

Blind fMRI Source Unmixing via Higher-Order Tensor Decompositions

Christos Chatzichristos^{a,b,*}, Eleftherios Kofidis^{a,c}, Manuel Morante^{a,b}, Sergios Theodoridis^{a,b,d}

^aComputer Technology Institute & Press “Diophantus” (CTI), Greece

^bDept. of Informatics and Telecommunications, National and Kapodistrian University of Athens, Greece

^cDept. of Statistics and Insurance Science, University of Piraeus, Greece

^dChinese University of Hong Kong, Shenzhen, China

Abstract

-Background: The growing interest in neuroimaging technologies generates a massive amount of biomedical data of high dimensionality. Tensor-based analysis of brain imaging data has been recognized as an effective analysis that exploits its inherent multi-way nature. In particular, the advantages of tensorial over matrix-based methods have previously been demonstrated in the context of functional magnetic resonance imaging (fMRI) source localization. However, such methods can also become ineffective in realistic challenging scenarios, involving, e.g., strong noise and/or significant overlap among the activated regions. Moreover, they commonly rely on the assumption of an underlying multilinear model generating the data.

-New Method: This paper aims at investigating the possible gains from exploiting the 4-dimensional nature of the brain images, through a higher-order tensorization of the fMRI signal, and the use of less restrictive generative models. In this context, the higher-order Block Term Decomposition (BTD) and the PARAFAC2 tensor models are considered for the first time in fMRI blind source separation. A novel PARAFAC2-like extension of BTD (BTD2) is also proposed, aiming at combining the effectiveness of BTD in handling strong instances of noise and the potential of PARAFAC2 to cope with datasets that do not follow the strict multilinear assumption.

-Comparison with Existing Methods: The methods were tested using both synthetic and real data and compared with state of the art methods.

-Conclusions: The simulation results demonstrate the effectiveness of BTD and BTD2 for challenging scenarios (presence of noise, spatial overlap among activation regions and inter-subject variability in the Haemodynamic Response Function (HRF)).

Keywords: fMRI, tensors, PARAFAC2, Block Term Decomposition (BTD), BTD2

1. Introduction

Functional Magnetic Resonance Imaging (fMRI) is a noninvasive brain imaging technique, which indirectly studies brain activity, by measuring fluctuations of the Blood Oxygenation Level Dependent (BOLD) signal [1]. BOLD fluctuation usually occurs between 3 to 10 seconds after the stimulus, and this effect is modeled by the so-called Haemodynamic Response Function (HRF). During an fMRI experiment and while the subject performs a set of

tasks responding to external stimuli (task-related fMRI) or no tasks (resting-state fMRI), a series of 3-D brain images is acquired. The localization of the activated brain areas is a challenging Blind Source Separation (BSS) problem [2], in which the sources consist of a combination of spatial maps (areas activated) and time-courses (timings of activation). fMRI data involve multiple modes, such as trial, session and subject, in addition to the intrinsic modes of time and space [3]. Up to now, multivariate bi-linear (i.e., matrix-based) methods, based on the concatenation of different modes, have been the state of the art in fMRI BSS [4, 5]. However, by definition, such methods fall short in exploiting the inherently multi-way nature of fMRI data.

On the other hand, the multi-way nature of the

*Corresponding author at: National and Kapodistrian University of Athens, Department of Informatics and Telecommunications, Panepistimiopolis, Ilisia, 15784, Athens, Greece. Room I3, Tel: +302107275104.

Email address: chatzichris@cti.gr, chrchat@hotmail.com (Christos Chatzichristos)

data is preserved in multi-linear (tensor) models, which, in general, a) produce unique (modulo scaling and permutation ambiguities) representations under mild conditions [6], b) can improve the ability of extracting spatiotemporal modes of interest [3, 7, 8], and c) facilitate neurophysiologically meaningful interpretations [3]. The state-of-the-art in tensorial methods for analyzing multi-subject fMRI data include the Canonical Polyadic Decomposition (CPD)-based analysis [3] and the Tensor Probabilistic Independent Component Analysis (TPICA) [9]. Both methods view the multi-subject fMRI data as a 3rd-order tensor, namely as space \times time \times subjects. Although the subjects mode corresponds to a third dimension, the methods share an initial step of unfolding each of the original 3-D images of fMRI, one per time instance, into a single vector. However, with this type of unfolding, inherited from the matrix-based methods, the intrinsic geometry of the original problem is neglected and not taken into full consideration. In contrast, in this paper, our intention is to exploit the fact that every single brain image (per time point) is a 3-D tensor; to this end, we will resort to higher- (than three) order tensor models, in the multi-subject case.

Furthermore, both CPD and TPICA assume that the data obey a multilinear model; in other words, the underlying signal sources (spatial maps and time-courses) are the same for the different subjects, up to a scaling. Of course, this presupposes that physiological artifacts, which are not likely to satisfy this assumption, have been removed prior to the analysis, and, also, that different subjects share the same HRF. The latter assumption, of a global HRF, is certainly not a valid one, since intra-subject and inter-subject variability is known to exist [10]. Hence, more flexible models, which can accommodate such variations, need to be considered.

In order to better exploit the spatial information that resides in the available data, our kick-off point will be to bypass the initial step of unfolding the 3-D brain images into vectors. Furthermore, the Block-Term Decomposition (BTD) model [11, 12, 13] will be adopted, for the first time in fMRI BSS, in view of its higher modeling potential and its reported robustness to noise. PARAFAC2 [14], which is a model appropriate for multi-way data that do not admit a perfect multilinear representation and allows one of the modes to vary, will be also studied. In addition to its wide use in chemometrics [15], PARAFAC2 has only been recently adopted in fusing electroencephalography (EEG) and fMRI [16]

and, also, has been shown to be effective in the analysis of the functional connectivity in resting-state fMRI [17].

These two tensorial methods (BTD and PARAFAC2) will be tested via extensive simulations with respect to their potential to overcome drawbacks of the state-of-the-art techniques, through improving upon the accuracy of their decomposition results. This will be especially demonstrated in challenging scenarios that involve high noise level and variations in the HRFs of the subjects, respectively. A new method, called BTD2, which extends the rationale behind the PARAFAC2 to the BTD setup so that the latter be able to cope with data that do not comply with a (strict) multilinear representation, is also proposed. BTD2 is a novel tensor model and could be potentially applied in a broader range of applications such as hyperspectral imaging [18] or multi-set factor analysis [19] or other biomedical applications, such as analysis of EEG [20] and electrocardiography (ECG) [21].

The main contributions of the paper are: (a) the adoption of a higher-order unfolding of the spatial domain of the brain, and the use of a more flexible model, namely BTD; (b) the use of non-multilinear models, like PARAFAC2, in task-related fMRI, and (c) the introduction of BTD2, a new tensor model, that combines the advantages of (a) and (b). A heuristic for the estimation of the rank of the block terms in BTD is also proposed. Preliminary shorter versions of the work presented here have previously appeared in [22] and [23].

1.1. Notation

Vectors, matrices and higher-order tensors are denoted by bold lower-case, upper-case and calligraphic upper-case letters, respectively. For a matrix \mathbf{A} , \mathbf{A}^\top and \mathbf{A}^\dagger denote its transpose and pseudo-inverse, respectively. An entry of a vector \mathbf{a} , a matrix \mathbf{A} , or a (3rd-order) tensor \mathcal{A} is denoted by a_i , $a_{i,j}$, or $a_{i,j,k}$, respectively. Matlab notation is used to denote a column of a matrix \mathbf{A} , namely $\mathbf{A}(:, j)$ is its j th column. \mathbf{I}_m is the m th-order identity matrix and $\mathbf{1}_m$ denotes the $m \times 1$ vector of all ones. The symbols \otimes and $*$ denote the Kronecker and the Hadamard (elementwise) products, respectively. The column-wise Khatri–Rao product of two matrices, $\mathbf{A} \in \mathbb{R}^{I \times R}$ and $\mathbf{B} \in \mathbb{R}^{J \times R}$, is denoted by $\mathbf{A} \odot \mathbf{B} = [\mathbf{a}_1 \otimes \mathbf{b}_1, \mathbf{a}_2 \otimes \mathbf{b}_2, \dots, \mathbf{a}_R \otimes \mathbf{b}_R]$, with $\mathbf{a}_j, \mathbf{b}_j$ being the j th columns of \mathbf{A}, \mathbf{B} , respectively. The outer product of two tensors is denoted

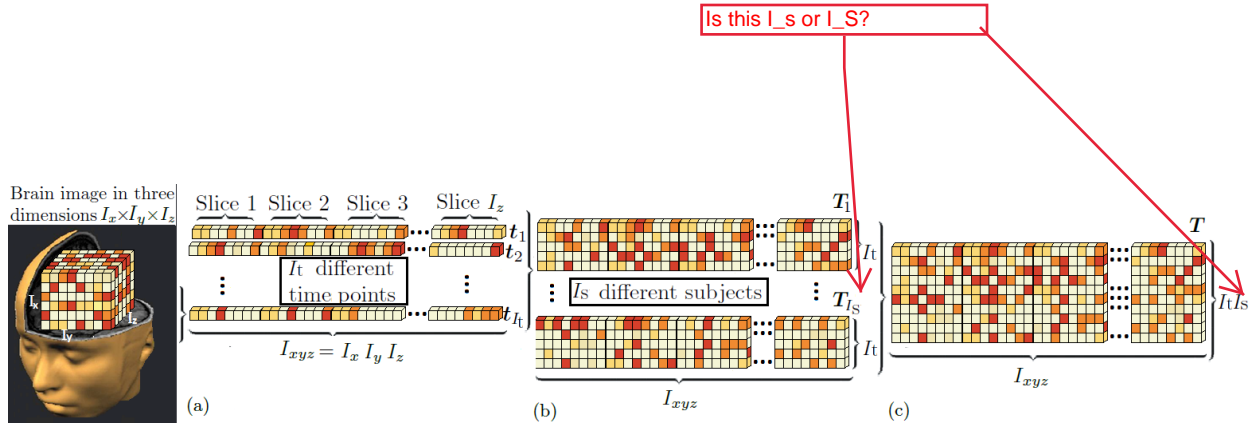


Figure 1: (a) Brain images unfolded in vectors and stacked in matrices (b) per subject and (c) for the multi-subject case.

by \circ . For an N th-order tensor, $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \dots \times I_N}$, $\mathbf{A}_{(n)} \in \mathbb{R}^{I_n \times I_1 I_2 \dots I_{n-1} I_{n+1} \dots I_N}$ is its mode- n unfolded (matricized) version (whose rank is known as mode- n rank), which results from mapping the tensor element with indices (i_1, i_2, \dots, i_N) to a matrix element (i_n, j) , with $j = 1 + \sum_{k=1, k \neq n}^N [(i_k - 1)J_k]$,

$$J_k = \begin{cases} 1, & \text{for } k = 1 \text{ or } k = 2 \text{ and } n = 1, \\ \prod_{m=1, m \neq n}^{k-1} I_m, & \text{otherwise.} \end{cases}$$

The multilinear rank of a tensor is the N -tuple of its N mode- n ranks [11].

2. Methods and Materials

2.1. Tensors and (un)folding of fMRI data

Traditionally, after acquiring a 3-D fMRI image (with spatial dimensions $I_x \times I_y \times I_z$) at a time instance n (Fig. 1), the data (referred to here as *folded* data) is reshaped to a lower dimension (*unfolded*), giving rise to a sequence of vectors, \mathbf{t}_n , for $n = 1, 2, \dots, I_t$ (with $I_{xyz} = I_x \cdot I_y \cdot I_z$ voxels each). These I_t vectors (3-D images at different time instants) are stacked together to form a matrix (Fig. 1(a)). Such a matrix is formed for each one of the subjects, i.e., $\mathbf{T}_k, k = 1, 2, \dots, I_s$ (I_s different subjects in the multi-subject case, Fig. 1(b)). These I_s matrices are in turn concatenated to form $\mathbf{T} \in \mathbb{R}^{I_t I_s \times I_{xyz}}$ (Fig. 1(c)), for which a decomposition is sought such that:

$$\mathbf{T} \approx \mathbf{M} \mathbf{A}^\top, \quad (1)$$

with $\mathbf{A} \in \mathbb{R}^{I_{xyz} \times R}$ containing the weights of the spatial maps and $\mathbf{M} \in \mathbb{R}^{I_t I_s \times R}$ containing the concatenated time-courses of all subjects, R being the estimated number of sources [3, 1]. Note that, in practice, the decomposition cannot be exact due to unmodeled phenomena including noise. In this way, the intrinsically 5th-order (dimension $x \times$ dimension $y \times$ dimension $z \times$ time \times subjects) problem of a multi-subject fMRI analysis has been transformed into a 2nd-order one. This type of unfolding of higher-order data into two-way arrays leads

to decompositions that are non-unique, unless specific assumptions on the involved factors are made. Moreover, and most importantly, such an unfolding can result in a loss of underlying informative correlations that may exist, because the neighborhood information is not respected. In this context, the approaches most frequently pursued are the Independent Component Analysis (ICA) [4, 24], in particular Group ICA (GICA) [25] for the multi-subject case, and Dictionary Learning [5, 26, 27] (or combinations thereof [28]). ICA solves Eq. (1) by assuming that the matrix \mathbf{A} contains statistically independent spatial maps in its columns, each one corresponding to a time-course in the associated column of the (mixing) matrix \mathbf{M} .¹ On the other hand, dictionary learning capitalizes on the assumption of sparsity for the rows of \mathbf{A} . An alternative to GICA for multi-subject cases is Independent Vector Analysis (IVA), which maximizes independence among source signals of each subject represented as random vectors, and dependence among the source signals within the vector (across subjects) [30, 31].

Although research on tensor decompositions has been active for several decades, it is only recently that such methods have attracted a high interest in a large number of applications. The main reason for their increasing popularity is their potential to extract information hidden in the correlations that underlie multidimensional data sets. The tensorial formulation of the data may be suggested either from the nature of the problem under study (e.g., biomedical imaging applications, the 3-D spatial structure of the human body) or due to the specific design of the experiment (e.g., using the data from multiple subjects, which perform the same task). Often, tensor tools can be applied after “tensorizing” the data. [Tensorization transforms the vector or the matrix under consideration to a tensor, via](#)

¹This formulation of ICA in fMRI is called spatial ICA. Temporal ICA, where independence in the time-courses is assumed instead, can be also applied [29].

some kind of folding, or by applying an appropriate transformation (e.g., Wavelet transform, Hankelization, Loewnerization, etc.). Such techniques can be useful when the data itself has a latent low-rank structure, which would then translate into a low-rank tensor [32, 33].

Following the tensorial rationale, in contrast to the previously discussed unfolding, instead of forming a matrix \mathbf{T} by concatenating the matrices $\mathbf{T}_{k=1,2,\dots,I_s}$ in Fig. 1, the latter can be arranged to form a third-order tensor $\mathcal{T} \in \mathbb{R}^{I_{xyz} \times I_t \times I_s}$ and hence a tensor decomposition method can be mobilized for the BSS task [34] of (1). Tensorial methods provide, in most of the cases, improved spatial and temporal localization of the activity, compared to the matrix-based approaches [3, 9]. However, such tensor formulation approaches inherit from their matrix-based counterparts the initial step of the unfolding of the 3-D spatial data into a *vector* \mathbf{t}_n . That is, they do not fully exploit the multi-way nature of the acquired data, which seems to be the natural path to follow for the task at hand.

In addition, the unfolding into vectors misses to fully reveal the low-rank content of the spatial signal, which can only be unveiled through a multi-way model. In fact, the main argument for the use of a matrix (instead of a tensor) model for the single-subject fMRI signal has been that there is no additional low-rank content than that represented ; (1) [35]. The low rankness of the fMRI signal per slice has been used in image reconstruction techniques for compressed sensing applications [36] (usually in the z-axis). Moreover, low-rank constraints (in the spatial domain of fMRI) have been very recently used in a multi-way alternative of the classical General Linear Model (GLM) [37] for monitoring brain responses. Further evidence that motivates the use of low-rank constraints in the spatial domain, per slice, is provided by the Multi-Subject Dictionary Learning (MSDL) probabilistic atlas [38]. Every slice of that atlas consists of 48×48

voxels (and hence has a maximum rank of 48). The maximum slice rank computed among all the Regions Of Interest (ROIs) is 15 (the detailed table with the maximum rank per ROI is presented in the Supplementary Material); hence, all the ROIs are of low rank. Even when different ROIs (of similar function) are combined, the combinations remain of low rank, with maximum rank equal to 17 (combined Language regions). Note that this combination has been considered as an extreme case, since there is not yet evidence that all the Language regions can be activated simultaneously (as one source)²

Furthermore, as it has been shown by Phan et al. [40, 41], unfolding higher-order noisy data to lower-order tensors generally results in a loss of the accuracy in the respective decomposition. The extent of this loss in accuracy depends on the degree of collinearity among the columns of the unfolded mode. Also, as pointed out in [42], the potential of multi-way representations that lead to more robust predictions, compared to their two-way counterparts, seems to increase with the noise level. The use of higher-order tensor models could, therefore, improve the separation performance, both in terms of accuracy and robustness to noise.

A possible way to benefit from the findings mentioned above is to adopt an alternative type of data unfolding. For the unfolding proposed in this paper, we adopt the mode-1 (frontal) matricization of the respective data tensor, \mathbf{A}_n (Fig. 2(a)) (by symmetry, mode-2 or mode-3 can also be used with similar results). By stacking the I_t matrices together, a 3rd-order tensor, $\tilde{\mathcal{T}}_k$ (all the tensors generated by the suggested alternative unfolding will be denoted with a tilde), is formed for the k th subject (Fig. 2(c)). For I_s different subjects, a 4th-order

²For example, Broca's area, which is responsible for the production of language, follows the activation of Wernicke's [39] area, which is responsible for the comprehension of language.

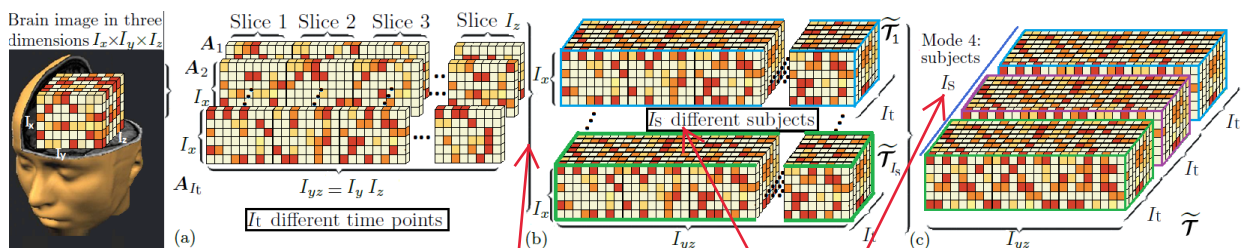


Figure 2: (a) Brain images unfolded in matrices and stacked in (b) 3rd-order tensors per subject and (c) 4th-order tensors for a number of subjects.

