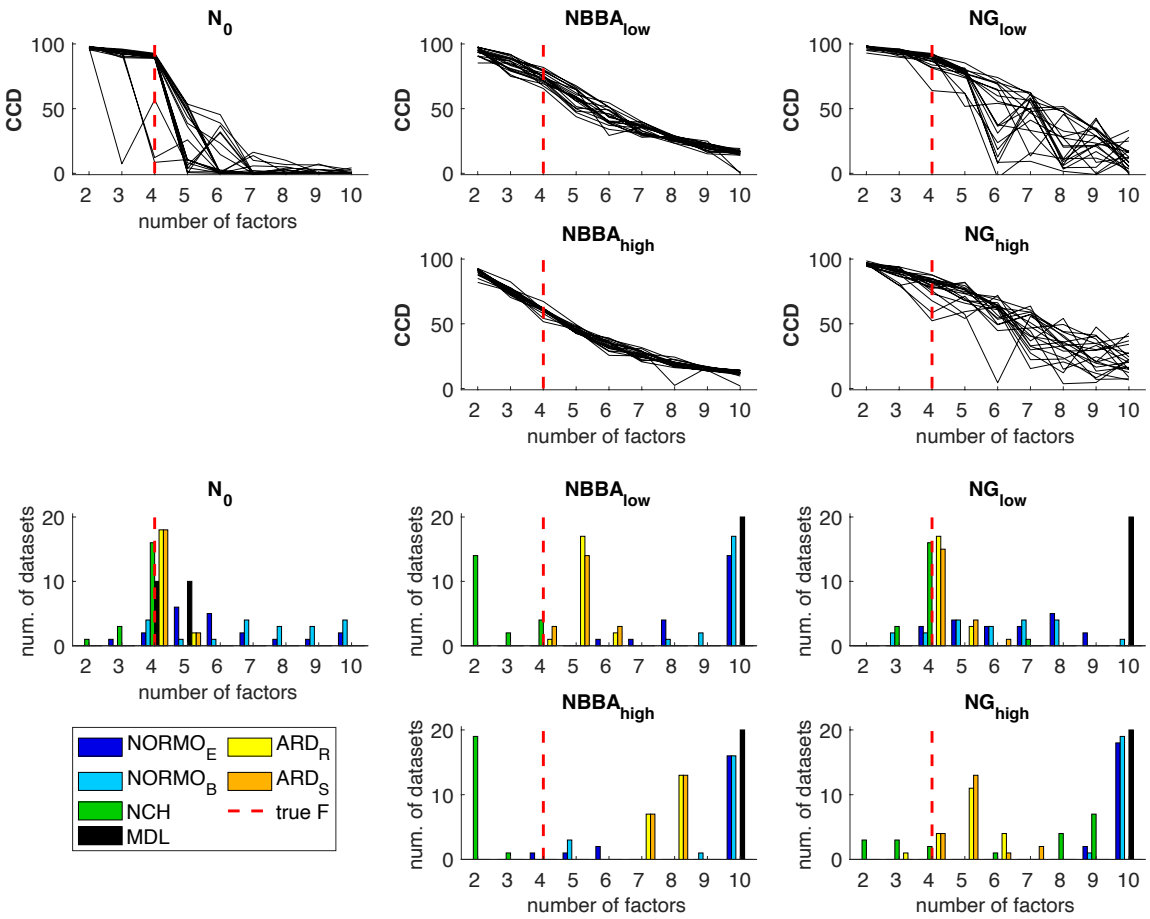

Fig. 1: *First and second row:* The core consistency diagnostics (CCD) of PARAFAC with two to 10 factors. *Third and fourth row:* Histograms of selected factor number in 20 datasets by the non-redundant model order estimator (NORMO_B, NORMO_E), numerical convex hull (NCH), minimal description length (MDL) and automatic relevance determination (ARD_S, ARD_R). Red dashed vertical lines represent the correct factor number $F = 4$. Three columns represent data with no noise (N_0), low and high broadband brain activity ($NBBA_{low}$,$NBBA_{high}$) or Gaussian noise (NG_{low},NG_{high}).

Results and Discussion

A visible rapid drop in CCD (Fig. 1, first and second row) allowed us to select $F = 4$ only for N_0. For NG_{low} data, the number of factors was overestimated ($F = 5$). The CCD values decrease relatively linearly for the other data types, and it was not possible to visually select the best F.

