Citation:
T. Adali, F. Kantar, M. A. B. S. Akhonda, S. Strother, V. D. Calhoun and E. Acar, "Reproducibility in Matrix and Tensor Decompositions: Focus on model match, interpretability, and uniqueness," in IEEE Signal Processing Magazine, vol. 39, no. 4, pp. 8-24, July 2022, doi: 10.1109/MSP.2022.3163870.

DOI:
https://doi.org/10.1109/MSP.2022.3163870

Reproducibility in Matrix and Tensor Decompositions: Focus on Model Match, Interpretability, and Uniqueness

Tülay Adali, Furkan Kantar, and M. A. B. Siddique Akhonda
Department of CSEE, University of Maryland, Baltimore County, Baltimore, MD, USA

Stephen Strother
Rotman Research Center, Baycrest, and Department of Medical Biophysics, University of Toronto, ON, Canada

Vince D. Calhoun
Tri-institutional Center for Translational Research in Neuroimaging and Data Science (TReNDS), Georgia State University, Georgia Institute of Technology, Emory University, Atlanta, GA

Evrim Acar
Simula Metropolitan Center for Digital Engineering, Oslo, Norway

1 Introduction and motivation

Data-driven solutions are playing an increasingly important role in numerous practical problems across multiple disciplines. This shift away from the traditional model-driven approaches to those that are data driven naturally emphasized the importance of explainability of our solutions, as in this case, the connection to a physical model is often not obvious. Explainability is a broad umbrella and includes interpretability but also implies that the solutions need to be complete, in that one should be able to "audit" them, ask appropriate questions and hence gain further insight about their inner workings [1]. Thus, interpretability, reproducibility, and ultimately, our ability to generalize these solutions to unseen scenarios and situations are all strongly tied to the starting point of explainability.

Model-based solutions, through their natural connection to the physical model, are fully interpretable. As data-driven solutions, matrix and tensor decompositions (MTD) [1] provide an attractive middle ground. While allowing for discovery of structure in the data in an unsupervised manner, they can also result in fully interpretable solutions, by which we mean we can associate the rows/columns of the final factor matrices with (physical) quantities of interest. In other data-driven solutions like multilayered neural networks, interpretability takes an indirect form and generally requires a second-level analysis, e.g., generation of heat maps in multilayered neural networks [2]. In MTD, interpretability is direct due to their intimate connection to the linear blind source separation problem, where the assumption is that there are a number of linearly mixed latent variables of interest. This assumption has proven useful in an array of applications, and MTD are being adopted across multiple domains such as medical image analysis and fusion, healthcare, remote sensing, chemometrics, metabolomics, recommender systems, natural language processing, and physical sciences [3–10].

While this rapidly growing interest in the development and use of factorization methods is very encouraging, a serious concern is the lack of formalizations such as a "reproducibility checklist", which have been developed for supervised learning, see e.g., [11]. The basics of reproducibility in this checklist that relate to the description and specification of models, algorithm, datasets, and the code are common in all machine learning approaches including the MTD. However, the last group of items in the checklist of [11] about reported experimental results, i.e., *computational reproducibility*, has not been emphasized for unsupervised learning, and for the MTD. While many success stories are reported using factorization

[1] We use matrices for data arranged as 2-way arrays, and tensors for N-way arrays where N>2.

of given set of observations across application areas, e.g., noting identification of putative biomarkers of disease in the analysis of neuroimaging data, in many instances, results are reported without much consideration to their computational reproducibility. The literature that acknowledges this important topic for MTD is currently rather limited and sparse, see e.g., [12–17]. Thus, our goal in this special issue paper is to identify critical issues for guaranteeing the reproducibility of solutions in unsupervised matrix and tensor factorizations, with an applied focus, considering the practical case where there is no ground truth. While simulation results can easily demonstrate advantages of a given model and support the given theoretical development, when the model parameters are not known—the case in most practical problems—their estimation and performance evaluation is a difficult problem. In this paper, we review the—*currently rather limited*—literature on the topic, discuss the proposed solutions, make suggestions based on those, and identify topics that require attention and further research.

We adopt the definition for (computational) reproducibility given in the report by the US National Academies of Sciences, Engineering, and Medicine [18] where it is defined as *obtaining consistent results using the same data and code*—i.e., the method as in the original study. In the same report, the closely related concept of replicability is defined as obtaining consistent results across studies aimed at answering the same scientific question using new data, which addresses generalization, a key topic in machine learning. Given that the cost functions for most matrix and tensor decomposition methods are non-convex, one can only guarantee convergence to a local optimum. Since in most cases, closed form solutions do not exist, iterative techniques are employed, and most commonly, using random initializations [15–17, 19, 20]. Even when all algorithmic quantities are fixed, and the only variability is due to random initializations, the resulting decompositions can be quite different as we also show with numerical examples in this paper. Still, in most of the literature reporting results with real data, a mechanism for selecting a single, *"most reproducible"* run from multiple runs has been rarely defined and used. In many instances, results are reported using a single run of the selected iterative algorithm. Hence, presenting a comprehensive review of the current solutions along with a set of suggested best practices, we believe is very important. Within the limitations of special issue papers in terms of length, we also touch upon the important related concepts, especially replicability.

When the emphasis is on interpretability in a given solution, it is important to have theoretical guarantees for having one set of factor matrices (subject to ambiguities) for a given dataset, i.e., to ensure the uniqueness of the model. Two decompositions that have proven useful in an array of applications and have uniqueness guarantees under relaxed conditions are the independent component analysis (ICA) and the canonical-polyadic decomposition (CPD). ICA poses the problem as that of separation of statistically independent latent variables (sources) with a matrix representation and provides a unique solution under very general conditions. On the other hand, CPD takes the multi-linear structure of the data into account with non-negative CPD (N-CPD) providing *almost always* uniqueness guarantees [21]. To introduce and solidify the important concepts of model match for reproducibility, we select two practical problems to demonstrate these ideas: analysis of functional magnetic resonance imaging (fMRI) data for ICA and unmixing of fluorescence spectroscopy data for N-CPD. We make use of these two applications to explain the important property of *model match*, which provides the link between interpretability and uniqueness. Besides, with better model match, the reproducibility of solutions tends to increase as we demonstrate. While we focus on ICA and CPD in much of the discussion, our conclusions are applicable to other decompositions such as non-negative matrix factorization (NMF), sparse decompositions, and

tensor decompositions like PARAFAC2 and Tucker [3], especially when their uniqueness guarantees can be established for a given problem.

The remainder of the paper is organized as follows. First, we introduce the three key concepts in our development: uniqueness, interpretability, and model match. To help make these ideas concrete, we introduce two applications as examples: fMRI data analysis using ICA and fluorescence spectroscopy data analysis using CPD. We then introduce the conditions for uniqueness for ICA and CPD, in Section 3. Section 4 provides the concrete definitions for reproducibility and replicability, Section 5 addresses replicability and Section 6 reproducibility. In Section 7, we provide numerical examples for the two motivating examples introduced in Section 2 to highlight our main messages. Finally, we conclude the paper with a summary and discussion.

2 Uniqueness, interpretability, and model match

To introduce the inter-related concepts of uniqueness, interpretability, and model match, we make use of two examples, analysis of fMRI data and fluorescence spectroscopy data. The first example makes use of ICA and the second one, of (non-negative) CPD, two models with uniqueness guarantees under general conditions.

An example in medical imaging: FMRI analysis

FMRI has enabled direct study of temporal and spatial changes in the brain either in response to various stimuli, or when the brain is at rest or in a neutral condition such as during sleep [22]. FMRI data is four dimensional, as volume data is captured as a function of time, and through its analysis with a focus on functional connectivity, we obtain estimates of intrinsic functional networks (FN), i.e., brain regions that might be physically disconnected but operate synchronously in that they have similar temporal fluctuations [23]. Their identification is a key step in the study of human brain function, with a goal of understanding the properties of the developing or aging brain, or mechanisms of addiction, mental illness and other disorders. Relatively low image contrast-to-noise ratio of the blood oxygenation level dependent fMRI signal, head movement, and undesired physiological sources of variability such as cardiac and pulmonary, make detection of the activation-related signal changes challenging. Hence, the robust identification of functional networks given an fMRI scan over a time period, P, is critical.

In Figure 1, we show two approaches for the analysis of fMRI data, i.e., estimation of FN: (i) a model-based approach based on linear regression, the general linear model (GLM), and (ii) a data-driven approach based on ICA applied in the spatial dimension [22,24]. As shown in the figure, both the GLM and the ICA approaches start with the same dataset, the fMRI data, $\boldsymbol{X}$, arranged as a matrix of time points P by volume elements, *voxels*, V, by flattening the volume as a row at each time instant, and make use of a simple linear mixing model

$$\boldsymbol{X} = \boldsymbol{A}\boldsymbol{S} \tag{1}$$

as shown in Figure 1.