Again, is this I_s or I_S ?

what is this brace needed for?

tensor, $\tilde{\mathcal{T}}$, is created, by stacking together all 3-way tensors, $\tilde{\mathcal{T}}_k = \tilde{\mathcal{T}}(:, :, :, k)$. It is of interest to note the close connection of this alternative unfolding with what is called segmentation in the recent tensorization literature [43] and its primary role in translating the BSS problem (1) to the 4-way BTD model of Section 4. A 5th-order tensor (using directly the 3-D spatial brain images) could also be considered, albeit at a complexity increase. In the following sections (2.2–3.1.4), the 4th-order tensor will be considered. The synthetic data used in Sections 3.2–3.4 (obtained from [7, 8, 44]) simulate the brain as a single slice brain, and hence in order to compare directly our results with [7, 8, 44] a 4th-order tensor is considered (since the z dimension of the spatial domain does not exist).³ In the real data section (Section 3.2), tensorizations of both 4th-order and the 5th-order are considered. Yet the results are very similar. It should be noted that all the equations presented in the following sections can be naturally extended to tensors of higher order.⁴

2.2. Models of tensorial fMRI Analysis

2.2.1. Canonical Polyadic Decomposition (CPD)

CPD (or PARAFAC) [45, 3] approximates the 3rd-order tensor of fMRI data, $\mathcal{T} \in \mathbb{R}^{I_{xyz} \times I_t \times I_s}$, by a sum of R (estimated number of sources) rank-1 tensors, namely

$$\mathcal{T} \approx \sum_{r=1}^R \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r. \quad (2)$$

The above can be equivalently written as

$$\mathbf{T}_{(1)} \approx \mathbf{A}(\mathbf{C} \odot \mathbf{B})^\top, \quad (3)$$

and for the k th frontal slice of \mathcal{T} :

$$\mathbf{T}_k \approx \mathbf{A} \mathbf{D}_k \mathbf{B}^\top, \quad k = 1, 2, \dots, I_s, \quad (4)$$

where $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_R]$ is a matrix that contains the R spatial components (I_{xyz} voxels per component) and $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_R]$, $\mathbf{C} = [\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_R]$ are similarly defined matrices, which contain the associated time-courses (I_t time

points) and the subject activation levels (I_s subjects), respectively. \mathbf{D}_k is the diagonal matrix with the elements of the k th row of \mathbf{C} on its diagonal. The main advantage of the CPD, besides its simplicity, is the fact that it is unique (up to permutation and scaling) under mild conditions [46]. Uniqueness of CPD is crucial to its application in fMRI. In fact, it was demonstrated [7, 8] that CPD with fMRI data is robust to overlaps (spatial and/or temporal). On the other hand, the result of CPD is largely dependent on the correct estimation of the tensor rank, R [47, 48].

2.2.2. Tensor Probabilistic Independent Component Analysis (TPICA)

ICA is a powerful tool for separating a multivariate signal into additive components based on the assumption that they are statistically independent. In other words, the assumption of uncorrelated components of the well-known Principal Component Analysis (PCA) is strengthened to statistically independent components. ICA has been demonstrated to provide good results in the characterization of fMRI data [4]. TPICA, as proposed in [9], is essentially a hybrid of the Probabilistic ICA (PICA) [49] method and the CPD method for multi-subject analyses.⁵ Given a 3rd-order tensor of fMRI data, \mathcal{T} , TPICA approximates it as (see (3)):

$$\mathbf{T}_{(1)} \approx \mathbf{A} \mathbf{M}^\top, \quad (5)$$

where the columns of \mathbf{A} are assumed to be samples of independent, non-Gaussian [49] random variables and $\mathbf{M} = \mathbf{C} \odot \mathbf{B}$ is a Khatri–Rao structured mixing matrix (cf. (1)). TPICA decomposes the tensor \mathcal{T} in two steps: an ICA step, which estimates \mathbf{M} and \mathbf{A} , and a Khatri–Rao factorization of \mathbf{M} (using Singular Value Decomposition (SVD)) to determine \mathbf{B} and \mathbf{C} . These two steps are performed iteratively, in an alternating fashion, until convergence. It has been shown in [8, 51] that iterations in TPICA are redundant and, hence, it was proposed that the algorithm be used with only one iteration. TPICA is more robust than CPD to rank estimation errors but it exhibits inferior performance in the presence of overlap among the sources and/or strong noise [7, 8].

³The 5th-order tensorization has been also tested in preliminary simulations not reported here, with cylindrical 3-D brains, with only small gains in the unmixing performance (and of course an increase in the computational cost).

⁴For the rank- $(L, L, L, 1, 1)$ case the two-fold segmentation algorithm, as proposed at [45] has been used.

⁵TPICA has been selected as the main tensorial ICA method, for comparison since it is the state-of-the-art tensorial method used in the fMRI community [8, 7]. In fact, it is the only tensorial method implemented in the FMRIB’s Software Library (FSL [50]), a widely employed tool for fMRI analysis.

2.2.3. Block Term Decomposition (BTD) for fMRI

Following the arguments presented in Section 2.1, an alternative unfolding and a higher than 3rd-order tensor model is a more natural way to perform the unmixing of the sources. However, the use of CPD (and hence TPICA) with such a formulation can be problematic in cases where the components are not of rank one. Let us consider the following example. Assume that only one source of activation exists (i.e. $R = 1$) and CPD is used for the 4-way tensor (Fig. 2(c)), $\tilde{\mathcal{T}}$. The CPD for the 4-D data can then be expressed as [52] (with $R = 1$ in our case):

$$\tilde{\mathcal{T}} \approx \sum_{r=1}^R \mathbf{x}_r \circ \mathbf{y}_r \circ \mathbf{b}_r \circ \mathbf{c}_r. \quad (6)$$

Thus, the spatial map of the source (since now we have two spatial modes) is necessarily expressed as an outer product, $\mathbf{x}_r \circ \mathbf{y}_r$. In cases where the spatial map is indeed of “spatial rank” 1 (we will refer to the rank of the 2-D spatial image as “spatial rank” for simplicity), CPD will succeed in decomposing correctly the data (S_1 source in Fig. 3(a), like the motion perception areas simulated in [8]). However, if the spatial activation is of higher “spatial rank” (S_2 source, Fig. 3(b)), CPD would result in phantoms (resulting from the multiplication of x_a with y_b , x_b with y_a and y_c , etc.). This is because it is impossible to express such a spatial map (image) as an outer product of vectors. An example of a source of the form of S_2 is that of the motion perception areas, which, of course, are not two equally sized perfect rectangles as simulated in [8] but of higher spatial rank, as can be viewed in Fig. 4. Hence, the constraint imposed by CPD (that all terms should be of rank one) can be too restrictive in such cases. On the other hand, it seems less restrictive to decompose the tensor in low multilinear rank terms, which, however, are not necessarily restricted to be of rank one.⁶ This enhances the potential for modeling more general phenomena [53]. As an alternative to CPD, the use of Block Term Decomposition (BTD) [11, 12, 13] in the 4-way (and 5-way) tensorization of the fMRI signal is investigated in this work, for the first time. The adoption of BTD is dictated by the need of a more flexible model that reveals the low rankness of the spatial mode [43].

BTD is a generalization of CPD, which can capture latent factors of rank higher than one in each component. In particular, the rank- $(L_r, L_r, 1)$ BTD

⁶Note that a rank-1 tensor can be seen as a multilinear rank- $(1, 1, \dots, 1)$ tensor [12].

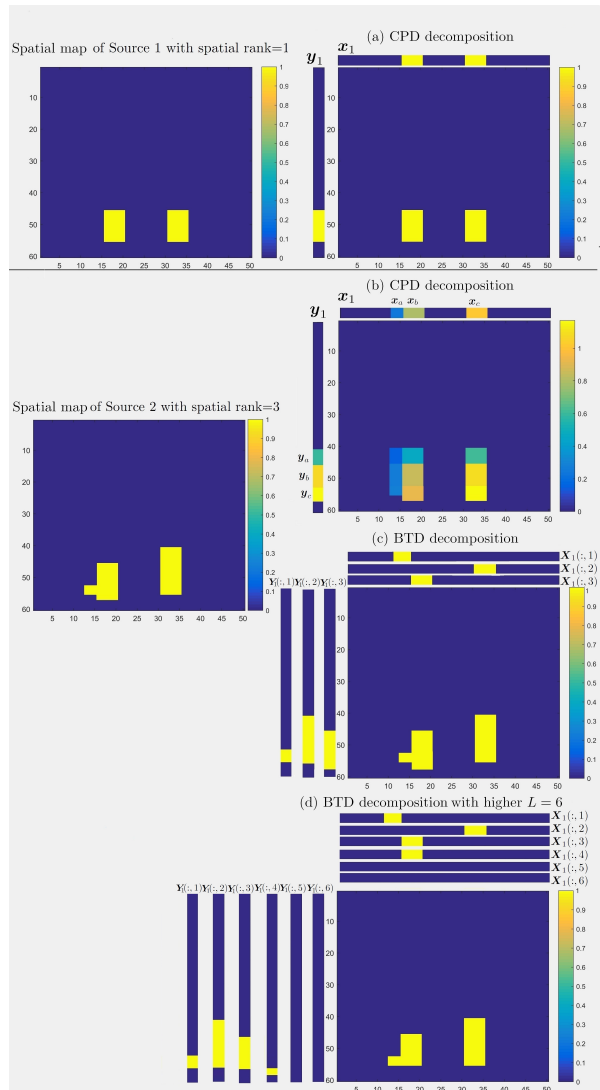


Figure 3: Decomposition of sources S_1 and S_2 .

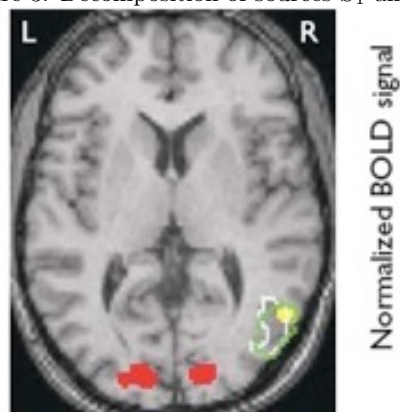


Figure 4: Real motion perception areas of “spatial rank” higher than 1. Picture obtained from [54].

of the tensor $\tilde{\mathcal{T}}_k \in \mathbb{R}^{I_x \times I_{yz} \times I_t}$ (in Fig. 2) is given by

$$\tilde{\mathcal{T}}_k \approx \sum_{r=1}^R \mathbf{A}_r \circ \mathbf{b}_r = \sum_{r=1}^R (\mathbf{X}_r \mathbf{Y}_r^\top) \circ \mathbf{b}_r, \quad (7)$$

$k = 1, 2, \dots, I_s$, where $\mathbf{A}_r = \mathbf{X}_r \mathbf{Y}_r^\top \in \mathbb{R}^{I_x \times I_{yz}}$ and the matrices $\mathbf{X}_r \in \mathbb{R}^{I_x \times L_r}$ and $\mathbf{Y}_r \in \mathbb{R}^{I_{yz} \times L_r}$ are of full column rank. BTD, for the case of a 3rd-order tensor, has been successfully applied in modeling epileptic seizures in EEG [33] and proved capable of modeling non-stationary (in frequency or in space) seizures, better than other techniques. A generalization of BTD has also been used in EEG Motor Imagery Data (MID) [55] and in ECG signal analysis [56]. BTD has not been previously applied in fMRI analysis, to the best of the authors' knowledge. In this work, it is proposed to decompose the 4th-order data (Fig. 2(c)) BTD:

$$\tilde{\mathcal{T}} \approx \sum_{r=1}^R \mathbf{A}_r \circ \mathbf{b}_r \circ \mathbf{c}_r = \sum_{r=1}^R (\mathbf{X}_r \mathbf{Y}_r^\top) \circ \mathbf{b}_r \circ \mathbf{c}_r. \quad (8)$$

The two spatial factors are of low rank (L_r) while the time and subject factors are of rank one (those modes have not been folded and the assumption of rank one is still valid). Thus, in the example of Fig. 3(c), source S_2 could be perfectly represented by a rank-(3,3,1,1) BTD (of a single term). The use of this model offers the ability to estimate sources of "spatial rank" higher than one (which are the cases where CPD fails) correctly, using the proposed unfolding. The rank- $(L_r, L_r, L_r, 1, 1)$ BTD will be employed when the 5th-order tensor is considered. For simplicity, and as it is the standard practice in the BTD applications, we choose all L_r 's in the following to be equal (to L).

Estimating L is an open problem in the BTD literature. Greedy algorithms, like [57], could be used. Furthermore, model order estimation techniques, dictated from corresponding CPD approaches have been proposed [33] as well as hybrids, depending on the application at hand [58]. In the case of fMRI, and since L is connected with the "spatial rank" of every slice, brain atlases can be used to provide indications for the rank of the areas under consideration. The result of the decomposition in most of the BTD applications, as reported in the literature, is relatively insensitive to overestimation of L (the relative robustness to the

choice of L has been also demonstrated in [59]). Of course, higher L values result in increased complexity. A compromise between accuracy and complexity must be made. In order to obtain an indication of the appropriate L (basically select a value close to the maximum "spatial rank" of the sources) for the BSS problem of fMRI, a heuristic is proposed here. When L is set higher than the actual "spatial rank" (e.g., $L = 6$ in Fig. 3(d) instead of 3 in Fig. 3(c)), some column pairs of \mathbf{X}_r and \mathbf{Y}_r do not contain any signal information (e.g., $\mathbf{X}_1(:, 5), \mathbf{Y}_1(:, 5)$ and $\mathbf{X}_1(:, 6), \mathbf{Y}_1(:, 6)$ in Fig. 3(d)) and some splits occur (e.g., $\mathbf{Y}_1(:, 3)$ of Fig. 3(c) into $\mathbf{Y}_1(:, 3)$ and $\mathbf{Y}_1(:, 4)$ of Fig. 3(d)).

This procedure (stated bellow as Heuristic 1) aims at identifying the pairs of columns which do not contain any useful signal information and decreases the L_r value accordingly. After computing

$\phi \leftarrow$ probability density function of normal distribution
 $(:)$ \leftarrow vectorization

Input : A 4th-order tensor $\tilde{\mathcal{T}} \in \mathbb{R}^{I_x \times I_{yz} \times I_t \times I_s}$, α (usually set equal to 0.05), L_{init} , and R

Output: L

- 1 Set L_{init} (overestimate)
- 2 $\tilde{\mathcal{T}} = \sum_{r=1}^R (\mathbf{X}_r \mathbf{Y}_r^\top) \circ \mathbf{b}_r \circ \mathbf{c}_r$ (BTD)
- 3 $p = \alpha / (I_x \cdot I_{yz})$ (Bonferonni corrected p-value)
- 4 $Zn = \phi^{-1}(p)$ (Compute Z threshold)
- 5 **for** $r \leftarrow 1$ **to** R **do**
- 6 **for** $j \leftarrow 1$ **to** L_{init} **do**
- 7 $\mathbf{Im} = \mathbf{X}_r(:, j) \mathbf{Y}_r(:, j)^\top$
- 8 $\mathbf{Q}(:, j) = \mathbf{Im}(:, :)$
- 9 **end**
- 10 $\mu = \text{mean}(\mathbf{Q}(:, :))$
- 11 $\sigma = \text{std}(\mathbf{Q}(:, :))$
- 12 $L_r = 0$
- 13 **for** $j \leftarrow 1$ **to** L_{init} **do**
- 14 $i = 1$
- 15 **do**
- 16 $z_{i,j} = (q_{i,j} - \mu) / \sigma$ (Compute z-stat)
- 17 **if** $z_{i,j} < Zn$ **then**
- 18 $L_r = L_r + 1$
- 19 $i = I_x \cdot I_{yz}$
- 20 **end**
- 21 $i = i + 1$
- 22 **while** $i < I_x \cdot I_{yz}$
- 23 **end**
- 24 **end**
- 25 $L = \max(L_r)$

when using a different L_r per source, line 25 is omitted

Heuristic 1: Estimation of L .

the factors, and following the assumption that all the voxels of the sources obtained can be considered independent of each other [7, 9], a Z-test in all the voxels is performed. The new L_r value is set equal to the number of columns which contain

at least one significant voxel (a voxel for which the null hypothesis is rejected) after applying the Bonferroni correction in order to correct the Family Wise Error (FWE) [60]. Bonferroni correction is considered conservative and other correction procedures could be used instead [61]. As a last step, in the case that a single L is used, it is obtained as the maximum of the L_r .

Regarding **the uniqueness of BTD**, it was proved in [11] that the decomposition in (7) is (essentially) unique, provided that the matrices $[\mathbf{X}_1 \ \mathbf{X}_2 \ \cdots \ \mathbf{X}_R]$ and $[\mathbf{Y}_1 \ \mathbf{Y}_2 \ \cdots \ \mathbf{Y}_R]$ are full column rank and the matrix $\mathbf{B} = [\mathbf{b}_1 \ \mathbf{b}_2 \ \cdots \ \mathbf{b}_R]$ does not contain collinear columns, up to the following indeterminacies: a) scaling and permutation, as in CPD, and b) the simultaneous post-multiplication of \mathbf{X}_r by a non-singular matrix \mathbf{F} and at the same time the pre-multiplication of \mathbf{Y}_r by its inverse. Note that the latter indeterminacy is of no consequence to our results, since it is the product $\mathbf{A}_r = \mathbf{X}_r \mathbf{Y}_r^\top$ (our spatial map) that is of interest in our study. An argument showing that this type of uniqueness can be extended to the rank- $(L, L, 1, 1)$ case follows.

Proof. As suggested in [6], the uniqueness of a higher-order tensor decomposition can be shown through a reduction to a third-order tensor, which is “the first instance of multilinearity, for which uniqueness holds and from which uniqueness propagates by virtue of Khatri–Rao structure” [6]. Assume that the matrices $[\mathbf{X}_1 \ \mathbf{X}_2 \ \cdots \ \mathbf{X}_R]$ and $[\mathbf{Y}_1 \ \mathbf{Y}_2 \ \cdots \ \mathbf{Y}_R]$ are of full column rank and the matrices $\mathbf{B}, \mathbf{C} = [\mathbf{c}_1 \ \mathbf{c}_2 \ \cdots \ \mathbf{c}_R]$ do not contain collinear or null columns (a realistic assumption for matrices that represent time and subjects, assuming that the correct R has been selected). In view of the above, uniqueness for (8) can be proved via the uniqueness of a 3-way counterpart of $\tilde{\mathcal{T}}, \tilde{\mathcal{T}}_{(1,2)} \in \mathbb{R}^{I_x \times I_{yz} \times I_s I_t}$, created by the concatenation of the third and fourth modes, such that the fourth mode is nested into the third one:

$$\tilde{\mathcal{T}}_{(1,2)} = \sum_{r=1}^R \mathbf{A}_r \circ \mathbf{g}_r, \quad (9)$$

where $\mathbf{G} = [\mathbf{g}_1 \ \mathbf{g}_2 \ \cdots \ \mathbf{g}_R] = \mathbf{B} \odot \mathbf{C}$ is the Khatri–Rao product of two matrices of neither null nor collinear columns, and hence it does not involve null or collinear columns either [62, Proposition 1]. Following Theorem 4.1 of [11],

since the matrices $[\mathbf{X}_1 \ \mathbf{X}_2 \ \cdots \ \mathbf{X}_R]$ and $[\mathbf{Y}_1 \ \mathbf{Y}_2 \ \cdots \ \mathbf{Y}_R]$ have full column rank and the matrix \mathbf{G} has no collinear columns, the decomposition is essentially unique. \square

2.3. Non strictly multilinear tensor models

2.3.1. PARAFAC2

PARAFAC2 [14] differs from CPD in that strict multilinearity is no longer a requirement. CPD assumes the *same* factors across *all* the different modes, whereas PARAFAC2 relaxes this constraint and allows variation across one mode (in terms of the values and/or the size of the corresponding factor matrix). For this reason, PARAFAC2 is not a tensor decomposition model in the strict sense, as it can represent both regular tensors, with weaker constraints than CPD, as well as irregular tensors (collections of matrices of different dimensions) with size variations along one of the modes (Fig. 5). It can be written in terms of the (here frontal) slices of the permuted tensor $\mathcal{T}^{(p)} \in \mathbb{R}^{I_t \times I_{xyz} \times I_s}$ (the tensor is permuted because the multilinearity constraint will be relaxed in the time domain per subject, so as to account for the inter-subject variability of the HRF) as:

$$\mathbf{T}_k^{(p)} \approx \mathbf{B}_k \mathbf{D}_k \mathbf{A}^\top, \quad k = 1, 2, \dots, I_t, \quad (10)$$

and allowing \mathbf{B}_k to be different for different k 's (\mathbf{D}_k and \mathbf{A} are matrices defined as in (4)). This type of decomposition is clearly non unique [14]. Thus, in order to allow for uniqueness, it has been proposed to add the constraint that the cross products $\mathbf{B}_k^\top \mathbf{B}_k$ be constant over k , a much more relaxed constraint than the requirement for equal \mathbf{B}_k 's in CPD. This has been shown [18] to be equivalent with $\mathbf{B}_k = \mathbf{P}_k \mathbf{H}$, where the $R \times R$ matrix \mathbf{H} is the same for all slices, while the variability is represented by the columnwise orthonormal $I_t \times R$ matrix \mathbf{P}_k . Under this constraint, one has to fit the equivalent model

$$\mathbf{P}_k^\top \mathbf{T}_k^{(p)} \approx \mathbf{H} \mathbf{D}_k \mathbf{A}^\top, \quad k = 1, 2, \dots, I_s. \quad (11)$$

As shown in [18], \mathbf{P}_k can be computed as $\mathbf{P}_k = \mathbf{V}_k \mathbf{U}_k^\top$, where \mathbf{U}_k and \mathbf{V}_k are the left and right singular matrices of $\mathbf{H} \mathbf{D}_k \mathbf{A}^\top \mathbf{T}_k^{(p)\top}$. As can be seen from Eq. (11), the problem of fitting PARAFAC2 has been transformed into that of fitting a CPD model to *transformed* data.