Both NORMO versions detected the correct F for N_0 and NG_{low} in less than one-third of datasets (Fig. 1, dark and light blue). Due to BBA or Gaussian noise, the estimated factors showed numerically weak correlation (< 0.5) despite many physiological similarities detected by visual inspection. Consequently, NORMO resulted in F close to the maximal allowed value 10 for $NBBA_{low}$, $NBBA_{high}$ and NG_{high}.

The correct $F = 4$ was detected by NCH in 16 N_0 and NG_{low} datasets (Fig. 1, green). However, $F = 4$ was chosen in only four cases for $NBBA_{low}$ and two factors were incorrectly chosen in all $NBBA_{high}$ datasets. For NG_{high}, the selected F varied between two and 10.

MDL failed to select the correct F in one-half of N_0 datasets (Figure 1, black). BBA or Gaussian noise's presence forced the F selection to the maximal allowed value equal to 10.

ARD_R and ARD_S performed well for N_0 and NG_{low}, the correct $F = 4$ was selected in at least 15 of 20 datasets (Figure 1, yellow and orange). For $NBBA_{low}$ and NG_{high} the performance had decreased and led to $F = 5$ or $F = 6$. However, only 14 Hz oscillation was present in the estimated five or six factors. The other factors represented higher oscillations or noise. The ARD method was not able to recover the correct number of factors for $NBBA_{high}$.

Conclusions

We compared the performance of five methods for selecting the number of factors in PARAFAC on generated multichannel EEG data. MDL and NORMO methods fail to choose the correct number of factors already for noiseless data. Increasing the level of broadband brain activity (BBA) or Gaussian noise deteriorates the other methods' performance. The best, but still far from ideal, results were obtained by ARD. We can conclude that none of the considered methods provides satisfactory results.

Moreover, we observed inferior results in data with BBA compared to data with Gaussian noise. We hypothesise, that this is due to the methods' assumption of the tensor trilinear structure and presence of Gaussian noise. This assumption is not met in the data with BBA.

In real EEG data, BBA and a measurement noise make detecting narrow-band scalp oscillations harder. Due to the investigated methods' failure to determine correct F in well-controlled generated data, we expect similar sub-optimal performance when applied to real EEG data. In [6], we addressed the problem of the factor number selection differently. We ran PARAFAC models with different F and applied the cluster analysis on obtained decompositions. This allows us to identify the most dominant clusters representing the subject-specific narrow-band scalp EEG oscillations.

Nevertheless, selecting the number of factors in the PARAFAC model remains an open problem, and new approaches are needed.

Acknowledgements

This research was supported by the Slovak Research and Development Agency (grant APVV-16-0202) and by the VEGA grant 2/0081/19.

References

[1] Mørup, M., Hansen, L.K. (2009). Automatic relevance determination for multi-way models. *Journal of Chemometrics*, 23(7-8), 352–363.

[2] Harshman, R.A. (1970). Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multimodal factor analysis. *UCLA Working Papers in Phonetics*, 16, 1–84.

[3] Ceulemans, E., Kiers, H. A.L. (2006). 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.

[4] Liu, K., da Costa, J., So, H.C., Huang, L., Ye, J. (2016). Detection of number of components in CANDECOMP/PARAFAC models via minimum description length. *Digital Signal Processing*, 51, 110–123.

[5] Rosipal, R., Rošťáková, Z., Trejo, L.J. (2021). Tensor decomposition of human oscillatory EEG activity in frequency, space and time. *PsyArXiv*.

[6] Rošťáková, Z., Rosipal, R., Seifpour, S., Trejo, L.J. (2020). A comparison of non-negative Tucker decomposition and Parallel Factor Analysis for identification and measurement of human EEG rhythms. *Measurement Science Review*, 20(3), 126–138.

[7] Bro, R., Kiers, H. A.L. (2003). A new efficient method for determining the number of components in PARAFAC models. *Journal of Chemometrics*, 17(5), 274–286.

[8] Fernandes, S., Fanaee-T, H., Gama, J. (2020). NORMO: A new method for estimating the number of components in CP tensor decomposition. *Engineering Applications of Artificial Intelligence* 96.