The model-based GLM approach, which can be implemented through the Statistical Parametric Mapping (SPM) toolbox [25], makes use of a user-defined design matrix. The columns of this matrix are defined using the information in the time courses for a given task, and hence the approach is widely

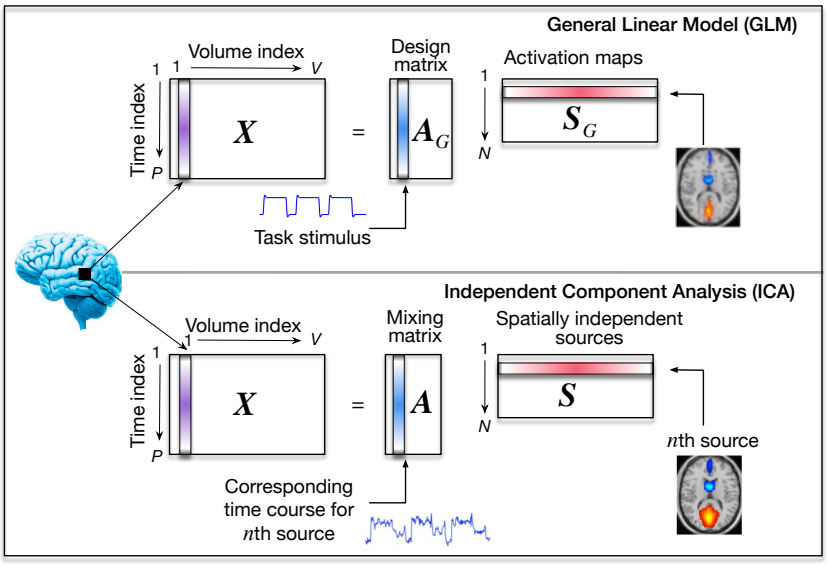


Figure 1: Given the fMRI data $\boldsymbol{X}$ arranged as a matrix, a model-based and a data-driven approach for estimating the functional networks. Here, N represents the number of sources, i.e., FN, in ICA such that $N < P$, and it is given by the number of regressors in GLM.

used for the analysis of task fMRI data. Given the user-defined mixing matrix, $\boldsymbol{A}_G$, the spatial maps are given by the least squares solution

$$\boldsymbol{S}_G = (\boldsymbol{A}_G^{\mathsf{T}}\boldsymbol{A}_G)^{-1}\boldsymbol{A}_G^{\mathsf{T}}\boldsymbol{X}. \tag{2}$$

The ICA approach, on the other hand, alleviates the need to specify timecourses and estimates both the mixing matrix $\boldsymbol{A}$ and the spatial maps given through matrix $\boldsymbol{S}$. Hence, it is attractive for the analysis of resting-state data where one cannot specify a design matrix $\boldsymbol{A}_G$, and even for task-related fMRI studies, leads to increased sensitivity in the results through more flexible modeling of the timecourses [22].

An example in spectral unmixing: Fluorescence data analysis

Fluorescence spectroscopy measures the intensity of the light emitted at different emission wavelengths when a fluorophore (i.e., a fluorescent chemical compound) is excited at various excitation wavelengths. Fluorescence spectroscopy measurements of a sample (e.g., a mixture) with multiple fluorophores are in the form of a two-way excitation emission matrix (EEM) containing the contributions (signals) from all fluorophores. In the presence of several samples, measurements can be arranged as a third-order tensor with modes: *samples*, *emission wavelengths*, and *excitation wavelengths*. Such measurements of various types of samples (e.g., human plasma samples, environmental samples or food samples) are often analyzed with the goal of revealing the chemical contents (e.g., sources) of the samples for monitoring or diagnosis purposes.

The CPD model has proved to be a successful data-driven approach for the analysis of fluorescence spectroscopy measurements [26, 27]. Given a third-order tensor $\mathcal{X} \in \mathbb{R}^{I\times J\times K}$ with modes: *samples*, *emission wavelengths*, *excitation wavelengths* containing EEM matrices from multiple samples, an R-component CPD model represents $\mathcal{X}$ as follows:

$$\mathcal{X} \approx \sum_{r=1}^{R} \boldsymbol{a}_r \circ \boldsymbol{b}_r \circ \boldsymbol{c}_r = [\![\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}]\!], \tag{3}$$

where $\circ$ denotes the vector outer product, $\boldsymbol{a}_r, \boldsymbol{b}_r, \boldsymbol{c}_r$ correspond to the rth column of factor matrices $\boldsymbol{A} \in \mathbb{R}^{I \times R}$, $\boldsymbol{B} \in \mathbb{R}^{J \times R}$, and $\boldsymbol{C} \in \mathbb{R}^{K \times R}$ in the *samples*, *emission*, and *excitation* modes, respectively.

Model match:
The success of these two approaches, use of ICA for analysis of fMRI data and CPD for fluorescence data, is due to the match of these two mixing models to the way the observations are generated. The fMRI data can be modeled through the linear mixing of FN with their temporal modulations [28] as in (1), and the fluorescence data through the multi-linear model in (3) where $\boldsymbol{b}_r$ and $\boldsymbol{c}_r$ reveal the emission and excitation spectra of the chemical compound captured by the rth component, and $\boldsymbol{a}_r$ shows the relative concentrations of that chemical compound in the samples [27]. In addition, since relative concentrations and spectra are expected to be non-negative, non-negative CPD becomes a particularly well suited approach for spectral unmixing.

Importance of order selection for model match:
An important parameter for model match in these two examples is the dimension of the signal subspace, i.e., number of FN, N, for the ICA model and the number of different emission/excitation spectra R in the mixture for the CPD. Here, N, corresponds to the size (and rank) of the mixing matrix in ICA and R to the rank of tensor in CPD.

For ICA, there are a number of data-driven solutions for estimating N given $\boldsymbol{X} \in \mathbb{R}^{P \times V}$ where $P \geq N$. A typical approach is use of information theoretic criteria such as minimum description length [29] (equivalently the Bayesian information criterion [30]) as in [31], or using a correction that takes the fact that samples are correlated into account when computing the likelihood function [32, 33]. In this case, principal component analysis (PCA) is used to project the data onto the dimension-reduced space. In [34], this problem is posed as one of generalizable patterns in neuroimaging using PCA.

For CPD, model selection refers to estimation of the rank R. In this case, the noise and interference are directly modeled by the residual error as the typical cost function is given as

$$\min_{\boldsymbol{A},\boldsymbol{B},\boldsymbol{C}} \|\boldsymbol{\mathcal{X}} - [\![\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}]\!]\|^2 \tag{4}$$

where $\|\cdot\|$ is typically the Frobenius norm. There is no polynomial-time algorithm to determine the rank of a given tensor—it has been shown that estimating the tensor rank is NP (nondeterministic polynomial time)-hard [35]. There are, however, various approaches such as checking the model fit with increasing number of components to see how much of the data the model explains and whether additional components contribute significantly. Another approach is the use of a core consistency diagnostic called CORCONDIA [36]. Bayesian approaches relying on automatic relevance determination [37] as well as missing data estimation strategies [38–40] have also provided promising performance. In general, how to determine the best model order in MTD is an important challenge, e.g., different approaches may produce different solutions, e.g., [34], so it requires special attention. Here, we have just scratched the surface on the topic in this brief overview.

Selecting the right model order plays a key role in model match. When there is a good model match, the stability of the solutions increases, as there is an agreement with the way observations are generated. We demonstrate this point with examples in the section on numerical examples.

Uniqueness and interpretability:
A key property in any analytical model is uniqueness. Given the factorization in (1), it is easy to see that for any $N \times N$ invertible matrix $\boldsymbol{Z}$, we can always write

$$\boldsymbol{X} = \boldsymbol{A}\boldsymbol{S} = (\boldsymbol{A}\boldsymbol{Z})(\boldsymbol{Z}^{-1}\boldsymbol{S}) = \tilde{\boldsymbol{A}}\tilde{\boldsymbol{S}}. \tag{5}$$

Since the pairs of matrices $\{\boldsymbol{A}, \boldsymbol{S}\}$ and $\{\tilde{\boldsymbol{A}}, \tilde{\boldsymbol{S}}\}$ cannot be distinguished, this simple two-factor matrix decomposition is highly non-unique. It is with additional constraints on either one, or both of these matrices, $\boldsymbol{A}$ and $\boldsymbol{S}$, such as orthogonality and sparsity, one can guarantee uniqueness, i.e., can uniquely identify each corresponding column/row of the factor matrices.[2]

Uniqueness is an important property supporting interpretability, i.e., attaching a physical meaning to the rows/ columns of the factor matrices in a given decomposition [41,42]. In the fMRI analysis example in Figure 1, we associate the rows of matrix $\boldsymbol{S}$ with FN and the corresponding columns in the mixing matrix $\boldsymbol{A}$ as their temporal modulation for a given subject's fMRI data. These, for example, can be used to identify biomarkers for a given disorder such as schizophrenia. Similarly, we associate the columns of factor matrices in (3) with spectra of the emission and excitation modes, and their corresponding proportions in the samples/mixtures for the CPD model. Whenever a reference signal is available, for example if we have the spectra of interest, interpretability can be quantified using a similarity measure like correlation with this reference.