Applications of PARAFAC2 in fMRI analysis include [16] and [17]. The former study is concerned with joint EEG-fMRI analysis in single-subject cases, with the different modes being time-courses, spatial maps per slice, and slices, and allowing the mode that represents the spatial maps to vary over slices. In [17], the modes considered are time, subjects and space, as in this paper. In that work, PARAFAC2 was evaluated along with other models w.r.t. its effectiveness in capturing the brain’s functional connectivity.

2.3.2. Block Term Decomposition 2 (BTD2)

As described in Section 2.2.3, the use of the alternative unfolding, which has been proposed (Fig. 2) or the consideration of the 3-D brain image (without the application of any unfolding) dictates the use of a more flexible model than CPD. Similarly to CPD, PARAFAC2 is not an appropriate model, again due to the rank-1 assumption. Hence, a BTD analog of PARAFAC2 is proposed here aiming at the combination of the advantages of BTD with the simultaneous use of higher-order tensors (robustness to the noise) and of PARAFAC2 (better performance in non strictly multilinear datasets with realistic scenarios, involving different time-courses per subject due to inter-subject HRF variability).

Using the (mode-1) unfoldings, $\tilde{\mathbf{T}}_{k(1)}^{(p)}$, of the $\tilde{\mathcal{T}}_k^{(p)} = \tilde{\mathcal{T}}^{(p)}(:, :, :, k) \in \mathbb{R}^{I_t \times I_x \times I_{yz}}$ tensors (Fig. 2) yields the BTD2 decomposition:

$$\tilde{\mathbf{T}}_{k(1)}^{(p)} = (\mathbf{B}_k \mathbf{S}) \tilde{\mathbf{D}}_k (\mathbf{X} \odot \mathbf{Y})^\top, k = 1, 2, \dots, I_t. \quad (12)$$

The matrices $\mathbf{S} = \text{blockdiag}(\mathbf{1}_L^\top, \mathbf{1}_L^\top, \dots, \mathbf{1}_L^\top)$ and $\tilde{\mathbf{D}}_k = \text{blockdiag}(c_{k1} \mathbf{I}_L, \dots, c_{kR} \mathbf{I}_L)$ appear in formulating BTD as CPD, as in [63]. Spatial matrices

are defined as $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_R]$, $\mathbf{Y} = [\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_R]$ with $\mathbf{X}_i \in \mathbb{R}^{I_x \times L}$ and $\mathbf{Y}_i \in \mathbb{R}^{I_{yz} \times L}$. Extending the PARAFAC2 direct fit algorithm from [18] to the above 4-way model and appropriately adapting the Alternating Least Squares (ALS) iterations from CPD to BTD as in [63, 11] results in Algorithm 1.⁷ First, the gradient of the cost function with respect to each component matrix is calculated (following Eq. (12) for all the possible unfoldings). Then, at a second step, the ALS update rules for the proposed cost-function are obtained by setting the calculated gradients to zero (lines 13–19 of Algorithm 1). Matrices $\mathbf{W}_x, \mathbf{W}_y, \mathbf{W}_h$ and \mathbf{W}_c result from the well-known property of the pseudoinverse of Khatri–Rao product of two matrices \mathbf{A} and \mathbf{B} , $(\mathbf{A} \odot \mathbf{B})^\dagger = ((\mathbf{A}^\top \mathbf{A}) * (\mathbf{B}^\top \mathbf{B}))^\dagger (\mathbf{A} \odot \mathbf{B})^\top$ [64].

The bottleneck of the CPD-ALS algorithm (and of its extensions, BTD-ALS and BTD2-ALS) is the computation of the Matricized-Tensor-Times-Khatri–Rao-Product (MTTKRP). For example, in Algorithm 1, it corresponds to $\mathbf{Z}_{(1)}((\mathbf{C}\mathbf{S}) \odot (\mathbf{H}\mathbf{S}) \odot \mathbf{Y})$ when solving for \mathbf{X} (line 13) and similarly in lines 15, 17 and 19 when solving for the other modes. For large tensors, the brute-force computation of the MTTKRP is very costly in terms of both computation and memory. In order to reduce the MTTKRP complexity and fully parallelize the algorithm with respect to the I_s subjects, a MTTKRP kernel can be used, similarly to [65].

Concerning uniqueness of PARAFAC2, concrete results have only been reported for the special cases of rank $R = 2$ or 3 [66, 19]. For the general

⁷Note that the scaling of the columns of two of the factor matrices, commonly performed in CPD ALS to avoid over-/under-flow, is translated in [63, 13] to a column orthonormalization (via QR decomposition).

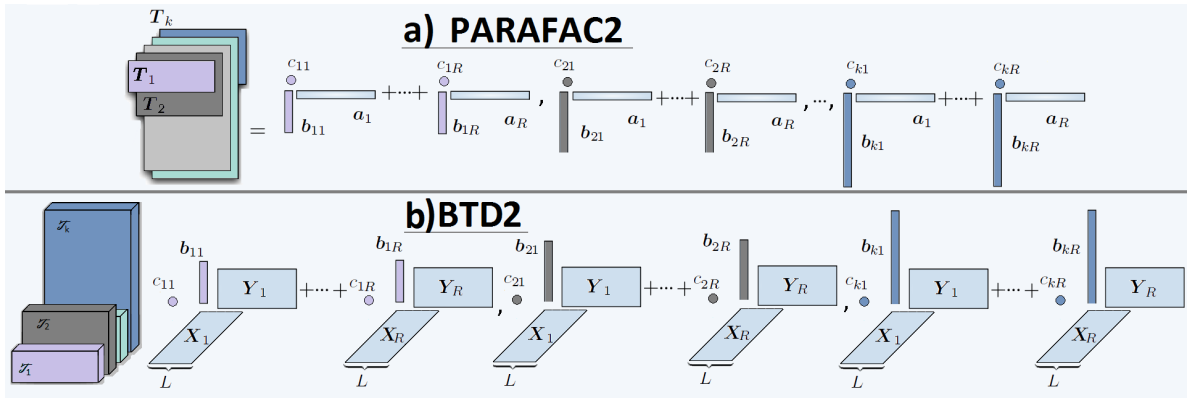


Figure 5: Non strictly multilinear tensor models: PARAFAC2 and BTD2.

case, a sufficient (but not necessary) condition was proved in [67]. All of these conditions can be extended to BT2. For instance, the condition in [67], which is based on the maximum number of unique combinations of four diagonal elements (with possible repetition) of \mathbf{D}_k , can be generalized as $I_t \geq \bar{R}(\bar{R} + 1)(\bar{R} + 2)(\bar{R} + 3)/24$, with $\bar{R} = LR$. Furthermore, it is of interest to observe that the rank- $(L, L, 1, 1)$ BT2 will still be unique, even if no constraints are imposed on \mathbf{B}_k , since the rank- $(L, L, 1, 1)$ unconstrained BT2 is equivalent to a rank- $(L, L, 1)$ BT2 of the 4-way data unfolded to a 3-way tensor (with dimensions $I_x \times I_{yz} \times I_t I_s$, check Fig. 3 of the Supplementary Material). This uniqueness property does not hold for the 3-way PARAFAC2 [15].

Input : A 4th-order tensor $\tilde{\mathcal{T}}^{(p)} \in \mathbb{R}^{I_t \times I_x \times I_{yz} \times I_s}$, R and L

Output: Spatial factors $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_R]$, $\mathbf{Y} = [\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_R]$ with $\mathbf{X}_i \in \mathbb{R}^{I_x \times L}$ and $\mathbf{Y}_i \in \mathbb{R}^{I_{yz} \times L}$, temporal factor $\mathbf{B} \in \mathbb{R}^{I_s \times R}$ and subject factor $\mathbf{C} \in \mathbb{R}^{I_t \times R}$.

```

1 Choose initial values of  $\mathbf{X}$ ,  $\mathbf{Y}$ ,  $\mathbf{H}$  and  $\mathbf{C}$  using
  multi-initialization technique
2  $\mathbf{S} \leftarrow \text{blockdiag}(\mathbf{1}_L^\top, \mathbf{1}_L^\top, \dots, \mathbf{1}_L^\top)$ 
3 repeat
4   for  $k \leftarrow 1$  to  $I_s$  do
5      $\tilde{\mathbf{D}}_k \leftarrow \text{blockdiag}(c_{k1} \mathbf{I}_L, \dots, c_{kR} \mathbf{I}_L)$ 
6     Form  $\tilde{\mathbf{T}}_{k(1)}^{(p)} \in \mathbb{R}^{I_t \times I_x \times I_{yz}}$  (12)
7     SVD  $(\mathbf{H}\mathbf{S})\tilde{\mathbf{D}}_k(\mathbf{X} \odot \mathbf{Y})^\top \tilde{\mathbf{T}}_{k(1)}^{(p)\top} = \mathbf{U}_k \mathbf{\Delta}_k \mathbf{V}_k^\top$ 
8      $\mathbf{P}_k \leftarrow \mathbf{V}_k \mathbf{U}_k^\top$ 
9      $\mathbf{Z}_{k(1)} \leftarrow \mathbf{P}_k^\top \tilde{\mathbf{T}}_{k(1)}^{(p)}$ 
      ( $\mathbf{Z}_k = \mathbf{Z}(:, :, :, k) \in \mathbb{R}^{R \times I_x \times I_{yz}}$ )
10    end
11    repeat
12       $\mathbf{W}_x \leftarrow ((\mathbf{S}^\top \mathbf{C}^\top \mathbf{C}\mathbf{S}) * (\mathbf{S}^\top \mathbf{H}^\top \mathbf{H}\mathbf{S}) * (\mathbf{Y}^\top \mathbf{Y}))$ 
13       $\mathbf{X} \leftarrow \mathbf{Z}_{(1)}((\mathbf{C}\mathbf{S}) \odot (\mathbf{H}\mathbf{S}) \odot \mathbf{Y})\mathbf{W}_x^\dagger$ 
14       $\mathbf{W}_y \leftarrow ((\mathbf{S}^\top \mathbf{C}^\top \mathbf{C}\mathbf{S}) * (\mathbf{S}^\top \mathbf{H}^\top \mathbf{H}\mathbf{S}) * (\mathbf{X}^\top \mathbf{X}))$ 
15       $\mathbf{Y} \leftarrow \mathbf{Z}_{(2)}((\mathbf{C}\mathbf{S}) \odot (\mathbf{H}\mathbf{S}) \odot \mathbf{X})\mathbf{W}_y^\dagger$ 
16       $\mathbf{W}_h \leftarrow \mathbf{S}((\mathbf{S}^\top \mathbf{C}^\top \mathbf{C}\mathbf{S}) * (\mathbf{Y}^\top \mathbf{Y}) * (\mathbf{X}^\top \mathbf{X}))\mathbf{S}^\top$ 
17       $\mathbf{H} \leftarrow \mathbf{Z}_{(3)}((\mathbf{C}\mathbf{S}) \odot \mathbf{Y} \odot \mathbf{X})\mathbf{S}^\top \mathbf{W}_h^\dagger$ 
18       $\mathbf{W}_c \leftarrow \mathbf{S}((\mathbf{S}^\top \mathbf{H}^\top \mathbf{H}\mathbf{S}) * (\mathbf{Y}^\top \mathbf{Y}) * (\mathbf{X}^\top \mathbf{X}))\mathbf{S}^\top$ 
19       $\mathbf{C} \leftarrow \mathbf{Z}_{(4)}((\mathbf{H}\mathbf{S}) \odot \mathbf{Y} \odot \mathbf{X})\mathbf{S}^\top \mathbf{W}_c^\dagger$ 
20    until stopping criterion has been met
21  until stopping criterion has been met
22  for  $k \leftarrow 1$  to  $I_s$  do
23     $\mathbf{B}_k \leftarrow \mathbf{P}_k \mathbf{H}$ 
24  end
```

Algorithm 1: BT2 algorithm for a 4th-order tensor.

2.4. ICA-based methods used in the simulations

Four different methods based on the independence assumption are tested and compared in the

results section. Namely, GICA [25], IVA [30], Sparse ICA (SICA) [28] and Canonical Parallel factor Analysis with independence constraint (ICAPFA) [51].

The main steps involved in GICA and IVA decomposition are as follows (for more details, see [25] and [30]).

GICA:

1. Subject-level PCA.
2. Temporal concatenation of the PCA-reduced subject data (Fig. 1).
3. Group-level PCA.
4. ICA (with Infomax algorithm [68]).
5. Back reconstruction using the spatio-temporal regression (STR) [69] method in order to compute the spatial map and time-course per subject. The subject-specific time-courses ($\check{\mathbf{B}}_k$) and the subject-specific spatial maps ($\check{\mathbf{A}}_k$) are given by $\check{\mathbf{B}}_k^\top = \tilde{\mathbf{T}}_k \mathbf{A}^\dagger$ and $\check{\mathbf{A}}_k = (\check{\mathbf{B}}_k^\top)^\dagger \tilde{\mathbf{T}}_k$ respectively (the same back reconstruction method is used for BT2, with \mathbf{A} computed from (8)).

IVA:

1. Subject-level PCA.
2. Concatenation of the PCA-reduced subject data along the third dimension.
3. IVA with multivariate Laplace prior [30] as implemented in GIFT [70].
4. Back reconstruction is not required for the subject-specific spatial maps since they are provided by IVA. Subject-specific time-courses are given as $\check{\mathbf{B}}_k^\top = \tilde{\mathbf{T}}_k \check{\mathbf{A}}_k^\dagger$.

Sparsity and independence (as mentioned before) are two constraints that have been proved useful in BSS. In [71], a unified mathematical framework that enables the exploitation of both independence and sparsity was introduced. The cost function in that approach, which from now on will be called Sparse ICA (SICA), has two terms, namely, an independence term and a sparsity term. The latter penalizes the ICA cost function through l_1 regularization. Two regularization parameters are used, ϵ and λ , for smoothing and sparsity, respectively. The tuning of these parameters is problem dependent. Indicative values are given in [71]. The code implementing SICA comes from [72].

As mentioned previously, TPICA imposes the independence of the sources and the trilinear structure of CPD in different steps. TPICA suffers from a divergence issue caused by the two different objectives, of the ICA step and the Khatri–Rao step (estimation of \mathbf{B} and \mathbf{C}). An alternative way of combining ICA and CPD was proposed in [51], as an attempt to overcome the disadvantages of TPICA. ICA-CPA proposes the incorporation of the trilinear structure during the ICA computation, which is shown to amount to the CPD of a 5th-order partially symmetric tensor containing the 4th-order cumulants of the 3-way data. In the simulation section, ICA-CPA has been implemented with the use of Tensorlab [73].

2.5. Implementation and Performance measures

2.5.1. Algorithms and Model Selection

Tensorlab 3.0 [73], a Matlab package for tensor algebra, is the tool that was mainly used for the tensor decompositions in this paper. The Non Linear Least Squares (NLS) method is employed, both for CPD (3-way, (2)–(4)) and BTM, as implemented in Tensorlab. PARAFAC2 was implemented based on [15] following an ALS implementation similarly with [74], hence in the related simulations the ALS implementation (and not the NLS one) of CPD [73] was adopted (in order to ensure that the performance differences are due to the models and not to the specific optimization method, ALS/NLS). TPICA was used in its non-iterative version [75], since the iterative version was shown to be flawed [8]. TPICA was tested with both FastICA (the algorithm originally proposed [9]) and Infomax [68] (as proposed in [8]), in order to assess the influence to the performance of the ICA algorithm. The implementation of IVA using a multivariate Laplace prior [30] included in Group ICA of the fMRI Toolbox (GIFT) [70] was adopted. Multiple convergence criteria have been used and the algorithms were terminated when at least one of those was met. Those criteria are: a) maximum number of iterations equal to 2000; b) a lower bound on the relative change in the value of the objective function during a step (meaning $|f(x_{i+1}) - f(x_i)|/|f(x_i)|$ equal to 10^{-6} and c) a lower bound on the size of the step (meaning $|x_{i+1} - x_i|$ equal to 10^{-12}).

Rank estimation in the simulations was performed with the aid of the Core Consistency Diagnostic (CorConDia) method [47] and the triangle method [48]. Following [7], the Contrast to Noise

Ratio (CNR) is defined as the Frobenius norm of the signal divided by the Frobenius norm of the noise. The estimated rank of the decomposition increases as the CNR levels decrease significantly, because some peaks of the noise have higher amplitude than the useful signal and are recognized as additional sources [7]. BTM and CPD are sensitive to the correct estimation of R (both overestimation and underestimation result in loss of accuracy). The performance of BTM is not sensitive to overestimation of L .

2.5.2. Initialization

Initialization is crucial for CPD, BTM and PARAFAC2 approximations, since they are known to suffer from local minima issues (also observed in our simulations). In order to overcome such problems, we used the multi-initialization technique. It consists of choosing first a “small” value for the associated convergence criterion (that is, a criterion that can be easily met) and running the algorithm multiple times (30 times in our case). Then, we select the initialization that gives the solution with the best fit, which is subsequently adopted to re-run the algorithm with a normal value for the convergence criterion. For CPD, the generalized eigenvalue decomposition (GEVD) method [76] was also considered, as implemented in Tensorlab 3.0 [73] (and is usually the one selected as “optimal” during the multi-initialization procedure). It was used as one of the possible initializations also for PARAFAC2 but not with equally good results (as it is designed for multilinear models). It has

Input : A 3rd-order tensor $\mathcal{T} \in \mathbb{R}^{I_{xy} \times I_t \times I_s}$, R and L

Output: Spatial factors $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_R]$, $\mathbf{Y} = [\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_R]$ with $\mathbf{X}_i \in \mathbb{R}^{I_x \times L}$ and $\mathbf{Y}_i \in \mathbb{R}^{I_{yz} \times L}$, temporal factor $\mathbf{B} \in \mathbb{R}^{I_t \times R}$ and subject factor $\mathbf{C} \in \mathbb{R}^{I_s \times R}$.

- 1 Use GEVD to compute $\mathbf{A} \in \mathbb{R}^{I_{xyz} \times R}$, $\mathbf{B} \in \mathbb{R}^{I_t \times R}$ and $\mathbf{C} \in \mathbb{R}^{I_s \times R}$ (Eq. (3))
- 2 Fold \mathbf{A} into $\tilde{\mathbf{A}} \in \mathbb{R}^{I_x \times I_{yz} \times R}$ (tensorization)
- 3 **for** $r \leftarrow 1$ **to** R **do**
- 4 SVD of $\tilde{\mathbf{A}}_r = \mathbf{U}_r \mathbf{\Delta}_r \mathbf{V}_r^\top (= \tilde{\mathbf{A}}(:, :, r))$
 (Keep the L most significant singular values)
- 5 $\mathbf{X}_r \leftarrow \mathbf{U}_r(:, 1 : L) \mathbf{\Delta}_r(:, 1 : L)$
- 6 $\mathbf{Y}_r \leftarrow \mathbf{V}_r(:, 1 : L)$
- 7 **end**

Algorithm 2: Computation of BTM based on GEVD for CPD.

been shown in [11] that, in the noise-free case, BTM can also be computed with the aid of a GEVD. Nevertheless, for the sake of simplicity, we used in

our simulations (as one of the possible initialization of the multi-initialization procedure) the scheme called here Algorithm 2, based on GEVD for CPD.⁸

2.5.3. Performance evaluation

The performance evaluation is based on the Pearson correlation, ρ (as in [7, 9]). In all cases, the sources with the highest correlation with the true ones are included in the correlation matrices (see Supplementary Material) and in the figures (e.g., in the case of $R = 3$) every source is matched to the one with which it has the highest correlation. If two or more obtained sources match the same artificial source, then the one with the highest correlation is selected and the second one is matched to the source with the second highest correlation. We define a metric called Absolute Concrete Correlation Distance (ACCD), which quantifies the deviation of the absolute Pearson correlation of the obtained sources (\mathbf{A}_o) with the actual ones (\mathbf{A}) from the absolute Pearson correlation of the actual ones with themselves:

$$\text{ACCD}_{i,j} = |\rho(\mathbf{A}_{o_i}, \mathbf{A}_j)| - |\rho(\mathbf{A}_i, \mathbf{A}_j)| + 1, \quad (13)$$


$$i, j = 1, 2, \dots, R.$$

For $i = j$, $\text{ACCD}_{i,i}$ is equal to the Pearson correlation. Usually, the metric employed for evaluation of the algorithms is the correlation of the source obtained with the true one, $\rho(\mathbf{A}_{o_i}, \mathbf{A}_j)$ for $i = j$, which would ideally be equal to 1. However, this metric does not provide all the required information, since it does not account for the residuals among the sources. In a correlation matrix, the residual (cross-talk between the maps) is represented at the off-diagonal elements, which are ideally zero, if the sources have no correlation. Hence, a good separation result corresponds to high diagonal and low off-diagonal elements. The use of ACCD allows an easier interpretation of the result; the correlation of the actual sources with themselves is taken into consideration and hence a perfect separation will result into a matrix with ones everywhere (no matter if there is overlap or not between sources). The ACCD of the diagonal elements will be called principal ACCD, while the ACCD of the off-diagonal elements will be called cross-talk ACCD. The use of the absolute value of the correlation makes this metric insensitive to the sign indeterminacy of the multi-way models. In the

⁸Thanks to Dr. Otto Debals, KU Leuven, for fruitful discussions about this question.

real data case, we will resolve the sign indeterminacies using the flip sign methods proposed in [77]. The standard deviations of the mean correlation coefficients were similar for all methods (slightly higher in datasets where the power of the signal of activation is lower) and hence they are not reported here.

2.5.4. Compression

Prior to decomposing large datasets, it is useful to compress the data to reduce the computational load. Assuming a tensor $\mathcal{T} \in \mathbb{R}^{I_1 \times I_s \times I_t}$ with $I_1 > I_t I_s$, the QR decomposition of $\mathbf{T}_{(1)}$ is given by the matrix product $\mathbf{Q}\mathbf{R}$, with $\mathbf{Q} \in \mathbb{R}^{I_1 \times I_t I_s}$ having orthonormal columns and $\mathbf{R} \in \mathbb{R}^{I_t I_s \times I_t I_s}$ being upper triangular. If \mathbf{R} has a perfect CPD fit, $\mathbf{R} = \hat{\mathbf{A}}(\mathbf{C} \odot \mathbf{B})^\top$, then $\mathbf{Q}\mathbf{R} = \mathbf{Q}\hat{\mathbf{A}}(\mathbf{C} \odot \mathbf{B})^\top$ and hence \mathcal{T} has a perfect CPD fit [7], with factors \mathbf{Q} , $\hat{\mathbf{A}}$, \mathbf{B} and \mathbf{C} . Thus, it is sufficient to calculate a CPD decomposition for \mathbf{R} , instead of the much bigger $\mathbf{T}_{(1)}$. In the data considered here, since $I_1 = 64 \times 64 \times 30 = 122800$ is larger than the product of $I_t = 370$ with $I_s = 18$, instead of computing the CPD of $\mathbf{T}_{(1)} \in \mathbb{R}^{122800 \times 6660}$ we will compute the CPD of $\mathbf{R} \in \mathbb{R}^{6660 \times 6660}$. In the absence of noise, the tensor could be further compressed, to $\hat{\mathcal{T}} \in \mathbb{R}^{R \times R \times R}$. The same compression ratio has been used for BTD throughout our simulations. For CPD and TPICA, a grey-matter mask is applied prior to the unfolding and to the compression done. For the application of BTD, a perfect square (or cube in 5-D data) is needed, and hence voxels, which belong to areas outside the brain or the ventricles, are included in the analysis. In order to decrease the effect of the non grey-matter voxels on the separation performance, two methods have been tested: (a) The grey-matter mask is applied on the data and all the non-grey matter voxels are set to zero. Compression is then performed (and hence the zero-valued voxels are suppressed). (b) The non-grey matter voxels are considered as missing values and the decomposition of an incomplete tensor is sought for [78, 73]. We noted that, although the second method is theoretically more accurate (since the non-white matter voxels do not participate in any stage of the analysis), the results were similar, with a significant difference in the computation time; method (b) (incomplete tensors) needs ~~time more than double than that~~ of method (a). Hence, to decrease the effect of the non grey-matter voxels, method (a) has been used in all the simulations presented next. 

3. Results

3.1. Synthetic data

The results of three different simulation studies are presented, with scenarios and data reproduced from [7, 8, 44]. We opt to consider these three different simulated experiments in order to exhibit the performance gain from the use of higher-order tensorization, along with BTD and BTD2, within different settings. In the first simulation study, indicated as “simulation of a perception study”, the importance of the higher-order tensorization is demonstrated and the effect of the choice of L on the performance of BTD is examined. TPICA with Infomax [68], CPD, ICA-CPA [51], SICA [28] and BTD are tested, with simulated data from [8]. In the section titled “Simulation with artifacts,” where datasets simulated with SIMTB are used, the effect of different types of physiological noise (artifacts) is reported with the aid of GICA [25], IVA [30], SICA and BTD. In that section, we also evaluate the performance of the higher-order unfolding against matrix-based ICA methods, in the presence of physiological artifacts of higher rank. After having verified that higher-order unfolding with BTD outperforms all the other methods in different settings, a last simulation with data used in [9, 7] is performed. In this last simulation study, indicated as “Multi-slice simulation”, different HRFs per subject are considered and the algorithms tested include TPICA (with FastICA, as in [9], in order to check for the effect of the ICA method adopted), CPD and BTD.

The noise added is white Gaussian (to follow the structure of the TPICA model, so that no further preprocessing is necessary), with different standard deviations per voxel.⁹ Preprocessed resting-state fMRI data (the same resting state data as in [9, 7]) were used to estimate background noise parameters, voxel-wise means and standard deviations.

3.1.1. Simulation of a perception study

The data of three subjects were simulated under the assumptions of a simplified version of a realistic perception study [8]. The simulated data used are a 60×50 axial slice of voxel activity from somewhere near the level of Broca’s area. The data from each subject contained three sources with different

activation levels; the three activation patterns have strengths (3, 4, 5), (2, 3, 4) and (2, 2, 3) times the average noise standard deviations for subjects 1–3, respectively. As stated in [7, 8, 9], these are reasonable subject activation weights for typical fMRI data.

The spatial maps and the time-courses of these sources are presented in Fig. 6. The first component represents the activation in the Broca’s area while the second and third components represent the visual and motion perception components, respectively. In [8], 4 different versions of Spatial Map 3 with different percentage of overlap (between maps 3 and 2) were used. In this paper, an adapted version of the maximum overlapped source (50% of shared active voxels and correlation with Spatial Map 2 equal to 0.4693) will be examined. Instead of S_1 of Fig. 3 (which has been used in [8]), we used the more realistic S_2 source (Figs. 3 and 6), which has higher spatial rank. The activity level at each active voxel was randomly sampled from a Uniform [0.8,1.2] distribution for each replication of each simulation condition. For later use (comparison using the correlations with the estimated sources), we provide here the correlations of the spatial maps, \mathbf{A} , and the time-courses, \mathbf{B} , with themselves:

$$\text{Cor}(\mathbf{A}) = \begin{bmatrix} 1 & 0.03 & 0.04 \\ 0.03 & 1 & 0.36 \\ 0.04 & 0.36 & 1 \end{bmatrix}, \text{Cor}(\mathbf{B}) = \begin{bmatrix} 1 & 0.14 & 0.06 \\ 0.14 & 1 & 0.46 \\ 0.06 & 0.46 & 1 \end{bmatrix}$$

All the previously stated rank estimation methods resulted in the same value for the rank R . The rank of the decomposition was estimated as $R = 3$ for high CNR, while for $\text{CNR} = 0.08$, for CPD and BTD it was $R = 4$. When values higher than $R = 6$ were used, the performance of the CPD and BTD decompositions was degraded, while the performance of TPICA was not affected from the overestimation of R . The influence of overestimation of R is similar for CPD and BTD and this is the

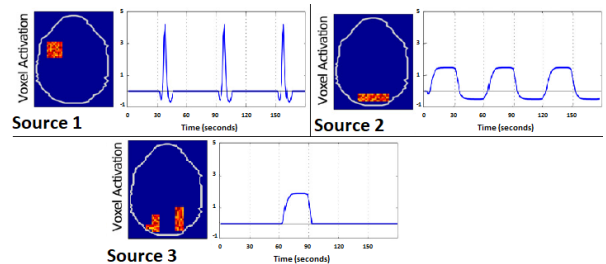


Figure 6: Spatial maps and time-courses of the three sources used in the perception study.

⁹Rician noise has also been tested with similar findings. We stick to the Gaussian model here for easier comparison with [7, 8, 9].

reason that no tables with correlation comparisons based on the rank are presented here (for such comparisons, see [7]). In order to select the appropriate L for BTD, different values were tested with different levels of noise. The best performance is obtained for L equal to the “spatial rank” of the third source ($L = 3$). As can be seen in Fig. 7, at high and medium CNRs (0.50 and 0.15), the result of the decomposition is robust to the overestimation of L . For values of L lower than 3, the separation of the third source is not good. For higher instances of noise (CNR= 0.08), it can be noted that, although values of L lower than 3 still give the worst results, the increase of L to values higher than 6 results also in a deterioration of the performance. As an example of the use of Heuristic 1, its result for the lowest CNR (CNR= 0.08) can be viewed in Supplementary Material.

In Figs. 8, 9 and Fig. 10, the performance gain of BTD over the other methods, as the noise power increases, can be observed (mean of 30 runs). Fig. 8 shows the mean (over the 3 different sources) principal ACCD (Pearson correlation) of the five methods tested (the standard deviation per source is also presented). For SICA, after fine tuning, the values of the regularization parameters were set to $\lambda = 0.01$ and $\epsilon = 0.1$. It must be noted that the performance of SICA is very sensitive (at least in this simulation setting) to the λ value but not so sensitive to the ϵ value (especially to its underestimation). In Fig. 9, the principal ACCDs for all the methods, for each source (not the mean), are shown with barplots while the mean cross-talk ACCDs (the mean off-diagonal elements for every source) are represented with diamonds. TPICA starts failing even at values of CNR around 0.25. Indeed, the second and third spatial maps are recognized

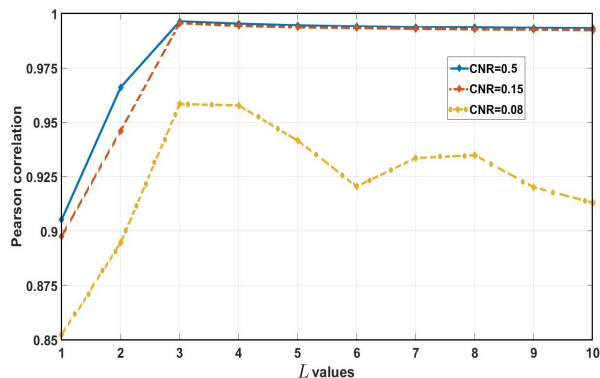


Figure 7: Decomposition of the data with different L values at different CNR values (perception study).

as one (Fig. 9). Although ICA-CPA seems to be more robust to noise than TPICA, it has the same behavior in the presence of overlap. In overlapped sources (Sources 2 and 3), the result of the decomposition is not accurate. All the methods succeed in recognizing the first source, which does not have any spatial overlap with the others, almost perfectly (small cross-talk is observed only at low CNR values). CPD, in line with the findings of [8], maintains the almost perfect separation of the sources at values of CNR ≈ 0.15 and starts having small problems with overlap at CNR = 0.12 (not shown in the figure). We can note that, especially for the time-course of Sources 1 and 2, the separation result is really bad at CNR = 0.08. SICA (if tuned properly) has comparable performance with CPD, and outperforms the other independence-based methods. Sources 2 and 3, which are spatially overlapped, are recognized correctly (with the aid of the sparsity regularizer) and the crosstalk is smaller than in CPD and BTD. On the other hand, the performance is less robust to the noise than in CPD (and BTD), since the level of noise impacts also the level of sparsity. For CNR = 0.08, the first two sources of TPICA and ICA-CPA are noise-like and this is the reason why the principal ACCD is so high, while a combination of the second and third sources is obtained in Source 3. The fact that BTD identifies the sources correctly in the overlapping areas, even at CNR = 0.08, must be emphasized. The performance of BTD starts degrading only at CNR < 0.06.

Split-half reliability analysis has shown that the tensor decomposition methods (CPD and BTD) will provide identical results if the correct number of components is chosen [79]. The same results are also obtained even if the sample size is reduced more than half (keeping in mind that it must remain $L_t > R_t$). In ICA-based methods, where statistical assumptions are crucial, there is a mild decrease in the correlation of the two datasets when fewer samples are used. In SICA, it was observed that the reliability of the algorithm deteriorates with large λ values.

In terms of their computational requirements, BTD and ICA-CPA are the most costly. Some indicative values of the required time per run are given in Table 1. It must be noted that, when compression is used, the time needed for both CPD and BTD decreases significantly. Taking into consideration the higher complexity of ICA-CPA as compared with TPICA, the similar performance of the

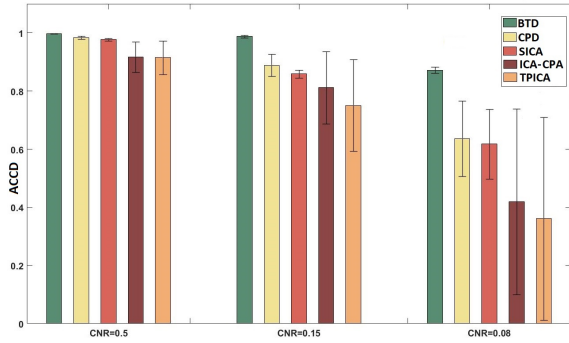


Figure 8: Mean ACCD at different values of CNR (perception study).

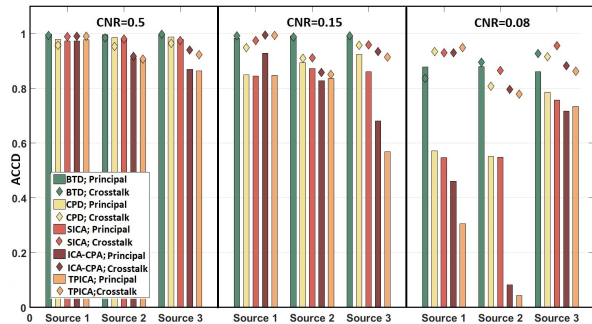


Figure 9: ACCD per source at different values of CNR (perception study).

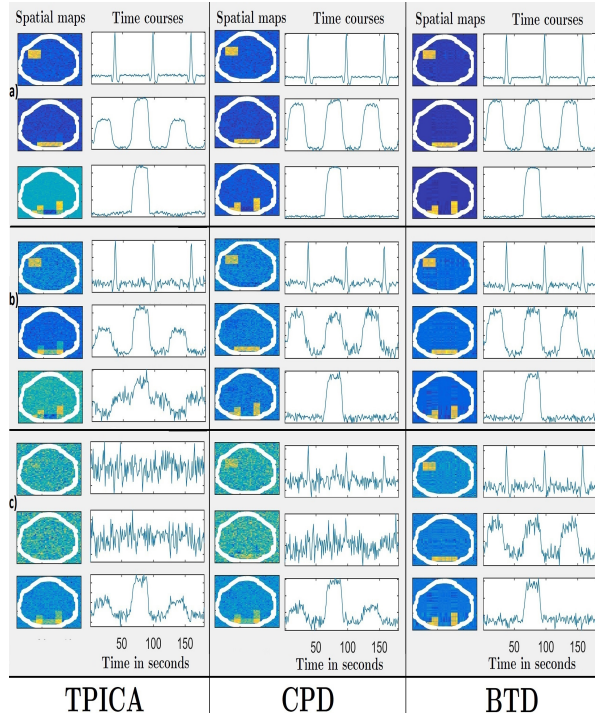


Figure 10: Decomposition of the data with different values of CNR. a) CNR=0.50 b) CNR=0.15 c) CNR=0.08 (perception study).

two algorithms in the presence of overlap, and the fact that TPICA is widely known in the fMRI community, TPICA will be considered instead of ICA-CPA, despite the superior performance of the latter in the presence of noise.

Methods		No Compression	Compression
TPICA		0.906655 s.	
SICA		1.234582 s.	
ICA-CPA*		13.583246 s.	
CPD		1.939442 s.	1.012235 s.
BTD	$L = 1$	1.832343 s.	0.984838 s.
	$L = 2$	2.123863 s.	1.323457 s.
	$L = 3$	5.402858 s.	1.844389 s.
	$L = 4$	5.735463 s.	2.343239 s.
	$L = 5$	8.862472 s.	4.65328 s.
	$L = 6$	11.810893 s.	5.203684 s.
	$L = 7$	18.45496 s.	6.123953 s.
	$L = 8$	23.253864 s.	8.426508 s.
	$L = 9$	26.831358 s.	8.910285 s.
	$L = 10$	30.513510 s.	10.023486 s.

*The code used in [51] is not available. The implementation, made by the authors, might not be the optimal in terms of efficiency.

Table 1: Mean computation time of each method in seconds (s).

3.1.2. Simulation with artifacts

SIM TB [80], a simulation toolbox running in the Matlab environment, which allows flexible generation of fMRI datasets under a model of spatiotemporal separability. The dataset used in this section has also been used in [44, 81] and it is publicly available in [82]. It consists of 8 sources (Fig. 11), one is task-related (1), two are transiently (occurring during only one or more portions of a task) task-related (2, 6), and five are artifact-related (3, 4, 5, 7, 8). Furthermore, five of them are super-Gaussian (1, 2, 5, 6, 8), two are sub-Gaussian (3, 7), and one is Gaussian (4) in the spatial domain. The characteristics of the various artifacts (physiological noise) are different. The time-course for head motion varies slowly with sudden transient fluctuations (8), while those of the respiratory and cardiac pulsation components appear to have random fluctuations (4, 5). Scanner drift is another artifact and is characterized by a slowly rising time-course (7). Each spatial map consists of a single slice of 60×60 voxels and each time-course lasts 100 seconds. Data for 5 different subjects were generated. The data of each subject consists of all the sources with different amplitude, and the amplitude per subject for each source was drawn from a half-normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 3$.