Without guarantees on the uniqueness of the factor matrices in the decomposition, it would make little sense to try to discuss their physical significance. For example, the least squares solution in (2) is unique when $\boldsymbol{A}_G$ is full column rank. In the next section, we review the basic conditions for uniqueness of the two models in our example, the ICA and the CPD.

3 Uniqueness guarantees: ICA and (N)-CPD

To state the conditions for uniqueness of ICA, we introduce the linear mixture model for ICA using the random process notation as that is the natural notation to introduce ICA and its properties. It is given by

$$\boldsymbol{x}(v) = \boldsymbol{A}\boldsymbol{s}(v), \tag{6}$$

where $\boldsymbol{x}(v)$ and $\boldsymbol{s}(v)$ are the N-dimensional mixture and source random vector processes, and $\boldsymbol{A}$ is the deterministic but unknown $N \times N$ mixing matrix, assumed to be full-rank. The index v might refer to space or time, where in the latter case, v is replaced by t for time. We used v in line with our fMRI example for voxels, where by using a random process notation rather than random variable, we can take statistical properties such as sample dependence and nonstationarity into account in addition to non-Gaussianity.

The main assumption in ICA is that the underlying latent random processes, i.e., *sources*, $s_n(v)$, $n = 1, \ldots, N$ of the random process vector $\boldsymbol{s}(v) \in \mathbb{R}^N$, are mutually *statistically independent*. This simple assumption enables essentially unique solutions, i.e., unique subject to the unavoidable permutation

[2] In general, we only require the decomposition to be *essentially unique* where *essential* uniqueness means that the individual factors in the decomposition are unique up to a common permutation and scaling/counter-scaling of the columns.

and scaling ambiguities. While independence might come across as a strong assumption, it is satisfied in many situations leading to a desirable model match.

ICA algorithms estimate a *demixing matrix* $\boldsymbol{W}$ using a cost function that maximizes independence, such as mutual information rate [7], and the source estimates $\boldsymbol{u}(v)$ are recovered via

$$\boldsymbol{u}(v) = \boldsymbol{W}\boldsymbol{x}(v). \tag{7}$$

If we consider a given set of V observations for each random process arranged as a matrix, $\boldsymbol{X} \in \mathbb{R}^{N\times V}$, we can write $\boldsymbol{X} = \boldsymbol{A}\boldsymbol{S}$ where $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ and $\boldsymbol{S} \in \mathbb{R}^{N\times V}$, and the source estimates for given $\boldsymbol{X}$ are recovered through

$$\boldsymbol{U} = \boldsymbol{W}\boldsymbol{X}. \tag{8}$$

In practice, most problems are overdetermined as in the case of fMRI analysis example, where there are more observations than the number of sources of interest. In this case, we typically perform dimension reduction, and project the original $P \times V$ data matrix to the lower dimensional matrix $N \times V$ matrix where $N \leq P$, and determines the dimensionality of the *signal subspace*. We refer to the problem of determining the order N as the model order selection problem, which plays a key role in *model match* as outlined above. Thus, in what follows, we assume the *determined* case where the number of mixtures $x_n(v)$ and sources $s_n(v)$ match, given by N, and hence $\boldsymbol{A}$ is square.

Hence, in ICA, under the assumption of independence, estimation of a single matrix, $\boldsymbol{W}$ is sufficient, as the second matrix in the decomposition, $\boldsymbol{U}$, estimate of $\boldsymbol{S}$, is recovered through a simple multiplication. This indicates a key difference for ICA, compared with other matrix decompositions such as dictionary learning and non-negative matrix factorization where both factors need to be estimated separately. Another important difference for ICA is that, uniqueness guarantees are established in a rather straightforward fashion by working within the maximum likelihood framework, and under very general conditions, while for other matrix decompositions can often only be established under specific conditions, e.g., [43].

While ICA has been around for some time, with first algorithms introduced in the late 1980s (see [5] for an engaging account of the early developments on ICA), a common misconception has been that it is primarily through the use of non-Gaussianity, i.e., higher-order statistical information, one can achieve ICA. This is the reason for the oft repeated misconception that ICA can identify only a single Gaussian source. ICA can be more generally achieved by using different statistical properties of the underlying sources in the mixture besides non-Gaussianity such as sample dependence and nonstationarity. We refer to these properties as *signal diversity*, and when multiple types of diversity are used in the development of the ICA algorithm, a broader class of sources—*including multiple Gaussians*—can be uniquely identified [5, 7]. Hence, the point of view that presents *higher-order statistics (HOS)* as the only means to achieve ICA presents a very limited view of ICA, and more importantly, of its capabilities.

Here, to simplify the development that is required for stating the results, we only consider sample dependence (non-whiteness) of the sources along with non-Gaussianity as types of signal diversity. We define the covariance function for the n-th (second-order stationary) source $s_n(v) \in \mathbb{R}$ as

$$c_{s_n}(\tau) = E\{s_n(v+\tau)s_n(v)\}.$$

Then, by writing the likelihood function, one can proceed by showing the necessary and sufficient conditions for the Fisher Information Matrix to stay positive definite, which results in the conditions given in Table 1 [5,7]. As observed, just by making use of non-whiteness, one can identify multiple Gaussians, and when jointly taken into account, sample dependence and HOS enable identification of a much broader class of signals. When other properties such as non-stationarity and non-circularity (when the data are complex valued), it can be shown that one can *uniquely* identify even a broader class of sources [44].

TYPE OF SIGNAL DIVERSITY	**IDENTIFICATION**
non-Gaussianity	Any source but only one Gaussian
non-whiteness	Any source except those with $c_{s_n}(\tau) = \alpha^2 c_{s_m}(\tau)$, for $n \neq m$ and $\alpha \neq 0$
non-Gaussianity + non-whiteness	Any source except Gaussians with $c_{s_n}(\tau) = \alpha^2 c_{s_m}(\tau)$, for $n \neq m$ and $\alpha \neq 0$

Table 1:
Using different types of diversity *jointly*, one can identify a broader class of signals including multiple Gaussians as long as they have different spectra across sources.

The most general uniqueness result for the CPD model is defined based on Kruskal rank (k-rank), which is defined such that the Kruskal rank of matrix $\boldsymbol{A}$, denoted $k_{\boldsymbol{A}}$, is the largest value of k such that any k columns of $\boldsymbol{A}$ are linearly independent [45]. Given this definition, for $\boldsymbol{\mathcal{X}}$ with factors $\boldsymbol{A}, \boldsymbol{B}$ and $\boldsymbol{C}$ given in (3), [45] states the sufficient condition for (essential) uniqueness of CPD as

$$k_{\boldsymbol{A}} + k_{\boldsymbol{B}} + k_{\boldsymbol{C}} \geq 2R + 2. \tag{9}$$

As in ICA, this is the condition for *essential uniqueness* as we can simply permute the order of rank-one terms, and also scale each column in a given rank-one term as in

$$\boldsymbol{\mathcal{X}} \approx \sum_{r=1}^{R} \boldsymbol{a}_r \circ \boldsymbol{b}_r \circ \boldsymbol{c}_r = \sum_{r=1}^{R} \alpha_r \boldsymbol{a}_r \circ \beta_r \boldsymbol{b}_r \circ \gamma_r \boldsymbol{c}_r$$

for $\alpha_r \beta_r \gamma_r = 1$. The sufficient condition in (9) has been generalized to N-way arrays and stronger conditions have been derived, see e.g., [3, 8]. In particular, for non-negative CPD, the approximation is shown to be well posed and the solution of N-CPD is almost always unique [21].

An important starting point for real data analysis is studying uniqueness properties of the selected decomposition. The conditions such as those given here can guide the study, help determine whether additional prior information such as smoothness or sparsity might need to be incorporated into the solution. E.g., in the case of ICA, these considerations can guide the algorithm choice, when an algorithm that accounts for multiple types of signal diversity is used a broader set of signals including multiple Gaussians can be identified. In addition, these all lead to better model match, and most often solutions that are more easily reproducible, and potentially replicable.

4 Definitions: (Computational) reproducibility and replicability

Following [18], we define *reproducibility* as obtaining consistent (same/similar) results using the same algorithm and data, and *replicability* as obtaining similar results—or more generally conclusions—across different datasets. The latter is a very general definition, and we would like to make the distinction about the use of a completely different dataset, e.g., use of neuroimaging data from a different site where the subjects and possibly their certain attributes are different as well, and the second scenario where the datasets are expected to come from the same distribution, which is the idea when implementing different sampling strategies for a given set of observations in replicability studies. The first scenario is different as in this case indeed one can reasonably expect only to have the main conclusions to be similar. E.g., in [46], a framework is presented that can effectively replicate estimation of brain network abnormalities of schizophrenia across different datasets. In what follows, we consider the second case when talking about replicability.

As one would expect, the terminology used for replicability and reproducibility widely varies in the literature. E.g., stability [47, 48], repeatability [49], similarity [17], and (algorithmic) reliability [50] have been all used to refer to reproducibility. At times, replicability and reproducibility have been used interchangeably [46], and reproducibility is used to refer to what we call replicability [12, 51]. Thus, while consulting other references, it is important to remember the definitions we make here, in this section.

Even though our focus is on reproducibility, because it is closely tied to replicability, which is important when determining the relevant parameters for decomposition, such as those for model order and regularization, we discuss replicability as well, next.

5 Replicability in MTD: Two solutions based on half-splits of observations

In MTD—for the unsupervised case, which we focus on—the use of different sampling strategies for cross validation is not well studied, especially when we compare it with their wide use in supervised machine learning. E.g., in sparse decompositions where there are a number of hyper-parameters to be determined, such as the sparsity level and the dictionary dimension, it has been rare seeing a clear account of the procedure for the selection of these hyper-parameters as in [43, 52] using training/validation/testing partitions. In certain studies, e.g., those in the medical field, it might be hard to achieve such partitions while keeping the statistical power of the study at an acceptable level if the number of subjects is relatively small. In this case, out-of-sample cross-validation has been used for evaluating replicability [53]. An attractive approach for replicability studies in MTD is using half-splits of the total observations, which we discuss next.

The NPAIRS (nonparametric prediction, activation, influence, and reproducibility resampling) framework [12] is originally introduced for neuroimaging data analysis, and for SPM which we introduced as the commonly used model-driven approach for fMRI analysis example. Since then, it has been used in a number of different contexts, especially for model selection in brain imaging studies, e.g., for determining the sparsity level [13], both sparsity and smoothness levels [54], the model order [55], or simply the model to be used [51]. The main argument in NPAIRS is selecting the solution that establishes a balance between accuracy (emphasizing model match, low deviation from the truth) and consistency of the

solutions (low variability), as this is the solution that corresponds to one that establishes the bias and variance trade-off.

The classical NPAIRS terminology refers to those two considerations as prediction accuracy vs reproducibility. We will instead refer to those in this discussion as prediction accuracy vs consistency, as we have a specific definition for reproducibility. The observations of a given dataset are split into two independent halves (e.g., across subjects): training and test sets. For example the model can be trained on the first split and the prediction accuracy estimated from the second split and vice versa, yielding two estimates of the prediction accuracy [13]. The process is repeated across such multiple halves and the averages are recorded. The solution that corresponds to high accuracy and highly consistent results is used in further analysis, e.g., if it is used to determine the model order, then the whole data is used with the order that corresponds to this solution.

As noted, NPAIRS makes use of split-half resampling to produce equal-sized training and test sets, hence ensuring that bias due to size differences in the two splits is eliminated and their independent error estimates can be directly compared. This has also been a common practice in the data-driven analysis of neuroimaging data when looking at differences across populations, such as a patient and a healthy control group, by making sure that the numbers of subjects in each group are balanced. This way, the power of each of the independent-group data analyses is maximized.

For tensors, Harshman and De Sarbo introduced the use of split-half procedure [56] for application of CPD to a practical problem with a small sample size, and noted the following:

> The strongest evidence for the "reality" of a factor—namely, that it is due to systematic influences and not just random noise—is the demonstration that the same or similar versions of the factor can be found in several independent samples of the data.

This is a very clear explanation of the model match problem, and given that it was written in 1984 and addressed small sample sizes, one could understand the use of *only several* independent samples. Today, our recommendation will be to make use of as many random samplings as the dataset allows. Since then, split-half experiments have been used in CPD for more general set of problems, for determining the correct number of components (rank, R) as well as for assessing the appropriateness of the CPD model for the given set of observations [26, 27, 56, 57]. Different strategies are proposed for generating the data halves [26, 56], and the argument as noted above is that if the model is representative of the sample population, the estimated components using independent data subsets should be highly similar. The arguments for using splits as noted in [56] is similar to that in NPAIRS, achieving better sensitivity of analysis.

6 Reproducibility: Selection of *"best run"* for interpretation

While the importance of making sure that all modeling and implementation details are accurate is being increasingly emphasized in literature [11, 18], *computational* reproducibility, which refers to obtaining consistent solutions, has not received the same attention, especially for MTD. However, it should be the first step also in replicability studies, e.g. when determining the hyper-parameters of an algorithm. For each run with different samplings of the data, selecting the most reproducible run will improve the

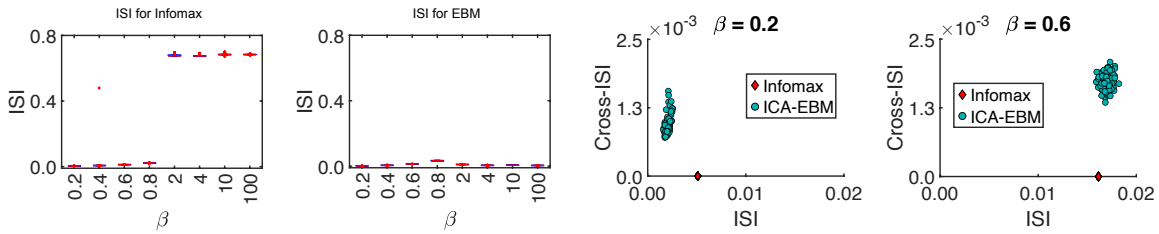
overall performance of the evaluations. In our discussion, we focus on the case where the dataset and algorithm fixed—which implies that all hyper-parameters for the algorithm are determined. Then, the remaining source of variability is due to random initialization of the algorithm.

For MTD, reproducibility is explicitly addressed only in a small subset of work that have interpretability as the main goal. Whenever it is addressed, this is achieved using multiple runs and defining a selection mechanism for identifying a single run as the desirable one, the one which will be interpreted, and/or studied further. We refer to this run as the *best run*. For tensor decompositions, and matrix decompositions that have a clear distance-based metric as in (4), the common approach has been to use multiple runs using the complete data and select the one with the minimum cost value for interpretation [19,20,58,59]. The second approach defines a metric such as the factor match score we define in (12) to measure the similarity of factors from different runs, or uses the similarity of estimated subspaces in order to evaluate the consistency of solutions. See e.g., [60] and the references therein, where a number of such measures are used for evaluating replicability (with our definition) for a dictionary learning solution, and [61] for selecting the order in an NMF solution. The proposed metrics in both references could also be used for reproducibility. This second approach has been commonly used for ICA for selecting the best run but not for tensor methods. Nevertheless, consistency of solutions has been used to select the number of components for CPD models [17]. As we demonstrate later with examples for fluorescence spectroscopy demixing introduced in Section 7.2, in fact both of these considerations should guide the selection of *a best run*. A highly consistent result might be due to high bias in the estimate, or a poor model match, and hence the cost function value should also guide the selection of the run for further evaluation.

Next, we review a number of solutions that have been proposed for evaluating the most consistent run, and though these solutions have been proposed in the context of ICA, these are adopted for other MTD as well, in particular the first solution ICASSO [50] has been used for a number of other decompositions [15, 48].

ICASSO: As we show with an example in this section in Figure 2, algorithmic variability in ICA might be due to modeling flexibility that in turn provides a better model match. Variability in the solutions can also result from the nature of updates. An early ICA algorithm, FastICA [62] is efficient but uses a fixed-point type iteration scheme with an orthogonality constraint, and hence is prone to instability [63]. It has motivated the introduction of a systematic approach for selecting a highly repeatable set of solutions, using a method named ICASSO [50], which also enables visualization of the solution space. The approach is based on performing multiple (M) runs of the given algorithm with different initializations[3]. All $N \times M$ source estimates where N is the number of sources are clustered based on their spatial correlation and are evaluated using a quality index based on the compactness and isolation of the clusters. The centroids of the clusters are selected as the final solution. One issue with the approach is that the selected set of sources does not necessarily allow one to reliably reconstruct the original dataset $\boldsymbol{X}$ using a single demixing matrix estimate, as the estimates typically do not come from the same run. As this is undesirable for applications where the goal is interpretation of the factors, in [64] a modified evaluation metric for ICASSO is introduced such that a single run can be selected, and used for ICA of fMRI. In [15], another modification of ICASSO is proposed where the number of clusters is determined using a low-

[3] In the introduction of ICASSO, random samplings of the data is included as another possibility.