Using the sources of [82] (Fig. 11), three different simulations with different datasets were performed: Dataset a is the one provided in [82] with two different instances of noise. In Dataset b, the Gaussian Source 4 was replaced by Source 9 (Fig. 11)

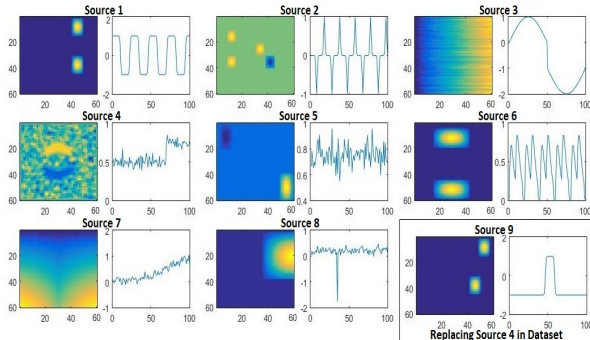


Figure 11: Sources used in [84].

with high spatial overlap (50% overlap) and temporal overlap (around 20%) with Source 1. Two different instances of noise were also considered. In Dataset c, subject variability was introduced in the spatial domain in two of the sources of interest (1, 6). For introducing subject variability to the dataset, rotation of the spatial map of Source 6 and simultaneous translation of that of Source 1 were performed (similarly to [83]). Two different levels of variability were examined. In the low level variability case, the first subject has 0 degrees of rotation and all the other subjects have a rotation in increments of 4 degrees (hence a rotation of 16 degrees for the fifth subject). In an analogous manner, subjects 2–5 have voxel shifts at increments of 2 voxels with respect to the 1st subject (hence a shift of 8 voxels in the 5th subject). In the high level variability dataset, the increments were set equal to 6 degrees for the rotation and 4 voxels for the shift.

The performance evaluation is once more based on ACCD and visual inspection, with a small difference to the previous sub-section. Instead of presenting the ACCD values for every source, a mean principal ACCD for all the sources will be given, since the number of sources (8) is large and the resulting figures would be hard to read. Furthermore, a mean principal ACCD for the sources of interest (1, 2, 6) will be presented (we have noted that IVA and GICA have poor performance for sources 3 and 7, which are sub-Gaussians with high overlap but not of high interest, since they are artifacts and hence they are omitted). The following color code is employed to highlight the most important table entries. Green (with a star) is used when a method is significantly better than the other(s), blue (with a double dagger) is used when the result (correlation) of a method is not good but still the source can be identified, while red (and a dagger) highlights cases where a method fails to identify a source or

has a very strong cross-talk (correlation with wrong sources). In cases where all the methods perform equally well, no color is used.

3.1.2.1. Dataset a [82]

For CNR=2, R was estimated equal to 8, while for CNR=0.80, R was set equal to 9. The performance of GICA and IVA is not sensitive to overestimation of R , while for BTM and R values higher than 11, the mean correlation starts decreasing. Source 4 has high “spatial rank” (equal to 58). Hence, in order to obtain good separation results with BTM, a high L is needed ($L = 40$). Of course, this renders it more complex (and hence expensive in terms of computational time) compared to IVA and GICA. As it can be noted from Table 2, all the methods succeed in identifying almost perfectly the sources of interest at high CNR. We can note that SICA, in contrast to the previous simulation, performs worse than the other ICA methods. This deterioration in the performance of SICA is due to the fact that some of the sources are dense (not sparse) and since they can not be well presented by a sparse model the residual of the source (the “left-overs” of the signal of the source which has not been modelled) causes a general drop in the performance. IVA and GICA have slightly worse results in the mean correlation of all sources (first two columns for each CNR value in Table 2) since they fail to distinguish Sources 3 and 7 (this was noted in simulations with all datasets so it will not be mentioned again). The extra noise added (CNR=0.80) influences the performance of the algorithms. GICA and IVA start suffering from cross-talk (mainly in the artifact sources) and in areas with overlap (even minimum), while the performance of BTM remains stable, apart from some minor problems with the time-courses, which are

Methods	CNR=2.00				CNR=0.80			
	All sources		Task related		All sources		Task related	
	Maps	Tcs	Maps	Tcs	Maps	Tcs	Maps	Tcs
IVA	0.88	0.82	0.92	0.90	†0.66	†0.60	0.78	†0.69
GICA	0.88	0.81	0.84	0.88	†0.75	†0.66	0.81	†0.75
BTM	0.88	0.88	0.95	*0.98	*0.83	*0.88	*0.92	*0.95
SICA	†0.68	†0.66	†0.82	†0.85	†0.56	†0.54	†0.58	†0.56

Table 2: Mean ACCD _{i,j} for $i = j$ of Dataset a.

slightly affected from the residual of Source 4

adding a small trend, similar to its time-course.¹⁰

3.1.2.2. Dataset b – Spatial and temporal overlap

Similarly to Dataset a, in Dataset b (where Source 4 has been replaced by a source of high overlap), R was estimated equal to 8 for $\text{CNR} = 2.00$, while for $\text{CNR} = 0.80$, R was set to 9. The way all algorithms perform with respect to the rank estimation is similar to that for Dataset a. From Table 3, it can be observed that, even at high CNR ($\text{CNR} = 2.00$) GICA and IVA have difficulties in separating correctly the sources. cross-talk between the overlapped areas is observed in Fig. 12 (a) and (b) (which depicts one representative run of the algorithms). This cross-talk increases at $\text{CNR} = 0.80$ (Fig. 13(a) and 13(b)). A difference in IVA and GICA can be seen here. Namely, the performance of IVA is affected more by the overlaps in the spatial maps while that of GICA in the time-courses. Notice that in Sim TB all the areas of activation are generated as 2-D Gaussian distributions, hence, although the overlap in percentage of voxels introduced is quite high, the fact that the Gaussian density is unimodal aids the algorithms. The same percentage of overlap introduced in a source with activation area created by a uniform distribution (as in Section 6.1) or by a Tuckey (tapered cosine) window causes even stronger cross-talk in the sources obtained by GICA and IVA. Again, the robustness of BTD can be observed. It is even more robust (in the time-course mode) compared to Dataset a, since now the high-rank Source 4 does not exist (the only source that sometimes is not perfectly recovered from BTB is Source 8, which has the lowest amplitude).

Methods	CNR=2.00				CNR=0.80			
	All sources		Task related		All sources		Task related	
	Maps	Tcs	Maps	Tcs	Maps	Tcs	Maps	Tcs
IVA	± 0.69	± 0.73	± 0.75	0.78	± 0.45	± 0.57	± 0.57	± 0.62
GICA	0.80	± 0.69	0.76	± 0.66	± 0.69	± 0.53	± 0.74	± 0.62
BTB	± 0.94	± 0.99	± 0.98	± 0.99	± 0.80	± 0.94	± 0.89	± 0.95

Table 3: Mean $\text{ACCD}_{i,j}$ for $i = j$ of Dataset b.

¹⁰Since the value of L used is smaller than the true “spatial rank,” a small residual exists. In Section 4.2 of the Supplementary Material, figures with the result of BTB with different L values are included and the effect of the underestimation of L can be noted therein. The effect could be avoided by increasing L even more, but the gain is not worth the corresponding increase in the complexity. Furthermore, uniqueness issues could then come up (affecting the column rank of matrices \mathbf{A} and \mathbf{B} and invalidating the assumption of Theorem 4.1 [11], since $\min(I_x, I_{yz})$ is less than LR) instead of being greater than or equal.

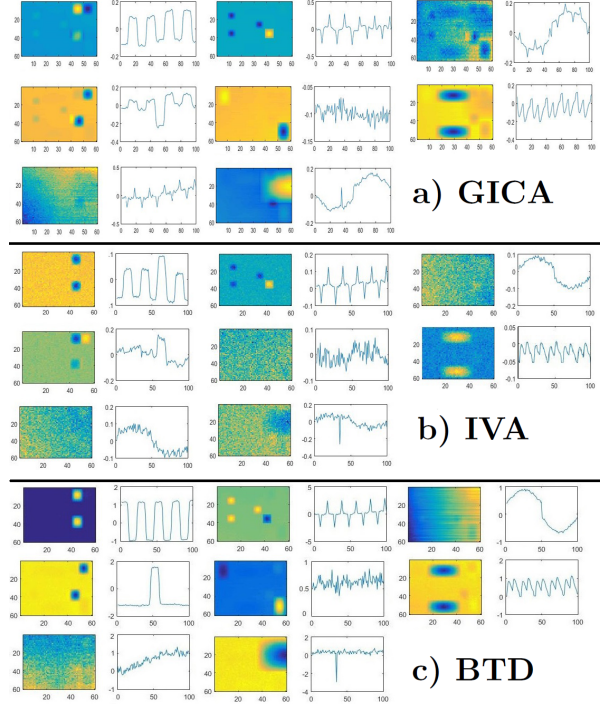


Figure 12: Results from the decomposition of Dataset b at CNR=2.00 a) GICA b) IVA c) BTB.

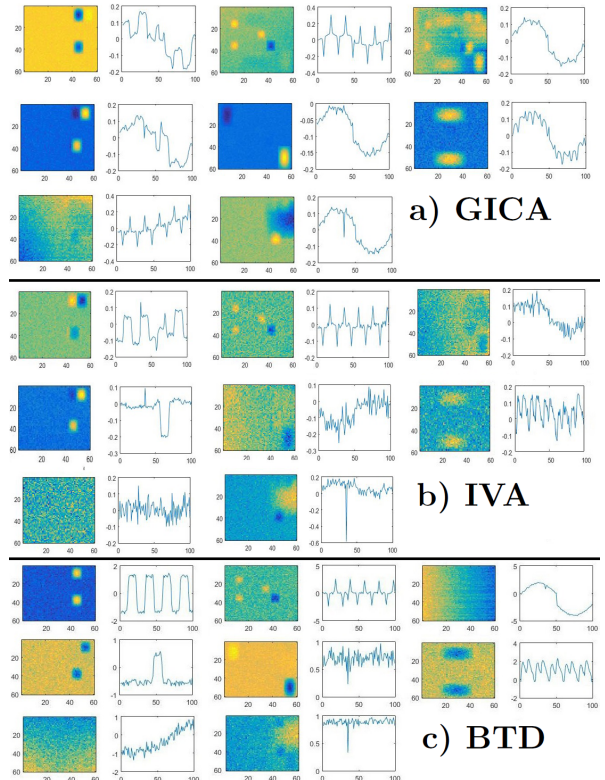


Figure 13: Results from the decomposition of Dataset b at CNR=0.80 a) GICA b) IVA c) BTB.

3.1.2.3. Dataset *c* – Subject variation

The performance of the algorithms with respect to the spatial subject variability is evaluated in this simulation. Source 9 with overlap is removed (now we have only 7 sources) and the simulation is performed only at high CNR (CNR = 2.00), to remove any influence from overlap and noise. Table 4 demonstrates the superiority of IVA in handling datasets where the variability of the spatial maps is high. At the low level variability, we can note that all methods perform similarly (note that without variability at this CNR, BTD would result in slightly higher correlations). With high level of variability, it is evident that IVA outperforms the other two methods, and BTD is the worst due to the assumption of multilinearity, since every subject must have the same (or at least very similar) spatial map for every source for the method to work properly.

Methods	Low variation				High variation			
	All sources		Task related		All sources		Task related	
	Maps	Tcs	Maps	Tcs	Maps	Tcs	Maps	Tcs
IVA	0.80	0.76	0.86	0.80	*0.74	*0.78	*0.81	*0.83
GICA	0.79	0.77	0.84	0.79	†0.75	†0.66	†0.68	†0.69
BTD	0.81	0.82	0.84	0.83	†0.54	†0.62	†0.61	†0.66

Table 4: Mean ACCD_{*i,j*} for *i* = *j* of Dataset *c*.

3.1.3. Multi-slice simulation

The signal consists of artificial voxel activation maps (of three different slices), time patterns and activation strengths for three subjects. In this simulation setting, the brain consists of 3 slices, so the cross-talk between different slices can be also checked. The voxel-wise noise mean and variance are the same for each subject. Beckmann and Smith [9] consider five different artificial fMRI

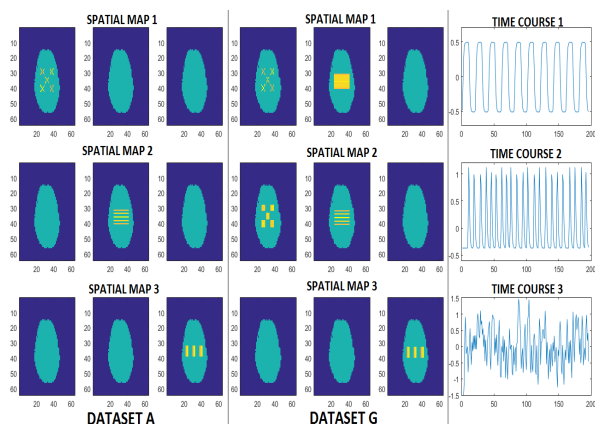


Figure 14: Spatial maps of Datasets A and G. Common time-courses.

datasets, named A–E, which differ only in their signal part and have no spatial overlap, while Steegman [7] added three more, F–H, with high percentage of overlap between the sources. In Dataset G, the first two spatial maps are a combination of Spatial Maps 1 and 2 of Dataset A, hence, the activity of the first two maps takes place in the first two brain slices, and they have 51% and 63% of their active voxels in common (high percentage of overlap), respectively. In Map 3, the time-courses \mathbf{B} , and the noise instances are the same as in Dataset A. Having been convolved with a canonical HRF, the time-courses and the spatial maps consist of three different spatiotemporal processes, which are present in every subject with different amplitude (same with Section 3.1.1 and [9, 7]). Figs. 15–16 depict the principal ACCDs and mean cross-talk ACCDs. As in Section 3.1.1, we will provide the correlation matrices of the actual sources with themselves for the sake of comparison:

$$\text{Cor}(\mathbf{A}) = \begin{bmatrix} 1 & 0.04 & 0.16 \\ 0.04 & 1 & 0.04 \\ 0.16 & 0.04 & 1 \end{bmatrix}, \text{Cor}(\mathbf{B}) = \begin{bmatrix} 1 & 0.02 & 0.16 \\ 0.02 & 1 & 0.05 \\ 0.16 & 0.05 & 1 \end{bmatrix}$$

$$\text{Cor}(\tilde{\mathbf{A}}) = \begin{bmatrix} 1 & 0.49 & 0.05 \\ 0.49 & 1 & 0.05 \\ 0.05 & 0.05 & 1 \end{bmatrix},$$

where \mathbf{A} stands for the spatial maps of Dataset A and $\tilde{\mathbf{A}}$ for those of Dataset G (as previously mentioned, the time-courses \mathbf{B} are common).

In this paper, Datasets A, G (lowest and highest spatial overlap) (Fig. 14) and C are used. Dataset C has the same spatial maps as A, but with different convolution parameters for the generation of the time-courses. In Datasets A and G, a canonical HRF is assumed for all subjects, while in Dataset C, the HRFs of every subject differ in mean lag and standard deviation (stdev = 3, 3.5, and 4 seconds, mean lag of 4, 5, and 6 seconds). This induces small differences in the temporal signal per subject. Using the same HRF parameters (as Dataset C) and spatial maps of Dataset G, an extra dataset, Dataset I, has been generated with different time-courses per subject with high spatial overlap. Furthermore, the mean lag in Datasets C and I has been increased to test the performance of the algorithms in such conditions. The data of Datasets C and I do not have a strict multilinearity structure and it has been shown [7, 8] that, in such cases, CPD fails to correctly identify the components. The noise instances used in all the datasets are the

same, however, the CNR values are different between Datasets A and G (and hence C and I), due to the fact that Dataset G has a stronger signal.

3.1.3.1. BTD

The rank R was set equal to 4 (estimated rank, as also shown in [7]), and the L value was estimated (with the use of Heuristic 1) equal to 10. At high and medium CNR values, the result is robust to overestimation of L even with values up to 30, while at the lowest CNR used in Figs. 15 and 16 a mild degradation of the performance is noted for $L > 16$.

Dataset A – Low spatial overlap. Note the stability in the performance of BTD compared to the other two methods. Furthermore, the different effect of noise in TPICA and CPD can be readily observed. In TPICA, the correlation between the estimated spatial map and the “true” one decreases dramatically as the level of the noise gets higher, while in CPD the decrease is slower but with a significant increase of the cross-talk (correlation among “wrong” spatial maps). In terms of cross-talk between different sources, TPICA performs better than CPD and BTD (if the sources are indeed independent as in Dataset A). The fact that no overlap exists also improves the result of the other methods (compared to cases with overlap), since the problem is now easier. Generally, when the components are truly spatially independent, TPICA (and all the other methods based on the independence assumption) yields better results in terms of crosstalk than unconstrained tensor decomposition methods. Employing constraints that agree with the data is expected to yield better results. The present case is an instance of this, since if no overlap exists, the independence assumption, on which the ICA-based methods rely, is valid.

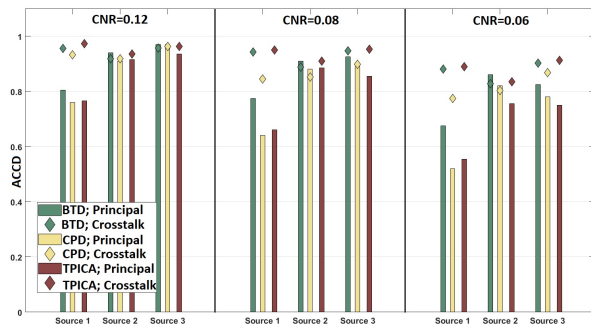


Figure 15: ACCD of the sources of Dataset A with different values of CNR.

Dataset G – High spatial overlap. Fig. 16 demonstrates the poor performance of TPICA compared to CPD and BTD in cases of high overlap, and the gains offered by BTD against the other two methods, in the presence of high levels of noise. Even at relatively high CNR ($=0.12$), TPICA fails to correctly separate the first two sources which overlap. Visual inspection shows that Spatial Map 1 of TPICA is an approximation of the common part of Sources 1–2. This could mean that TPICA splits the common part and the two unique parts of the sources. However, this is not the case since Spatial Map 2 is not the unique part of one of the sources but a combination of the two unique parts with the common part and Source 4 is noise. We can note that TPICA exhibits lower cross-talk for the independent Source 3 compared to the other methods (similarly to Dataset A) and higher correlation (at high CNR). Similarly to Dataset A, the main problem of CPD is the high cross-talk (between Sources 1 and 3). As the noise level increases, the difference in the performance of the decompositions increases in favor of BTD as the result of BTD is less affected by noise.