(a) ISI values (b) Cross-ISI vs ISI values

Figure 2: ISI and Cross-ISI for two ICA algorithms: Infomax that uses a fixed nonlinearity providing a good match only for super-Gaussian sources ($\beta < 1$) and EBM that uses an adaptive procedure for nonlinearity selection based on the maximum entropy principle. In addition, note that use of only a reproducibility metric such as Cross-ISI might not be always sufficient for selecting the best run.

rank graph approximation whereas in ICASSO, this is a user-defined parameter. The method is noted to yield more coherent clusters than ICASSO and is demonstrated for electroencephalogram (EEG) analysis using CPD.

MST: An approach introduced specifically for fMRI analysis, minimum spanning tree (MST) [65] aligns the components across multiple ICA runs using the linear assignment problem. The minimum cost of alignment and the corresponding alignment for each pair is computed, and is followed by identification of a central run as the run with minimum cost of alignment. The components in each run are reordered as per this central run. After alignment, a one-sample t-test is performed across runs in order to evaluate the reproducibility of the estimated components. The best run is selected as the run with highest correlation between the components and the corresponding T-maps.

Cross-ISI: Inter-symbol interference (ISI) is a frequently used global metric for performance evaluation in ICA when the ground truth is available. It is defined as

$$\mathrm{ISI}(\mathbf{G}) = \frac{1}{2N(N-1)} \cdot \sum_{n=1}^{N} \left(\sum_{m=1}^{N} \frac{\|g_{nm}\|}{\max_k \|g_{nk}\|} - 1 \right) + \frac{1}{2N(N-1)} \cdot \sum_{m=1}^{N} \left(\sum_{n=1}^{N} \frac{\|g_{nm}\|}{\max_k \|g_{km}\|} - 1 \right) \tag{10}$$

where $\mathbf{G} = \mathbf{AW}$ with elements denoted as g_{nm}, where $\mathbf{A}$ is the true mixing matrix and $\mathbf{W}$ is the estimated demixing matrix. If $\mathbf{W}$ is perfectly estimated, $\mathbf{G}$ is identity subject to permutation and scaling ambiguities, thus yielding zero ISI and indicating perfect separation. Therefore, the smaller the ISI, the closer the estimates are to the ground truth, and the metric is bounded such that $0 \leq \mathrm{ISI} \leq 1$. In [16], motivated by the definition of ISI, a new global metric, Cross-ISI, is introduced to assess the consistency of the components across runs. It is defined as $\mathrm{ISI}^{\mathrm{C}}_{ij} = \mathrm{ISI}(\boldsymbol{P}^{ij})$, where $\boldsymbol{P}^{ij} = \boldsymbol{A}_i \boldsymbol{W}_j$ with elements denoted as p^{ij}_{nm}, where $\boldsymbol{A}_i = \boldsymbol{W}_i^{-1}$ is the inverse of the demixing matrix of the ith run and $\boldsymbol{W}_j$ is the demixing matrix of the jth run. To measure the consistency of a single run to all the other runs, the cross-ISI of the current run is generated by averaging all its pairwise cross-ISI values

$$\mathrm{ISI}^{\mathrm{C}}_i = \frac{1}{K-1} \sum_{j=1, j\neq i}^{K} \mathrm{ISI}^{\mathrm{C}}_{ij} \tag{11}$$

where K is the total number of runs. Cross-ISI is a simple function of only the estimated demixing matrices, and is computationally efficient. Hence it can be computed when there is no ground truth that is available, while ground truth is needed in the evaluation of ISI.

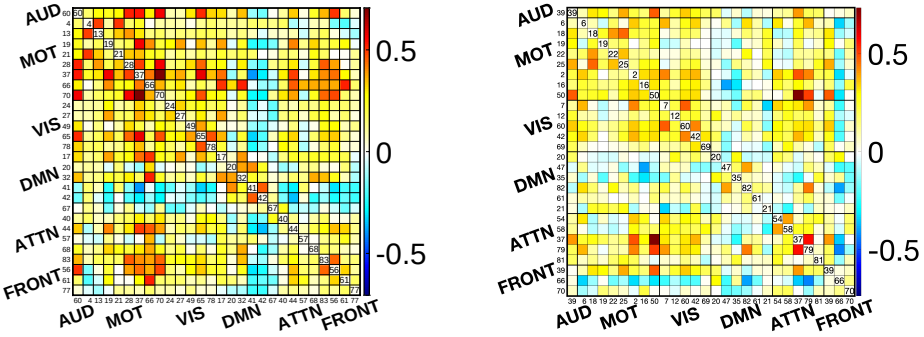
An example for Cross-ISI vs ISI: We end this section with an example to help explain the role of Cross-ISI as a reproducibility metric and its relationship with ISI, an accuracy metric. The example also helps us make a number of points, in particular (i) density estimation is another important aspect of model match in ICA; (ii) variations in the solution space might be desirable if it leads to better model match; and finally, (iii) for reproducibility as in the case of NPAIRS for replicability, we need to balance reproducibility considerations with a measure for accuracy.

We generated five sources, each with 10,000 samples from the generalized Gaussian distribution (GGD) varying the shape parameter, β, which controls the shape of the unimodal and symmetric GGD such that it is super-Gaussian for $0 < \beta < 1$, sub-Gaussian for $\beta > 1$, and Gaussian when $\beta = 1$. The mixing matrix is selected such that its condition number is less than 30, and then kept fixed. The only source of variability for each of the 100 runs is the random initializations, and the algorithm stopped because of the condition on the gradient.

In ICA, besides the estimation of the demixing matrix $\boldsymbol{W}$, matching of the underlying density of the sources is what helps achieve the desirable large sample properties of the likelihood function [7]. To demonstrate the importance of achieving model match through density estimation in ICA, we show the performance of a widely used ICA algorithm for fMRI analysis, Infomax [66] that uses a fixed nonlinearity matched to super-Gaussians and that of entropy bound minimization (EBM) [63], which uses an adaptive nonlinearity to match a relatively wide range of source distributions. As observed in Figures 2(a), since Infomax nonlinearity is a good match for super-Gaussian sources, the ISI values for Infomax are low when $\beta < 1$, and the sources are not successfully separated when $\beta > 1$. On the other hand, EBM yields low ISI values for all values of β shown here. For $\beta = 0.8$, both algorithms have slightly higher ISI as these are independent and identically distributed sources, and when the shape parameter is close to that of Gaussian ($\beta = 1$), performance slightly degrades. With an algorithm that takes sample dependence into account, we could identify multiple correlated Gaussians and not observe this penalty.

Figures (2(b)) show the cross-ISI values plotted against the ISI values for two shape parameters for which both Infomax and EBM provide good performance. Since, we would like to typically have both good separation performance and better reproducibility, we should choose a solution with low ISI and Cross-ISI, and hence looking at Cross-ISI alone might not be sufficient. This is especially important in challenging optimization landscapes, and in this case, if we were to plot Cross-ISI vs ISI for values of $\beta > 1$ where Infomax has a very poor model match and fails, this situation would be more pronounced. Obviously, ISI computation requires knowledge of ground truth but it can be easily replaced with a metric that corresponds to the cost function value, the goal in the decomposition, e.g., a measure of independence for ICA.

Another point important to note is that the EBM solutions are more variable compared with those for Infomax. Since EBM is adaptive, it is able to decrease the bias due to density mismatch leading to consistently low ISI values but at the expense of increased variability in the solutions. Such an adaptive solution to pdf estimation provides a better model match, and is shown to provide better estimation of



(a) Run with the lowest Cross-ISI (best run) (b) Run with the highest Cross-ISI

Figure 3: Functional network connectivity (FNC) maps for the best run and the one with the highest Cross-ISI. Note the best run result has the expected anti-correlations with DMN to other networks and stronger modularity overall compared with the highest cross-ISI result, hence leading to better interpretability. (AUD: Auditory; MOT: Sensorimotor; VIS: Visual; ATTN: Attentional, and FRONT: Frontal networks; DMN Default mode network.)