3.1.3.2. PARAFAC2

Even in datasets where the assumption of multilinearity is valid (Datasets A and G with the same time-course per subject) PARAFAC2 performs well, similarly with CPD, only with a slightly higher cross-talk (cf. Fig. 18). In Datasets C and I, where multilinearity is not satisfied, CPD completely fails to correctly extract three different sources, even if we consider an equivalent multilinear model of higher rank ($R = 9$). TPICA manages to handle the non-overlapped case of Dataset C (the result is slightly better than that of PARAFAC2, with better crosstalk ACCD), but in cases of high

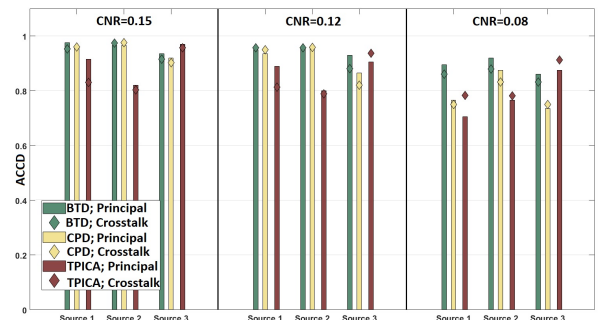


Figure 16: ACCD of the sources of Dataset G at different values of CNR.

overlap (Datasets G–I), it fails dramatically to separate the spatially overlapped sources. If we increase the lag introduced to HRF, the result of TPICA gets worse, while PARAFAC2 remains stable (as despite the lag introduced by HRF the cross product of time-courses is stable) and gets better than TPICA even for Dataset C. The rank R was set to 3 both for CPD and PARAFAC2.

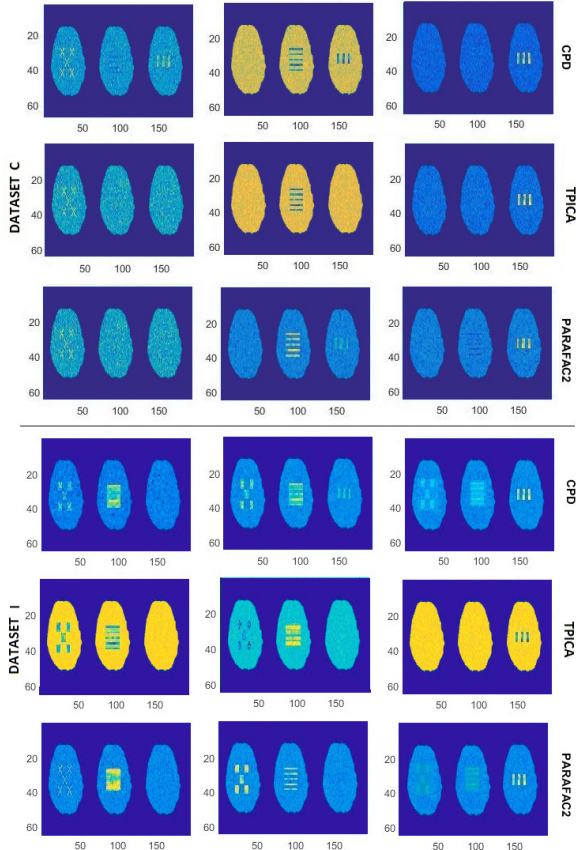


Figure 17: Spatial maps of Datasets C and I.

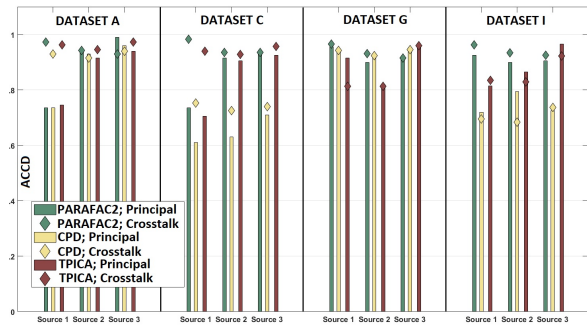


Figure 18: ACCD of the data at different Datasets.

3.1.3.3. *BTD2*

Having verified the effectiveness of BTD and PARAFAC2 in the multi-slice simulation scenarios of Section 5.2, the focus is now turned to the proposed BTD2 method using noise of varying strength. L was estimated to be equal to 30 and R to 3 at $\text{CNR} = 0.08$, for dataset C, and at CNRs 0.15 and 0.12 for Dataset I, while at CNRs 0.08 and 0.06, for Dataset C and $\text{CNR} = 0.08$ for Dataset I, the estimated value was equal to 4.

It can be seen in Figs. 19 and 20 that BTD2 is more robust to noise than PARAFAC2, especially in the noisier cases. The performance of PARAFAC2 deteriorates significantly at CNR values lower than 0.05 for Dataset C and 0.10 for Dataset I, while the result of TPICA worsens at even higher values of CNR . It can also be observed that the component which is less influenced by noise is component no. 3. This is quite reasonable, since it has the minimum spatial and temporal overlap with the other sources. The result of BTD2 is fairly insensitive to the overestimation of L , similarly to BTD. On the other hand, BTD2 is computationally more complex since more factors need to be computed.

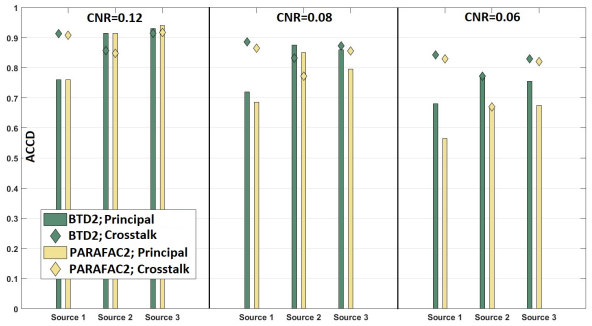


Figure 19: ACCD of Dataset C at different values of CNR (BTD2 - Section 3.1.4).

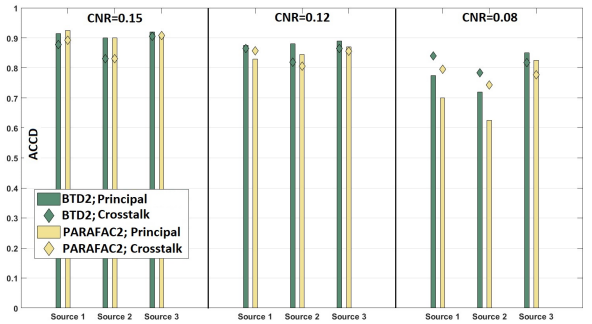


Figure 20: ACCD of Dataset I at different values of CNR (BTD2 - Section 3.1.4).

3.2. Real data

In this subsection, results from the analysis of real data will be presented. The dataset used was obtained from the OpenfMRI [85] database. Its accession number is ds000157 and it has been previously used in [86]. During the experiments, subjects alternately viewed 8 blocks of 24 seconds of palatable food images and the same number of non-food images (e.g., office utensils), with an intersection of 8–16 sec resting blocks (for more information about the task and the experiment, see [86]). It was demonstrated that visual stimuli connected with food elicit increased attention, since higher activation of the primary visual cortex has been noted. In addition, it was found that viewing images of palatable food results in activation of the self-regulation areas (e.g., the lateral prefrontal and orbitofrontal cortex) and the reward regions (e.g., caudate nucleus, ventral striatum and amygdala) [86].

The initial dataset involves 30 different subjects. After performing a quality assessment in the dataset, we identified the subjects with high percentage of movement (subjects 9, 16, 18, 24, 25 and 26) and we excluded them from the analysis. Furthermore, from the remaining 24 subjects, we noted that 6 of them (Subjects 1, 2, 3, 4, 8 and 10) had a different number of time points (378 instead of 370). Since in [86] no details were provided about this difference in the time points, those 6 subjects have been also excluded, leading to a dataset of 18 subjects. The spatial analysis for every image is $64 \times 64 \times 30$ with Repetition Time $TR = 1.6$ and voxel size 4 mm. fMRI data preprocessing was carried out using FMRI Expert Analysis Tool (FEAT) Version 6.00, part of FMRIB’s Software Library (FSL [50]). The following pre-statistics processing was applied. Motion correction using Motion Correction FMRI’s Linear Image Registration Tool (MCFLIRT [87]); no slice-timing correction; non-brain removal using Brain Extraction Tool (BET) [88]; spatial smoothing using a Gaussian kernel of Full Width at Half Maximum (FWHM) of 8 mm (twice the voxel size); grand-mean intensity normalisation of the entire 4-D dataset by a single multiplicative factor; highpass temporal filtering (Gaussian-weighted least-squares straight line fitting, with FWHM=100 seconds).

Rank estimation is a hard (open) problem for noisy multi-way data. With synthetic data, where the number of simulated sources is a-priori known, estimating the rank of the noisy tensor is easier. In

order to estimate the rank for real data, we resorted to different tools: the CorConDia) method [47], the triangle method [48], and the Automated Unsupervised Tensor Mining (AutoTen) method [89] (and the baseline methods described and made publicly available in [89]). The results of the methods were not consistent with each other as the estimated rank ranged from 5 to 30. We decomposed the data with all possible ranks (between 5 and 30) and checked the results. The relative error drops dramatically as the rank is increased up to the value of 6 and then a moderate decrease of the error is observed as the rank increases further. We selected $R = 8$. This is because, for ranks higher than 8, the main component (source of interest), which is the primary visual cortex activated by the blocks of images, might be split. Generally, at a reasonable CNR, adding more components can result in modelling noise and sources of no interest. Moreover, as it has been reported, the techniques which are trying to find a trade-off between complexity and fit might overestimate the rank [35]. For the estimation of L , different heuristic methods can be used [57]. After testing different values and with the application of Heuristic 1, L was set equal to 20 (the result of Heuristic 1 can be viewed in Section 1.1 of Supplementary Material).

As mentioned before, a higher activation of the primary visual cortex is expected when the food images appear. Since the visual component has a common spatial map for both the food and non-food images, the difference in the activation is depicted in the respective time-course. Hence, odd indexed blocks (food images) turn out to have higher amplitude than the even indexed ones (non-food images). The mean difference between the maximum of the (time-course of the) food blocks and the non-food blocks has been calculated. For TPICA, it is 0.3260, for CPD it is 0.2684 and for BTM, it is 0.3289. Although it seems that the ability to discriminate between food and non-food blocks in TPICA and BTM is quite similar (and better than CPD), this finding can be quite misleading. As it can be observed from Fig. 21 (where time-courses are presented detrended and low pass filtered with a Butterworth filter), the time-course from TPICA displays a large difference among the amplitudes of the first blocks. In contrast, in pairs 6 and 8, we can observe that the even block is higher than the odd block (hence, we can not discriminate the food or non-food blindly, based on the level of activation of the visual cortex). BTM is the only, among the

three methods, which results in higher amplitude in the odd (food) blocks in all of the pairs. For visualizing the components and computing the Z scores, the Display GUI of the GIFT toolbox [70] was used (and all of its post-processing tools; e.g., detrending and low pass filtering with a 5th-order Butterworth filter) and $Z > 2$ was selected. In all the methods, we can distinguish the main visual component. Two of the components for CPD are probably caused by motion and they have very high correlation with each other (they could be merged into a single component). Each method resulted in at least one component in which the reward regions can be observed. Since the component of interest is the visual cortex, we will focus on the comparison of the visual component obtained via the three methods (the rest of the components can be viewed at pix.sfly.com/bziPkBS8). In Figs. 22–25, the visual component of each method is depicted.¹¹ We note that the primary visual cortex is the main area activated and it is quite similar for all three methods with slightly higher activation for CPD and BTD (z slices -2 to -18). The main difference between CPD and BTD is the fact that, in CPD, the precuneus does not seem activated (z slices 14 to 42). It is activated in the visual component of TPICA but with significantly less power than BTD. The precuneus has been reported to be activated during food choice versus non-food choice [90] and is known for its involvement in attention and connected to visuospatial preprocessing and the recall of emotions [91]. Furthermore, it has been reported that it exhibits increased activation during inhibition versus imaginary eating [92]. ~~The visual component of GICA is characterized by the fact that is only activated at the lower slices and shows almost zero activation in the areas of the precuneus.~~

The visual component of BTD2 is shown in Fig. 26 (since the time-courses are different per subject, the common spatial map and the time-course of the first of the 18 subjects are only depicted) and can be noted that, similarly to BTD, the precuneus and the primary visual cortex are activated. We also note that Component 8 (Fig. 27) has a time-course with high correlation with the food blocks (the food peaks have been denoted with an F). The level of correlation differs per subject and the amplitude of the time-course is larger for subjects with higher appetite (based on the

¹¹The rank of some of the slices of the visual component of CPD is computed in the Supplementary Material, in order to demonstrate that the resulting regions are of low-rank.

appetite score provided with the metadata of the dataset). The spatial map of Component 8 indicates activation of the primary visual cortex, caudate nucleus, precuneus and the thalamus. The activation of the thalamus has been reported to be connected with the anticipation of food and/or drink consumption [93], and be higher in overweight and obese participants than healthy weight subjects [94]. This can be connected to the higher activation of Component 8 in subjects with higher appetite as observed. A possible drawback of BTD2 and PARAFAC2 (that has not been observed with simulated data) is the fact that the resulting components are corrupted by noise more than CPD and BTD. The fact that in some cases the time-courses obtained from non-multilinear models have such a noisy shape has been also previously reported [79]. This is an indirect effect of the weaker restrictions imposed on the time-courses. Nevertheless, since the time-courses follow the original profiles closely, this is not really an issue; a low pass filtering could overcome the problem. Furthermore, we observed that the reproducibility for R higher than 15 is not always achieved (which may indicate uniqueness problems for high values of R ; similar observations for PARAFAC2 have been made in [20]).

3.2.0.1. Augmented data

In real data, due to the lack of ground truth, it is difficult to identify which algorithm performs best and hence indirect comparisons are made. As an alternative way to make the comparisons we will create an “augmented” dataset. An extra simulated source (Fig. 28) is added to the real dataset, which has both temporal and spatial overlap with the primary visual component. The activity level at each active voxel was randomly sampled from Uniform [0.8,1.2] times the maximum activation of visual component of every subject, while the am-

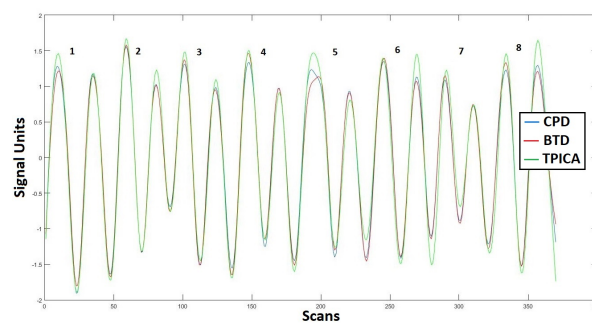


Figure 21: The time-courses of the visual component.

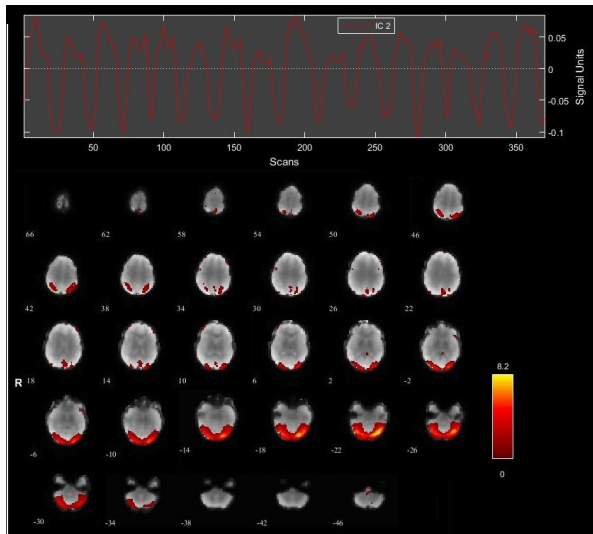


Figure 22: The visual component from TPICA, $R = 8$.

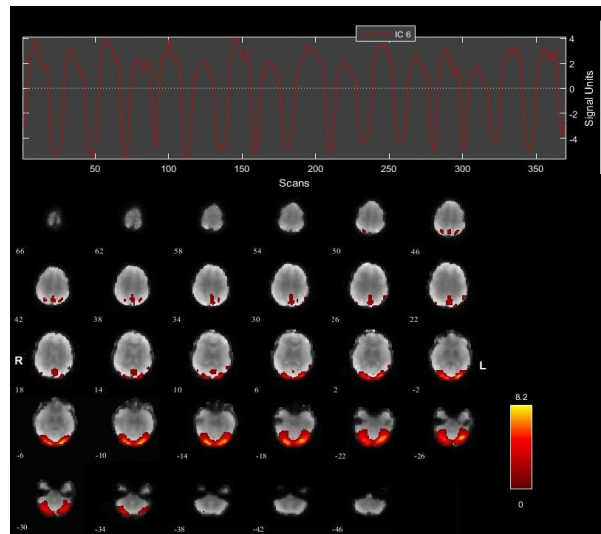


Figure 25: The visual component from BTM, $R = 8$.

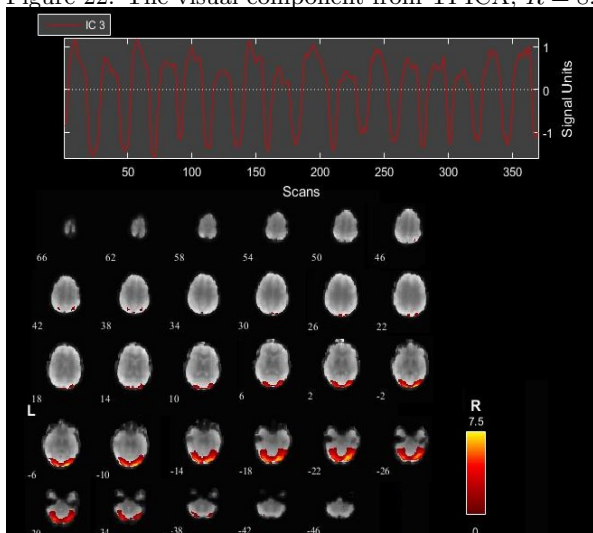


Figure 23: The visual component from GICA, $R = 8$.

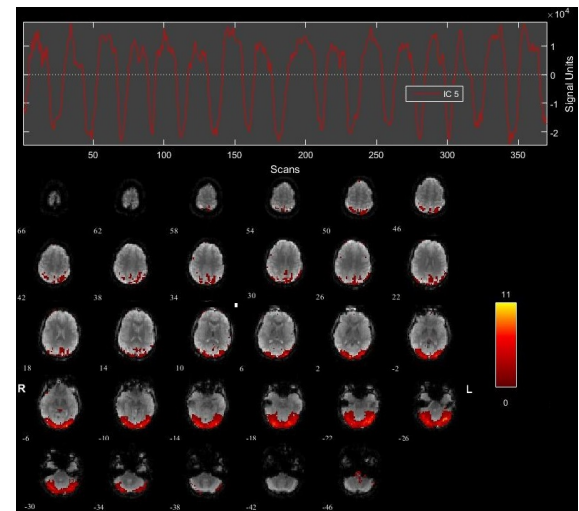


Figure 26: The visual component from BTM2, $R = 8$.

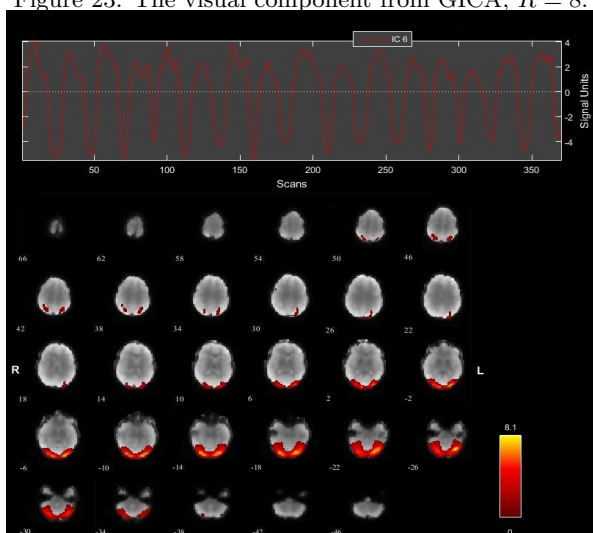


Figure 24: The visual component from CPD, $R = 8$.