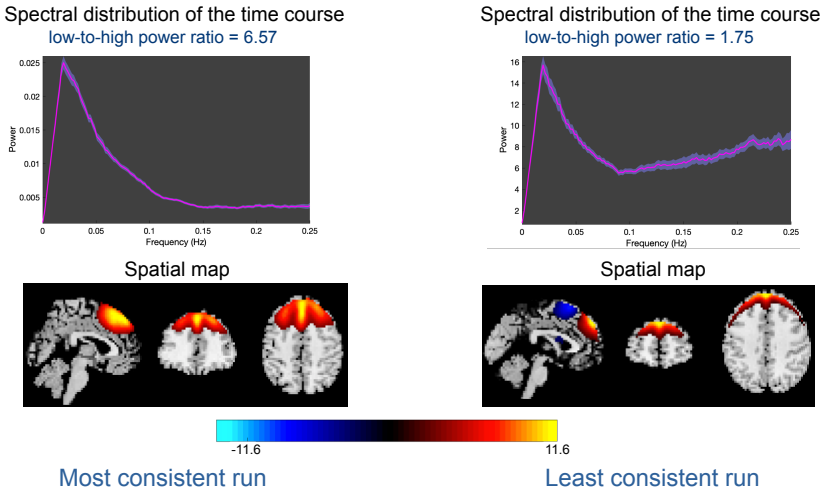
the underlying sources in a mixture with multiple examples including those in fMRI analysis [67].

7 Numerical examples

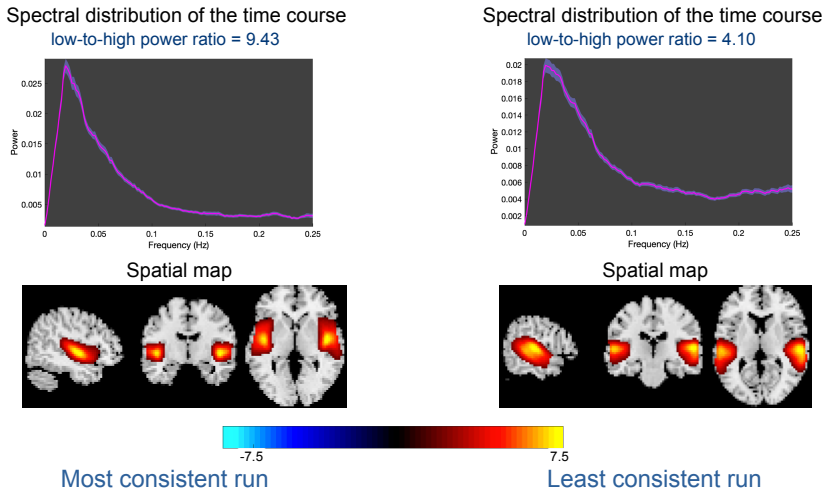
To demonstrate the importance of performing a *best run selection mechanism* in an application setting, we make use of the two motivating examples introduced in Section 2 using numerical examples. These examples also help us make the point that when there is a good model match, then the results are more interpretable and reproducibility is easier to achieve. The first example is based on ICA of fMRI data shown in Figure 1, and is a case where there is no ground truth. The second example is based on spectral unmixing using CPD, and is constructed using ground truth spectra so that besides importance of best run selection, we can also study the implications of poor model match due to incorrect order (rank) estimation.

7.1 ICA: Importance of best run selection for model match

We use resting-state fMRI data obtained from the Center of Biomedical research Excellence (COBRE), which is available on the collaborative informatics and neuroimaging suite data exchange repository (https://coins.trendscenter.org/) [68]. The dataset we used consists of 176 subjects including 88 healthy controls (HCs) and 88 patients with schizophrenia (SZs). Each resting state scan contains 150 timepoints and the first 6 timepoints were omitted to address the T1 saturation effect. Subjects were instructed to keep their eyes open during the whole scan. Motion correction, slice time correction and spatial normalization was applied to each subject as pre-processing steps. Furthermore, masking was performed to discard non-brain voxels and masked voxels were flattened to form the observation vector of V = 58,604 voxels for each subject. Group ICA, a widely used method for multi-subject fMRI data analysis using ICA, was implemented using the Group ICA of fMRI Toolbox (GIFT) [69] version GIFTv3.0c available at `http://trendscenter.org/software/gift`. Two stages of PCA are used where in the first stage 100 components are retained in the subject-level data reduction step and 85 components in the group level



(a) Sensorimotor component



(b) Auditory component

Figure 4: Comparison for the most and least consistent runs for two components estimated using EBM. Note that in the most consistent run, the sensorimotor component estimation is localized to gray matter whereas the least consistent has higher frequencies and edges, consistent with an artifactual component. For auditory component, similarly, the most consistent run has less high frequency power.

data reduction step following the analysis by other work that made use of this dataset [70], hence, estimating a total of 85 intrinsic FNs. The EBM algorithm which enables flexible density matching is used and run 100 times with random initializations. The best run—referred to in this section as the most consistent run—is selected using the run with the minimum cross-ISI value, as well as the least consistent run, the run with the maximum cross-ISI value. After the ICA step, a back-reconstruction step is applied to obtain subject-specific source estimates, i.e., spatial maps.

A total of 28 components were selected as functionally relevant and interesting [71] by inspecting the spatial maps of each component, the power spectra of the time courses, and dynamic range that is the difference between peak power and minimum power at frequencies that comes after the peak. Selected components had peak activations in gray matter and low frequency variations in time courses (TCs) indicating more blood oxygenation level dependent (BOLD) activity, as criteria for meaningful FN estimates. Peak activations of FNs of interest tend to have more low frequency power (<0.1Hz) and the

dynamic range of FNs (between the high and low frequency bands) is typically higher than 0.03. Since there is no ground truth, for the comparison of the decompositions obtained using the most and least consistent runs, we studied functional network connectivity (FNC), i.e., cross-correlation among ICA timecourses and the fractional amplitude of low-frequency fluctuations (fALFF) values.

An FNC map shows the pairwise level of co-activation between brain regions [72] and is obtained by calculating the Pearson correlations of the component time courses associated with the 28 selected FNs. Subject specific TCs were detrended, despiked and filtered before computing the pairwise correlations across the 28 FNs which were transformed to Z-scores. FNC of the most and least consistent runs are shown in Figure 3, which is organized such that we have FNs grouped with respect to functions domains. Within a given functional domain, we expect higher connectivity values, which is clearly observed for the most consistent run. In addition, the default mode network (DMN), as expected, shows higher negative correlation (blue values) with the sensorimotor network. These aspects are weaker in the FNC map for the least consistent run suggesting that indeed there is a good model match for ICA of fMRI data, and hence the most consistent run provides a better model match and a more interpretable result.

fALFF is the ratio of the power of the time course spectra in the low frequency band (<0.10 Hz) to high frequency band (>0.15 Hz). While low fALFF values are mostly associated with cardiac and respiratory noise, high fALFF values typically indicate true BOLD activation [73]. The average fALFF value for all 28 components for the most consistent run is 6.89 ± 3.44 while for the least consistent run, this value is 5.55 ± 2.93. Because these values are expected to be higher for BOLD-related activity, higher values here are again indicating that for the most consistent run, there is a better model match. We show the spatial maps of two estimated components for the most and least consistent runs as example, the auditory network and the sensorimotor network in Figure 4. For both components, the most consistent run results in a map with larger fALFF and dynamic range which increase the confidence in interpretability of the component. In Figure 4(b), besides the larger fALFF and dynamic range for the most consistent run, larger and more focal activation areas are clearly observed for the most consistent run further increasing our confidence in model match and interpretability of the component.

7.2 CPD: Importance of order and best run selection for model match

In this section, we analyze the well-studied amino acids dataset [27] with a known underlying ground truth using a non-negative CPD model, and demonstrate that (i) different initializations may lead to different solutions, (ii) the chance of ending up at a poor solution increases in the case of overfactoring, i.e., when the number of components is overestimated indicating poor model match, and (iii) while the solution corresponding to the minimum cost function value reveals the true factors in the noise-free case, it fails to capture the true factors in the presence of noise.