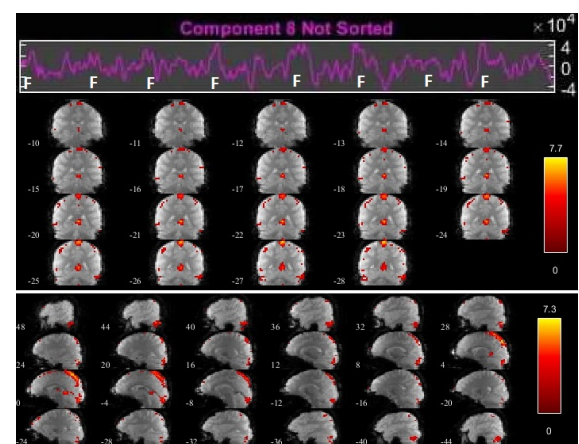


Figure 27: Component 8 from BTM2, $R = 8$.

plitude of the time-course is analogous to the visual component per subject. Ideally, the result for the sources originated from the real dataset should not be affected by the extra source and hence the sources obtained from each method with the initial dataset plus the simulated source can be considered as a ground truth.¹² Figs. 29–34 depict the results of the different decomposition methods. BTD separates almost perfectly the two sources and the correlation of the simulated source with the obtained source is 0.91. As expected, TPICA is the method that suffers most from the overlap with the simulated source. We can see that both the time-course of the visual component and the spatial map are contaminated from the artificial source. With CPD, the time-courses are recovered relatively well but the spatial maps have cross-talk with each other and the correlation of the obtained map with the ground truth for the spatial maps is only 0.72.

Similarly to the comparisons for BTD, augmented data will also be used for testing the non-multilinear methods. The source of Fig. 28 was also used in that case but with a significant difference. The time-course of the source is shifted randomly for every subject from -5 to +5 seconds (compared to the one used in Fig. 28). CPD fails in distinguishing correctly the two sources (Figs. 35 and 36) even if we increase the rank of the decomposition. On the other hand, BTD2 finds the spatial maps accurately (Figs. 37 and 38) but with some noise of

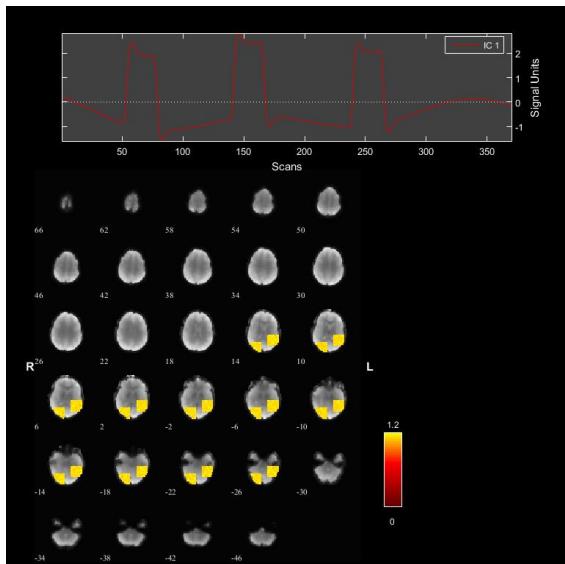


Figure 28: The simulated component added.

¹²An idea following the rationale of the super-position of pilots in communications.

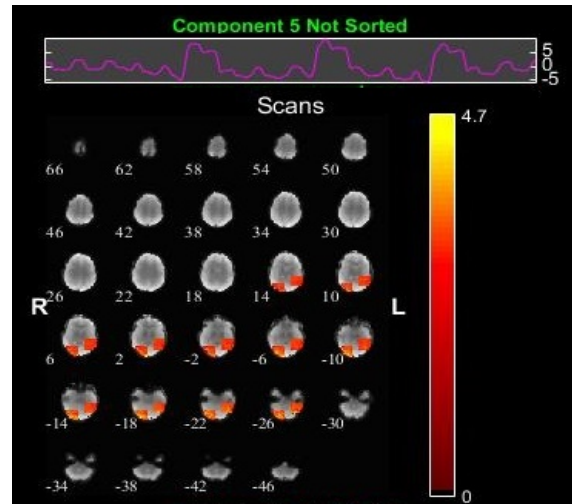


Figure 29: The simulated component from TPICA.

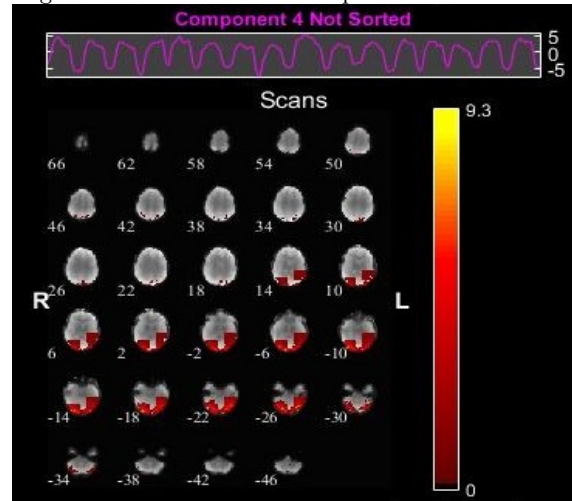


Figure 30: The visual component from TPICA.

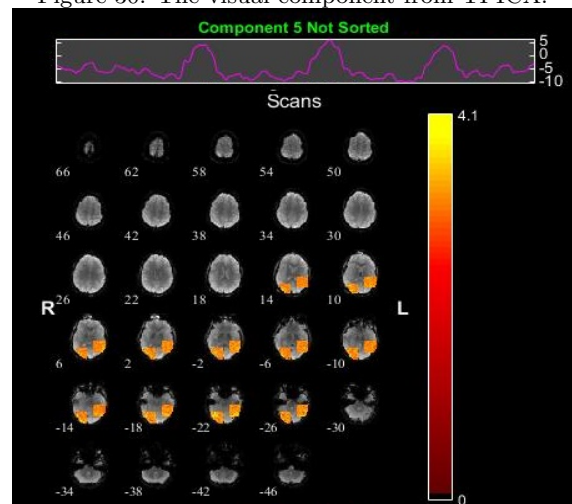


Figure 31: The simulated component from CPD.

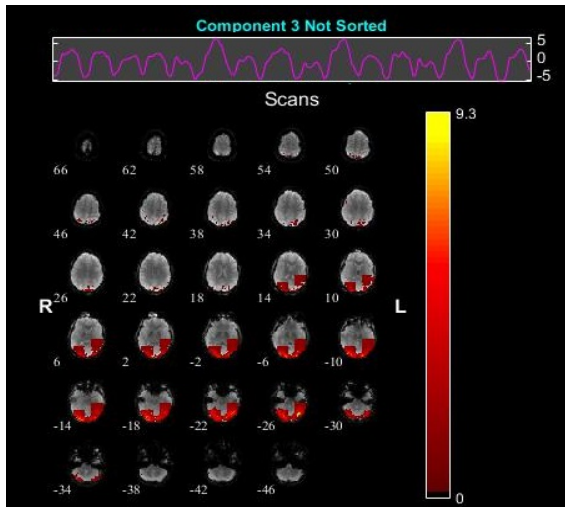


Figure 32: The visual component from CPD.

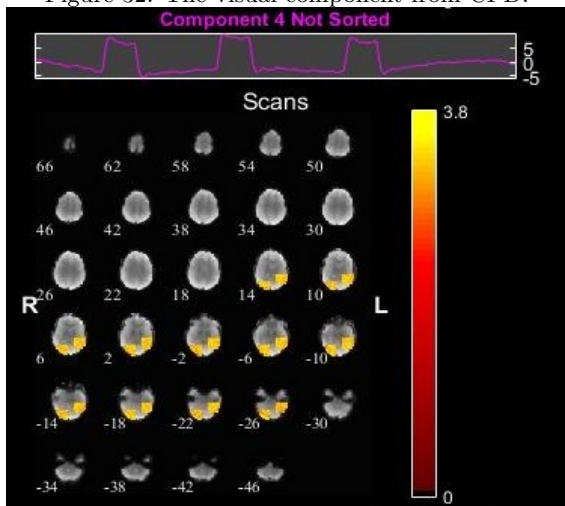


Figure 33: The simulated component from BTM.

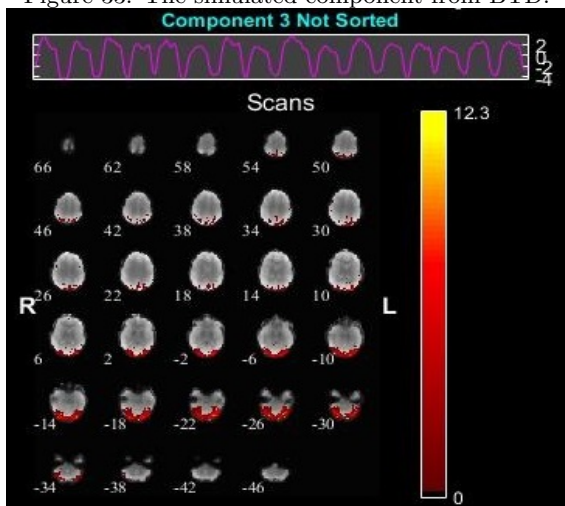


Figure 34: The visual component from BTM.

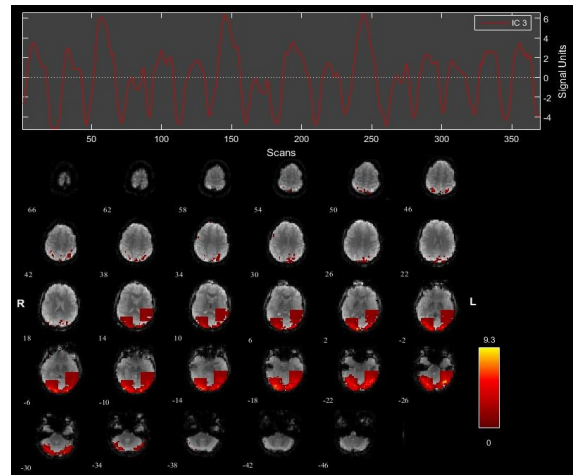


Figure 35: The visual component from CPD for the augmented data, with different HRF per subject.

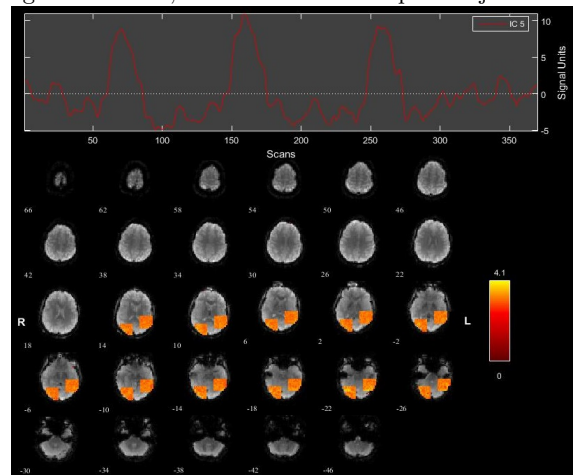


Figure 36: The simulated component from CPD for the augmented data, with different HRF per subject.

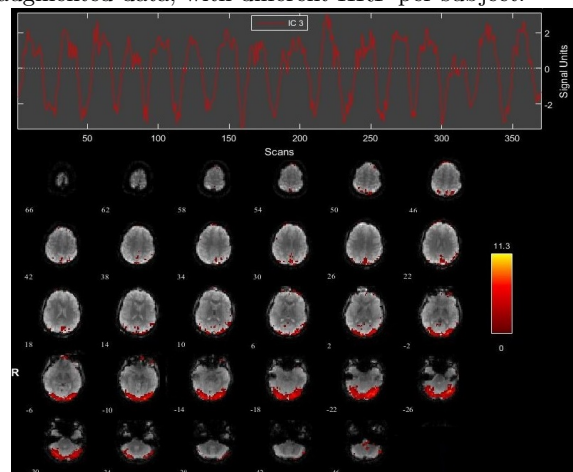


Figure 37: The visual component from BTM2 for the augmented data, with different HRF per subject.

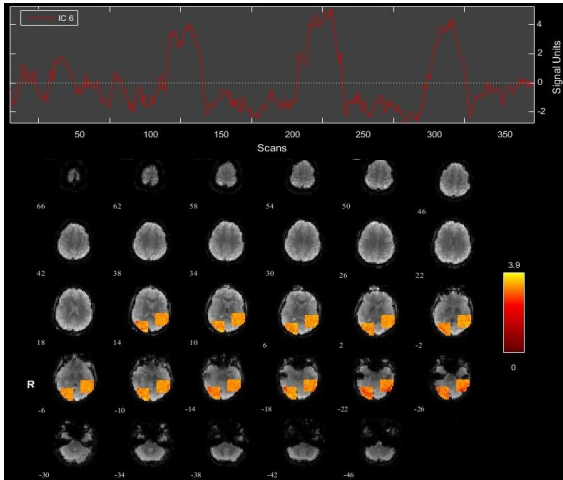


Figure 38: The simulated component from BTDD2 for the augmented data, with different HRF per subject.

high frequency in the time-courses (as previously noted). The constraint that the cross products $\mathbf{B}_k^\top \mathbf{B}_k$ are constant over k (as required in PARAFAC2 and BTDD2) is not satisfied in this case (since we only shift one of the sources). However, since this constraint is less strict than \mathbf{B}_k being the same for every subject, the performance of BTDD2 is still better than that of BTDD (see [15] for similar findings in a chemometrics application).

4. Conclusions

In this paper, a novel approach to blind fMRI source unmixing was presented. It is based on the combined use of higher (than three)-order tensors and a BTDD tensor model, in an effort to better exploit the original 3-D spatial structure of the data. Extensive simulation results demonstrated the enhanced robustness of the proposed method to the presence of noise, as compared with CPD- and ICA-based decompositions. In cases of spatial and/or temporal overlap, both CPD and BTDD give better results compared to ICA-based methods [8, 7]. **SICA is the only among the ICA-based methods, which exhibits good performance (albeit careful tuning) in the case of overlap when only sparse sources exist; on the other hand, problems arise when dense and sparse sources coexist.** CPD and BTDD are more sensitive to the tensor rank overestimation. In terms of computational cost, the complexity of BTDD is higher than that of CPD. A heuristic was proposed for the estimation of L for the fMRI source separation problem.

Furthermore, a new decomposition method, called BTDD2, was proposed, which is based on

BTDD while allowing variation across one mode. It was demonstrated in simulated scenarios that non-multilinear tensor decomposition methods, such as PARAFAC2 and BTDD2, are expected to be more suitable for fMRI BSS due to the variability of the HRF across subjects and that they result in improved separation performance compared to strictly multilinear methods like CPD and TPICA. The proposed tensor decomposition, BTDD2, combined with the use of higher-order tensors, shows significant robustness to noise compared to PARAFAC2, at the cost of a higher computational complexity. 13

Our approach, based on higher-order tensorization, has also some further important advantages in practical applications. In contrast to statistical approaches (notably ICA and IVA), BTDD only relies on the low rank assumption. No statistical assumptions (such as independence) nor statistics estimates (such as of correlations or cumulants) are necessary. The only reliance of the tensor decomposition methods (such as CPD and BTDD) on statistics is the assumption of additive white Gaussian noise, which underlies the least squares criterion that was chosen for their computation. On the other hand, ICA relies on the statistical independence of the sources and that is why it fails in scenarios involving strong spatial overlap among them.

In task-related fMRI, external information is often available. The time-courses of some of the sources of activation may be approximately known and/or indications about the areas, which are expected to be activated, may exist. Thus, future work may explore semi-blind (guided) versions of BTDD and BTDD2 [27], which will allow prior information to be exploited. Developing methods for estimating R and L , specific to the fMRI BSS problem, is another possible future research direction. Following a similar rationale to that of Heuristic 1, an optimization algorithm, using column pruning, can be sought for the estimation of L . Similar optimization techniques, based on the nuclear norm, have been used for the estimation of R for CPD [95].

Acknowledgments

The authors would like to thank Dr. A. Stegeman (<http://www.alwinstegeman.nl/>) and Prof. N. Helwig, Univ. of Minnesota for providing the

¹³The code for the algorithms used and for reproducing the figures is available at <https://github.com/chrichat/BTDD>.

datasets used in [7] and [8], respectively, Prof. S. Van Huffel, KU Leuven for her critical comments on earlier versions of this paper, Dr. N. Andreadis, Vioiatriki for his help in the interpretation of the results in the real data scenario and Y. Kopsinis, Libra MLI for fruitful discussions on the topic of BSS in fMRI. This research has been funded by the European Union’s Seventh Framework Programme (H2020-MSCA-ITN-2014) under grant agreement No. 642685 MacSeNet. Data used in Section 3.2.2.2 were provided by the Human Connectome Project, MGH-USC Consortium (Principal Investigators: Bruce R. Rosen, Arthur W. Toga and Van Wedeen; U01MH093765) funded by the NIH Blueprint Initiative for Neuroscience Research grant; the National Institutes of Health grant P41EB015896; and the Instrumentation Grants S10RR023043, 1S10RR023401, 1S10RR019307.

5. References

- [1] M. A. Lindquist, “The statistical analysis of fMRI data,” *Statistical Science*, vol. 23, no. 4, pp. 439–464, Jun. 2008.
- [2] S. Theodoridis, *Machine Learning: A Bayesian and Optimization Perspective*. Academic Press, 2015.
- [3] A. H. Andersen and W. S. Rayens, “Structure-seeking multilinear methods for the analysis of fMRI data,” *NeuroImage*, vol. 22, no. 2, pp. 728–739, Jun. 2004.
- [4] V. D. Calhoun and T. Adali, “Unmixing fMRI with independent component analysis,” *IEEE Eng. Med. Biol. Mag.*, vol. 25, no. 2, pp. 79–90, Mar. 2006.
- [5] L. Kangjoo, S. Tak, and J. C. Ye, “A data-driven sparse GLM for fMRI analysis using sparse dictionary learning with MDL criterion,” *IEEE Trans. Med. Imag.*, vol. 30, no. 5, pp. 1076–1089, Dec. 2011.
- [6] N. Sidiropoulos and R. Bro, “On the uniqueness of multilinear decomposition of N-way arrays,” *J. Chemometrics*, vol. 14, no. 3, pp. 229–239, May 2000.
- [7] A. Stegeman, “Comparing Independent Component Analysis and the PARAFAC model for artificial multi-subject fMRI data,” Unpublished Technical Report, Univ. of Groningen, Feb. 2007.
- [8] N. E. Helwig and S. Hong, “A critique of Tensor Probabilistic Independent Component Analysis: implications and recommendations for multi-subject fMRI data analysis,” *J. Neuroscience Methods*, vol. 213, no. 2, pp. 263–273, Mar. 2013.
- [9] C. F. Beckmann and S. M. Smith, “Tensorial extensions of independent component analysis for multisubject fMRI analysis,” *NeuroImage*, vol. 25, no. 1, pp. 294–311, Mar. 2005.
- [10] D. A. Handwerker, J. Ollinger, and M. Esposito, “Variation of BOLD hemodynamic responses across subjects and brain regions and their effects on statistical analyses,” *NeuroImage*, vol. 21, no. 4, pp. 1639–1651, Apr. 2004.
- [11] L. De Lathauwer, “Decompositions of a higher-order tensor in block terms—Part I: Lemmas for partitioned matrices,” *SIAM J. Matrix Anal. Appl.*, vol. 30, no. 3, pp. 1022–1032, Sep. 2008.
- [12] L. De Lathauwer, “Decompositions of a higher-order tensor in block terms—Part II: Definitions and uniqueness,” *SIAM J. Matrix Anal. Appl.*, vol. 30, no. 3, pp. 1033–1066, Sep. 2008.
- [13] L. De Lathauwer and D. Nion, “Decompositions of a higher-order tensor in block terms—Part III: Alternating least squares algorithms,” *SIAM J. Matrix Anal. Appl.*, vol. 30, no. 3, pp. 1067–1083, Sep. 2008.
- [14] R. A. Harshman, “PARAFAC2: Mathematical and technical notes,” *UCLA Working Papers in Phonetics*, vol. 22, pp. 30–44, 1972.
- [15] R. Bro, C. Andresson, and H. Kiers, “PARAFAC2 Part II: Modeling chromatographic data with retention time shifts,” *J. Chemometrics*, vol. 13, no. 3, pp. 295–309, May 1999.
- [16] S. Ferdowsi, V. Abolghasemi, and S. Sanei, “A new informed tensor factorization approach to EEG–fMRI fusion,” *J. Neuroscience Methods*, vol. 254, no. 1, pp. 27–35, Oct. 2015.
- [17] K. Madsen, N. W. Churchill, and M. Mørup, “Quantifying functional connectivity in multi-subject fMRI data using component models,” *Human Brain Mapping*, vol. 38, no. 2, pp. 882–899, Oct. 2016.
- [18] H. A. Kiers, J. T. Berge, and R. Bro, “PARAFAC2 Part I: A direct fitting algorithm for the PARAFAC2 model,” *J. Chemometrics*, vol. 13, no. 3, pp. 275–294, May 1999.
- [19] A. Stegeman and T. Lam, “Multi-set factor analysis by means of PARAFAC2,” *British J. Mathematical and Statistical Phys.*, vol. 69, no. 3, pp. 1–19, Nov. 2015.
- [20] L. Spyrou, M. Parra, and J. Escudero, “Complex tensor factorization with PARAFAC2 for the estimation of brain connectivity from the EEG,” arXiv preprint:1705.02019v1, May 2017.
- [21] G. Goovaerts, B. Vanderbeck, R. Willems, and S. Van Huffel, “Automatic detection of T wave alternans using tensor decompositions in multilead ECG signals,” *Physiol. Meas.*, vol. 38, no. 8, pp. 1513–1518, Jul. 2017.
- [22] C. Chatzichristos, E. Kofidis, Y. Kopsinis, M. M. Moreno, and S. Theodoridis, “Higher-order block term decomposition for spatially folded fMRI data,” in *Latent Variable Analysis and Signal Separation Conf. (LVA/ICA)*, Grenoble, France, Feb. 2017.
- [23] C. Chatzichristos, E. Kofidis, and S. Theodoridis, “PARAFAC2 and its block term decomposition analog for blind fMRI source unmixing,” in *Eur. Signal Process. Conf. (EUSIPCO)*, Kos, Greece, Aug.–Sep. 2017.
- [24] V. D. Calhoun and T. Adali, “A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data,” *NeuroImage*, no. 1, pp. 163–172, Mar. 2009.
- [25] V. D. Calhoun, T. Adali, T. Pearlson, and J. Pekar, “A method for making group inferences from functional MRI data using Independent Component Analysis,” *Human Brain Mapp.*, vol. 14, no. 3, pp. 140–151, Nov. 2001.
- [26] Y. Kopsinis, H. Georgiou, and S. Theodoridis, “fMRI unmixing via properly adjusted dictionary learning,” in *Eur. Signal Process. Conf. (EUSIPCO)*, Lisbon, Portugal, Sep. 2014.
- [27] M. Moreno, Y. Kopsinis, E. Kofidis, C. Chatzichristos, and S. Theodoridis, “Assisted dictionary learning for fMRI data analysis,” in *IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, New Orleans,

- USA, Mar. 2017.
- [28] Z. Boukouvalas, Y. Levin-Schwarz, V. D. Calhoun, and T. Adahi, “Sparsity and independence: Balancing two objectives in optimization for source separation with application to fMRI analysis.” *J. Franklin Inst.*, vol. 355, no. 4, pp. 1873–1887, Mar. 2018.
- [29] K. Petersen, L. Hansen, T. Kolenda, E. Rostrup, and S. Strother, “On the independent components of functional neuroimages,” in *Int. Conf. Ind. Comp. Anal. and Blind Signal Separ. (ICA)*, Helsinki, Finland, Jun. 2000.
- [30] J. Lee, T. Lee, F. Jolesz, and S. Yoo, “Independent Vector Analysis (IVA): Multivariate approach for fMRI group study,” *NeuroImage*, vol. 40, no. 1, pp. 86–109, Mar. 2008.
- [31] T. Adahi, M. Anderson, and G. S. Fu, “Diversity in Independent Component and Vector Analyses: Identifiability, algorithms, and applications in medical imaging,” *IEEE Signal Process. Mag.*, vol. 31, no. 3, pp. 18–34, Apr. 2014.
- [32] O. Debals and L. De Lathauwer, “Stochastic and deterministic tensorization for blind signal separation,” in *Latent Variable Analysis and Signal Separation Conf. (LVA/ICA)*, Liberec, Czech Republic, Aug. 2015.
- [33] B. Hunyadi, D. Camps, L. Sorber, W. Van Paesschen, M. De Vos, S. Van Huffel, and L. De Lathauwer, “Block term decomposition for modelling epileptic seizures,” *EURASIP J. Adv. Signal Process.*, 2014, doi:10.1186/1687-6180-2014-139.
- [34] A. Cichocki, D. Mandic, L. De Lathauwer, G. Zhou, Q. Zhao, C. Caiafa, and A. H. Phan, “Tensor decompositions for signal processing applications: From two-way to multiway component analysis,” *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 145–163, Mar. 2015.
- [35] B. Hunyadi, P. Dupont, W. Van Paesschen, and S. Van Huffel, “Tensor decompositions and data fusion in epileptic electroencephalography and functional magnetic resonance imaging data,” *WIREs Data Mining Knowl. Discov.*, vol. 7, no. 1, Feb. 2017.
- [36] P. Aggarwal and A. Gupta, “Accelerated fMRI reconstruction using matrix completion with sparse recovery via split Bregman,” *Neurocomputing*, vol. 16, no. 216, pp. 319–330, Aug. 2016.
- [37] T. Zhang, M. Pham, J. Sun, G. Yan, H. Li, Y. Sun, M. Gonzalez, and J. Coan, “A low-rank multivariate general linear model for multi-subject fMRI data and a non-convex optimization algorithm for brain response comparison,” *NeuroImage*, vol. 173, no. 2, pp. 580–591, Jun. 2018.
- [38] G. Varoquaux, A. Gramfort, F. Pedregosa, V. Michel, and B. Thirion, “Multi-subject dictionary learning to segment an atlas of brain spontaneous activity,” in *Inf. Process. Med. Imaging (IPMI)*, Kloster Irsee, Germany, Jul. 2011.
- [39] Wikipedia, “Wernicke’s area.” [Online]. Available: https://en.wikipedia.org/wiki/Wernicke%27s_area
- [40] A. H. Phan, P. Tichavský, and A. Cichocki, “CANDECOMP/PARAFAC decomposition of high-order tensors through tensor reshaping,” *IEEE Trans. Signal Process.*, vol. 61, no. 19, pp. 4847–4860, Oct. 2013.
- [41] P. Tichavský, A. H. Phan, and Z. Koldovsky, “Cramér-Rao-induced bounds for CANDECOMP/PARAFAC tensor decomposition,” *IEEE Trans. Signal Process.*, vol. 61, no. 8, pp. 1986–1997, Oct. 2013.
- [42] L. Norgaard, “Classification and prediction of quality and process parameters of thick juice and beet sugar by fluorescence spectroscopy and chemometrics,” *Zuckerindustrie*, vol. 120, no. 1, 1995.
- [43] M. Boussé, O. Debals, and L. De Lathauwer, “A tensor-based method for large-scale blind source separation using segmentation,” *IEEE Trans. Signal Process.*, vol. 65, no. 2, pp. 346–358, Jan. 2017.
- [44] N. Correa, T. Adahi, Y.-O. Li, and V. D. Calhoun, “Comparison of blind source separation algorithms for fMRI using a new Matlab toolbox: GIFT,” in *IEEE Int. Conf. on Acoustics, Speech, and Signal Process. (ICASSP)*, Philadelphia, USA, Mar. 2005.
- [45] R. A. Harshman, “Foundations of the PARAFAC procedure: Models and conditions for an “explanatory” multi-modal factor analysis,” *UCLA Work. Papers in Phonetics*, pp. 1–84, 1970.
- [46] J. B. Kruskal, “Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics,” *Linear Algebra and Its Applications*, vol. 18, pp. 95–138, 1977.
- [47] R. Bro and H. Kiers, “A new efficient method for determining the number of components in PARAFAC models,” *J. Chemometrics*, vol. 17, no. 5, pp. 274–286, May 2003.
- [48] J. L. Castellanos, S. Gómez, and V. Guerra, “The triangle method for finding the corner of the L-curve,” *Applied Numerical Mathematics*, vol. 43, no. 4, pp. 359–373, Dec. 2002.
- [49] C. F. Beckmann and S. M. Smith, “Probabilistic independent component analysis for functional magnetic resonance imaging,” *IEEE Trans. Med. Imag.*, vol. 23, no. 2, pp. 137–152, Feb. 2004.
- [50] M. Jenkinson, C. F. Beckmann, T. E. Behrens, M. W. Woolrich, and S. M. Smith, “FSL,” pp. 782–790. [Online]. Available: <https://fsl.fmrib.ox.ac.uk/fsl/fslwiki>
- [51] M. De Vos, M. Nion, S. Van Huffel, and L. De Lathauwer, “A combination of parallel factor and independent component analysis,” *Signal Processing*, vol. 92, no. 12, pp. 2990 – 2999, 2012.
- [52] G. Favier and A. L. F. de Almeida, “Overview of constrained PARAFAC models,” *EURASIP J. Adv. Signal Process.*, 2014, doi:10.1186/1687-6180-2014-142.
- [53] L. De Lathauwer, “Block component analysis: A new concept for blind source separation,” in *Latent Var. Anal. and Signal Sep. Conf. (LVA/ICA)*, Tel Aviv, Israel, Mar. 2012.
- [54] J. C. Martinez-Trujillo, J. K. Tsotsos, E. Simine, M. Pomplun, R. Wildes, S. Treue, H. Heinze, and J. Hopf, “Selectivity for speed gradients in human area MT/V5,” *NeuroReport*, vol. 16, no. 5, pp. 435–438, Apr. 2005.
- [55] A. H. Phan, A. Cichocki, R. Zdunek, and S. Lehky, “From basis components to complex structural patterns,” in *IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, Vancouver, Canada, May 2013.
- [56] L. N. Ribeiro, A. Almeida, and V. Zarzoso, “Enhanced block term decomposition for atrial activity extraction in atrial fibrillation ECG,” in *IEEE Workshop on Sensor Array and Multich. Signal Process. (SAM)*, Rio de Janeiro, Brazil, Jul. 2016.
- [57] A. Brockmeier, J. Principe, F. Gainesville, A. Phan, and A. Cichocki, “A greedy algorithm for model selection of tensor decompositions,” in *IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, Van-

- couver, Canada, May 2013.
- [58] V. Zarzoso, "Parameter estimation in block term decomposition for noninvasive atrial fibrillation analysis," in *Int. Workshop on Comput. Adv. in Multi-Sensor Adapt. Process. (CAMSAP)*, Curacao, Dutch Antilles, Dec. 2017.
- [59] Y. Qian, F. Xiong, S. Zeng, J. Zhou, and Y. Y. Tan, "Matrix-vector nonnegative tensor factorization for blind unmixing of hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 3, pp. 71–79, Mar. 2017.
- [60] Wikipedia, "Family-wise error rate." [Online]. Available: https://en.wikipedia.org/wiki/Family-wise_error_rate
- [61] Wikipedia, "Bonferroni correction." [Online]. Available: https://en.wikipedia.org/wiki/Bonferroni_correction
- [62] D. Brie, S. Miron, F. Caland, and C. Mustin, "An uniqueness condition for the 4-way CANDECOMP/PARAFAC model with collinear loadings in three modes," in *IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, Prague, Czech Rep., May 2011.
- [63] L. Sorber, M. V. Barel, and L. D. Lathauwer, "Optimization-based algorithms for tensor decompositions: Canonical Polyadic Decomposition, decomposition in rank- $(L_r, L_r, 1)$ terms, and a new generalization," *SIAM J. Optim.*, vol. 23, no. 2, pp. 695–720, Apr. 2013.
- [64] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Review*, vol. 51, no. 3, pp. 455–500, 2009.
- [65] I. Perros, E. Papalexakis, F. Wang, R. Vuduc, E. Searles, and J. Sun, "SPARTan: scalable PARAFAC2 for large and sparse data," in *Knowl. Discovery and Data Mining Conf. (KDD)*, Halifax, Canada, Aug. 2017.
- [66] J. T. Berge and H. Kiers, "Some uniqueness results for PARAFAC2," *Psychometrika*, vol. 61, no. 1, pp. 123–132, Feb. 1996.
- [67] R. A. Harshman and M. Lundy, "Uniqueness proof for a family of models sharing features of Tucker's three-mode factor analysis and PARAFAC/CANDECOMP," *Psychometrika*, vol. 61, no. 1, pp. 133–154, Mar. 1996.
- [68] A. J. Bell and T. J. Sejnowski, "An information maximisation approach to blind separation and blind deconvolution," *Neural Comput.*, vol. 7, no. 6, pp. 1129–1159, Nov. 1995.
- [69] C. F. Beckmann, C. E. Mackay, N. Filippini, and S. M. Smith, "Group comparison of resting-state fMRI data using multi-subject ICA and dual regression," in *Org. for Human Brain Mapping (OHBM)*, San Francisco, USA, Jun. 2009.
- [70] Medical Image Analysis Lab (MIALAB), "Group ICA of fMRI Toolbox (GIFT)." [Online]. Available: <http://mialab.mrn.org/software/gift/index.html>
- [71] Q. Long, S. Bhinge, Y. Levin-Schwartz, Z. Boukouvalas, V. D. Calhoun, and T. Adalı, "The role of diversity in data-driven analysis of multi-subject fMRI data: Comparison of approaches based on independence and sparsity using global performance metrics," *Human Brain Mapp.*, pp. 1–16, Sep. 2018.
- [72] Z. Boukouvalas and T. Adalı, "Sparse independent component analysis," 2018. [Online]. Available: http://mlsp.umbc.edu/SparseICA_EBM.html
- [73] N. Vervliet, O. Debals, L. Sorber, M. Van Barel, and L. De Lathauwer, "Tensorlab 3.0," Mar. 2016. [Online]. Available: <http://www.tensorlab.net>
- [74] C. A. Andersson and R. Bro, "The N-way toolbox for MATLAB," *J. Chem. and Intell. Lab. Systems*, vol. 52, no. 1, pp. 1–4, Aug. 2000.
- [75] H. Gøvert, J. Hurri, J. Sørøelø, and A. Hyvärinen, "Fastica package for Matlab," 2016. [Online]. Available: <http://research.ics.aalto.fi/ica/fastica/>
- [76] E. Sanchez and B. R. Kowalski, "Tensorial resolution: A direct trilinear decomposition," *J. Chemometrics*, vol. 4, no. 1, pp. 29–45, Jan. 1990.
- [77] R. Bro, R. Leardi, and L. G. Johnsen, "Solving the sign indeterminacy for multiway models," *J. of Chemometrics*, vol. 27, no. 3–4, pp. 70–75, Mar. 2013.
- [78] N. Vervliet, O. Debals, L. Sorber, and L. De Lathauwer, "Breaking the curse of dimensionality using decompositions of incomplete tensors: Tensor-based scientific computing in big data analysis," *IEEE Signal Process. Mag.*, vol. 31, no. 5, pp. 71–79, Aug. 2014.
- [79] R. Bro, "PARAFAC: Tutorial and applications," *Chemometrics and Intelligent Laboratory systems*, vol. 38, no. 2, pp. 149–171, 1997.
- [80] E. B. Erhardt, E. A. Allen, Y. Wei, T. Eichele, and V. D. Calhoun, "SimTB, a simulation toolbox for fMRI data under a model of spatiotemporal separability," *NeuroImage*, vol. 59, no. 4, pp. 4160–4167, Feb. 2012.
- [81] Y.-O. Li, T. Adalı, and V. D. Calhoun, "Estimating the number of independent components for functional magnetic resonance imaging data," *Human Brain Mapping*, vol. 28, no. 11, pp. 1251–1266, Nov. 2007.
- [82] T. Adalı, "Simulating fMRI-like sources," 2016. [Online]. Available: <http://mlsp.umbc.edu/resources.html>
- [83] A. M. Michael, M. Anderson, R. L. Miller, T. Adalı, and V. D. Calhoun, "Preserving subject variability in group fMRI analysis: performance evaluation of GICA vs. IVA," *Front. in Syst. Neurosc.*, vol. 8, Jun. 2014.
- [84] E. B. Erhardt, S. Rachakonda, E. J. Bedrick, E. A. Allen, T. Adalı, and V. D. Calhoun, "Comparison of multi-subject ICA methods for analysis of fMRI data," *Human Brain Mapping*, vol. 32, no. 12, pp. 2075–2095, 2011.
- [85] "OpenfMRI." [Online]. Available: <https://openfmri.org>
- [86] P. A. Smeets, F. M. Kroese, C. Evers, and D. T. de Ridder, "Allured or alarmed: counteractive control responses to food temptations in the brain," *Behavioural Brain Research*, vol. 248, no. 1, pp. 41–45, Jul. 2013.
- [87] M. Jenkinson, P. Bannister, J. Brady, and S. Smith, "Improved optimisation for the robust and accurate linear registration and motion correction of brain image," *NeuroImage*, vol. 17, no. 2, pp. 825–841, Dec. 2002.
- [88] S. Smith, "Fast robust automated brain extraction," *Human Brain Mapping*, vol. 17, no. 3, pp. 143–155, Nov. 2002.
- [89] E. Papalexakis, "Automatic unsupervised tensor mining with quality assessment," in *SIAM Int. Conf. Data Mining (SDM)*, Miami, USA, May. 2016.
- [90] L. Charbonnier, L. N. Van der Laan, M. Viergever, and P. A. Smeets, "Functional MRI of challenging food choices: Forced choice between equally liked high- and low-calorie foods in the absence of hunger," *PLoS ONE*, vol. 10, no. 7, Jul. 2015.
- [91] A. E. Cavanna and M. R. Trimble, "The precuneus: a review of its functional anatomy and behavioural correlates," *Brain*, vol. 129, no. 3, pp. 564–583, Jan. 2006.
- [92] J. Tuulari, H. Karlsson, J. Hirvonen, P. Salminen,

- P. Nuutila, and L. Nummenmaa, “Neural circuits for cognitive appetite control in healthy and obese individuals: An fMRI study,” *PLoS ONE*, vol. 2, Feb. 2015.
- [93] E. Stice and S. Spoor, “Relation of reward from food intake and anticipated food intake to obesity: A functional magnetic resonance imaging study,” *J. Abnorm. Psychol.*, vol. 117, no. 4, pp. 924–935, Nov. 2008.
- [94] K. M. Pursey, P. Stanwell, R. j. Callister, K. Brain, C. E. Collins, and T. L. Burrows, “Neural responses to visual food cues according to weight status: A systematic review of functional magnetic resonance imaging studies,” *Frontiers in Nutrition*, vol. 7, no. 1, Jul. 2014.
- [95] I. Tsaknakis, P. Giampouras, A. Rontogiannis, and K. Koutroumbas, “A computationally efficient tensor completion algorithm,” *IEEE Signal Process. Letters*, vol. 25, no. 8, pp. 1266–1270, Jul. 2018.