The amino acids dataset [4] consists of fluorescence spectroscopy measurements of five mixtures, each containing different amounts of three amino acids: tyrosine (Tyr), tryptophan (Trp), and phenylalanine (Phe). Measurements are arranged as a third-order tensor $\mathcal{X}$ in the form of a 5 mixtures × 201 emission wavelengths × 61 excitation wavelengths array. When this dataset is modeled using a 3-component CPD model using (3), each rank-one component models one of the amino acids (i.e., Tyr, Trp or Phe) with

[4] Available at `http://www.models.life.ku.dk/Amino_Acid_fluo`

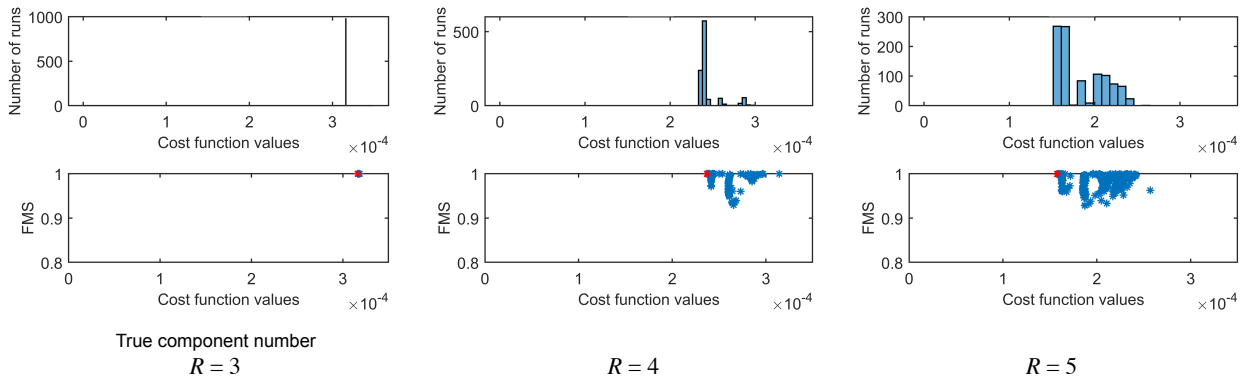


Figure 5: *Noise-free* case. Top plots: Histograms of cost function values reached using 1000 random initializations when a non-negative CPD model is fit to the original amino acids dataset using different number of components, R. Bottom plots: FMS values corresponding to different cost function values. The run that achieved the minimum cost function value is marked in red. In order to have the same scale for different number of components, few outliers are omitted.

$\boldsymbol{b}_r$ and $\boldsymbol{c}_r$ revealing the emission and excitation spectra of the amino acid modelled by component r while $\boldsymbol{a}_r$ showing the relative concentration of that amino acid in different mixtures. Non-negativity constraints are incorporated since factors in all modes are expected to be non-negative.

Here, we fit the non-negative CPD model using CP-OPT [74] from the Tensor Toolbox [75]. CP-OPT is a gradient-based all-at-once optimization approach solving the optimization problem (4) for all factor matrices simultaneously. In order to fit the model with non-negativity constraints, we use CP-OPT with the limited memory BFGS algorithm with bound constraints (LBFGS-B) [5]. We use 1000 random initializations, where entries in the factor matrices are drawn from uniform distribution. All runs stop due to either the gradient condition or the relative change in cost function value. We carry out the following experiments to study the "best run" selection problem when fitting the CPD model:

- *Noise-free case* : Original amino acids dataset ($\mathcal{X}$) is modelled using $R = \{3,4,5\}$ components. While we refer to this case as *noise-free*, it is a real experimental dataset and contains noise. However, the 3-component non-negative CPD model can model the original data well with a model fit of 99.9%, where the model fit is defined as $100 \times \left(1 - \frac{\|\mathcal{X}-\hat{\mathcal{X}}\|^2}{\|\mathcal{X}\|^2}\right)$ and $\hat{\mathcal{X}}$ denotes the approximation. Here, $R = 3$ corresponds to the true number of components, i.e., the number of amino acids, while $R > 3$ results in overfactoring.

- *Noisy case* : We add noise to the original dataset $\mathcal{X}$ and construct $\mathcal{X}_{\text{noisy}}$, where $\mathcal{X}_{\text{noisy}} = \mathcal{X} + \eta \frac{\mathcal{N}}{\|\mathcal{N}\|}\|\mathcal{X}\|$. Here, $\mathcal{N}$ is a tensor of the same size as $\mathcal{X}$ with entries drawn from the standard normal distribution, and η indicates the noise level. We use $\eta = 0.5$ in the experiments. The noisy tensor is then modelled using an R-component non-negative CPD model with $R = \{3,4,5\}$.

As the ground truth corresponds to the factors extracted by a 3-component non-negative CPD model, $\boldsymbol{A},\boldsymbol{B},\boldsymbol{C}$, we assess the quality of different solutions, $\hat{\boldsymbol{A}},\hat{\boldsymbol{B}},\hat{\boldsymbol{C}}$, using the factor match score (FMS) defined

[5] Available at `https://github.com/stephenbeckr/L-BFGS-B-C`

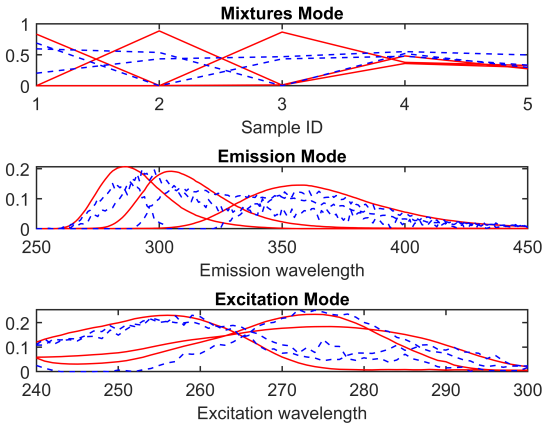


Figure 6: Components from a 3-component non-negative CPD model in the *noise-free* case. Ground truth is the solution corresponding to the minimum cost function value (shown in red). Blue plots show the components corresponding to the maximum cost function value (out of 1000 random initializations).

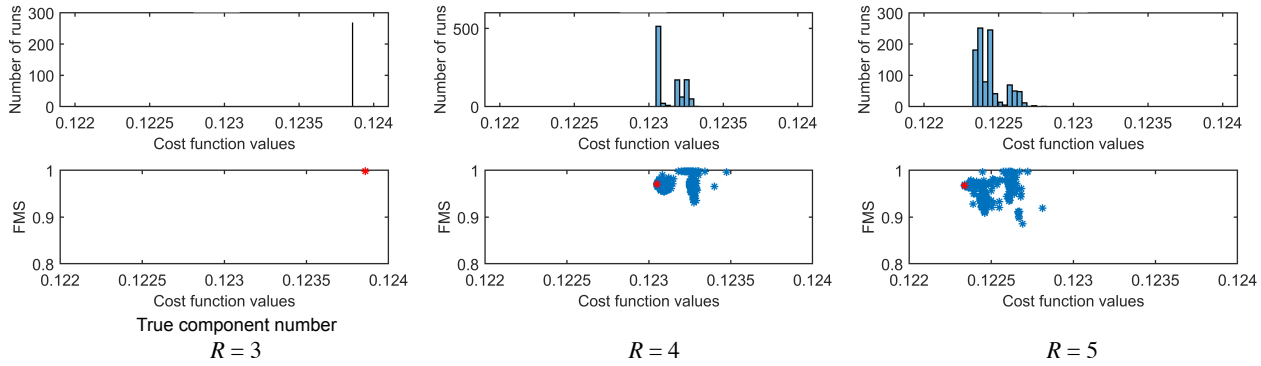


Figure 7: *Noisy* case. Top plots: Histograms of cost function values reached using 1000 random initializations when a non-negative CPD model is fit to the noisy amino acids dataset using different number of components, R. Bottom plots: FMS values corresponding to different cost function values. The run that achieved the minimum cost function value is marked in red. In order to have the same scale for different number of components, few outliers are omitted.

as follows (after finding the best matching permutation):

$$\text{FMS} = \frac{1}{R}\sum_{r=1}^{R} \frac{\boldsymbol{a}_r^\top \hat{\boldsymbol{a}}_r}{\|\boldsymbol{a}_r\|\|\hat{\boldsymbol{a}}_r\|} \frac{\boldsymbol{b}_r^\top \hat{\boldsymbol{b}}_r}{\|\boldsymbol{b}_r\|\|\hat{\boldsymbol{b}}_r\|} \frac{\boldsymbol{c}_r^\top \hat{\boldsymbol{c}}_r}{\|\boldsymbol{c}_r\|\|\hat{\boldsymbol{c}}_r\|}. \tag{12}$$

In the *noise-free* case, when the data is modelled using a 3-component non-negative CPD model, except for a couple of initializations, almost all initializations reach the same cost function value, see Figure 5. If we choose one of those solutions reaching the minimum cost function value, the true underlying factors, i.e., true emission and excitation spectra of amino acids as well as the true relative concentrations, are captured as shown in Figure 6 (in red). However, if we were to choose one of the unlucky solutions, e.g., the maximum cost function value, components will not reveal the true patterns (see the patterns in blue in Figure 6). Similarly, in the case of overfactoring, when we use multiple initializations, and pick the solution corresponding to the minimum value, the three components revealing the underlying patterns

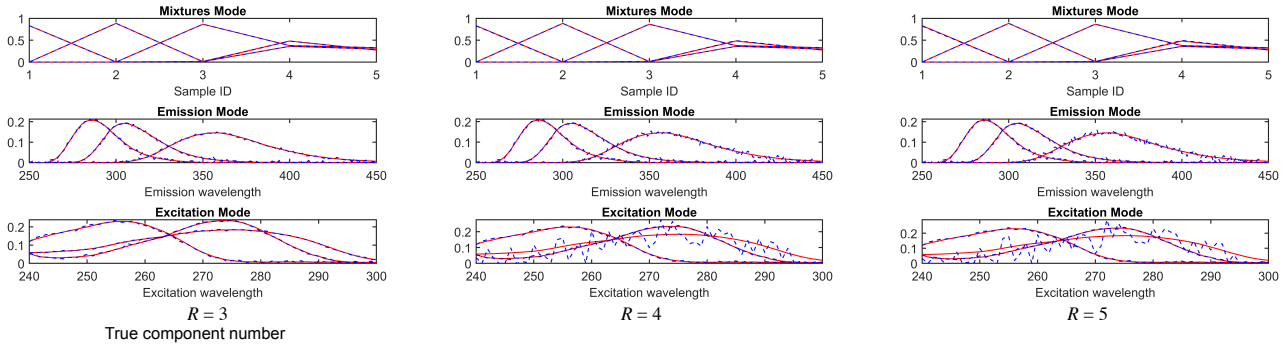


Figure 8: Components extracted using non-negative CPD models in the *noisy* case using different number of components, R. Ground truth is plotted in red while blue patterns show the components corresponding to the minimum cost function value (out of 1000 random initializations). In the case of $R = 4$ and $R = 5$, the three components that match best to the true components are plotted.

needed to identify the amino acids will be accurately captured. However, as we see in Figure 5 for $R = 4$ and $R = 5$, we may more easily end up at poor solutions.

When modelling $\mathcal{X}_{\text{noisy}}$ using a non-negative CPD model, for $R = 3$, again, almost always, we can reach the true solution (Figure 7 first column), and choosing the minimum cost function value reveals the true components as shown in Figure 8 first column (in blue) even though they are slightly noisier. As in the noise-free case, when the number of components is overestimated, it becomes more likely to end up at a different solution (Figure 7 for $R = 4$ and $R = 5$). Note that the solution corresponding to the minimum cost function value is no longer the best solution (i.e., the one with the highest FMS) when the number of components is overestimated, and it can no longer reveal the true patterns emphasizing the importance of model match. Much noisier patterns are extracted due to overfitting the noise as shown in Figure 8 for $R = 4$ and $R = 5$ using the solution corresponding to the minimum cost function value. In order to prevent overfitting, we could use ridge regularization on all factor matrices, and in that case, the solution corresponding to the minimum cost function value would reveal the true factors even in the case of overfitting providing a remedy for the best run selection problem.

8 Summary and discussion

In this paper, we focused on an important class of data-driven solutions, MTD for data analysis, where the goal is interpretability of the estimated factors. With this focus, we defined the inter-related concepts of uniqueness, interpretability, and model match. To solidify what we mean by these concepts, we made use of two examples where the MTD have found fruitful applications, analysis of fMRI data and unmixing of fluorescence data. We reviewed the uniqueness conditions for two decompositions used in these applications, ICA and CPD, and discussed the role of uniqueness. We reviewed the main solutions introduced to date for reproducibility, and the closely related concept of replicability for hyper-parameter selection by sampling of the available set of observations. We noted the limited nature of work in the area, and provided a critical view of the existing solutions, along with their promise and limitations. We then presented examples based on the two problems we introduced, analysis of fMRI and fluorescence data, which demonstrate the importance of

- using a mechanism for selection of a best run among a large number of independent runs for further analysis and interpretation;
- model match for reproducibility.

When there is a good model match, the results are more interpretable. This is demonstrated with the FNC maps and estimates for sample FNs for the fMRI analysis example, and with a better match to the truth for the spectral demixing example, which uses data with ground truth. This second example also lets us demonstrate the role of order selection (for CPD of rank R) for model match. When there is a good model match and the order is correct, the solution with minimum cost function value yields also the best solution (highest correlation with truth), but this is not the case when there is noise and model mismatch, i.e., order is chosen differently. A proposed solution for order selection for CPD deserves mention in this respect. In [17], evaluation of accuracy versus reproducibility is proposed to determine the model order. For a simulated example, it is shown that when the data is fixed and models are estimated using random initializations, the error decreases with increasing number of components (rank), i.e., bias decreases, while reproducibility, evaluated using a similarity metric for the estimated components across initializations decreases as well, i.e., variance increases. The true model order is the one that provides the trade-off between the two.

Methods such as regularization can help alleviate the negative effects of poor model match, at the expense of introducing new parameters to the algorithm. This in turn increases complexity, especially in terms of the need to tune additional user-defined parameters. In addition, sampling strategies for hyperparameter selection should be carefully employed in MTD, with split half methods being favored for multiple reasons including limited nature of observations. In addition, during the process, one should make sure to select the most reproducible run for each parameter setting and sampling choice.

More importantly, we noted that within the limited set of solutions that do account for reproducibility for MTD, many fall short of providing a complete picture. The two sets of solutions that are used to select a *a best run* among multiple runs either (i) using a metric to evaluate the consistency of the solution among multiple runs, or (ii) basing this decision on the minimum cost value. However, both of these considerations should guide the selection of the "best run" to be used for further interpretation as only together they provide a complete picture. When taking the cost function value one should be careful though as the use of regularization, and in the case of ICA, selection of different nonlinearities effectively changes the cost. For ICA, one can opt for use of a direct criterion related to the goal, such as use of mutual information (rate).

Next, we summarize our guideline for assuring reproducibility in MTD, which we believe is the first attempt of a checklist for this important class of data-driven solutions. The machine learning reproducibility checklist [11] has been compiled with supervised methods in mind, however the first set of recommendations on this list do hold for our case as well. These relate to the complete description of algorithms and models, theoretical claims, information on the datasets used, and the code that is shared. Our additional recommendations below relate to the experimental results that are reported, which are the last group of items of the list in [11].

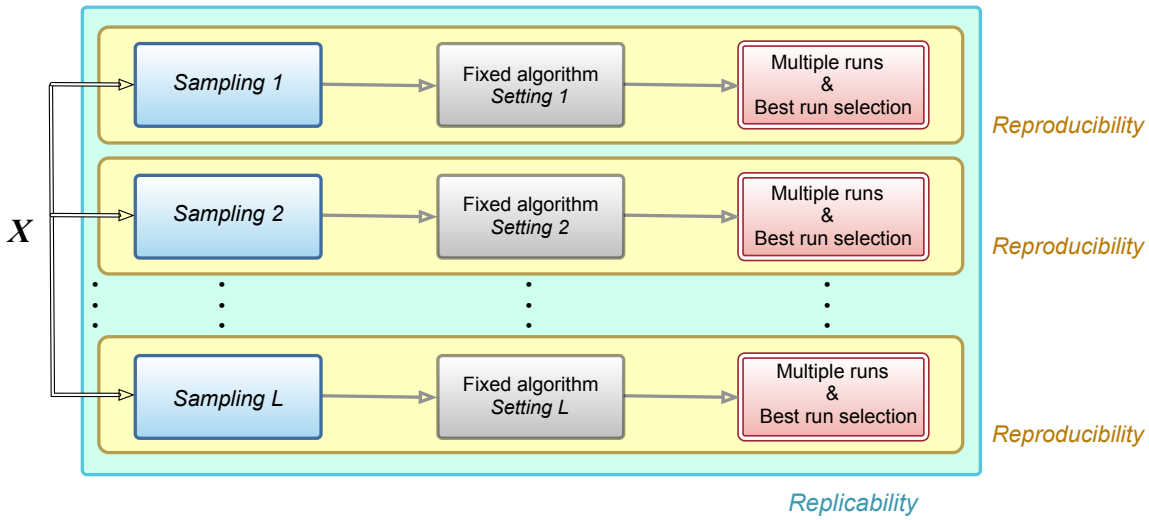


Figure 9: Summary of the steps for reproducibility and replicability for a given dataset $\boldsymbol{X}$ following the study of uniqueness conditions for the selected model. **Reproducibility** is represented by yellow blocks where all parameters, and hence method is fixed, but multiple runs are used.
Replicability is represented with the outer green block where different samplings from the dataset $\boldsymbol{X}$ are used.
Best run selection involves use of multiple runs with different initializations and choosing one using selected consistency metric(s) along with the cost function value.
• Fixed algorithm settings in the gray blocks can refer to the use of different orders, N, and/or regularization parameters such as those for sparsity or smoothness. Or, all parameters can be fixed to the same value (following an initial step to determine those) in settings 1 to L, and this can be used to study effects such as sample size.
• Samplings might make use of various strategies, such as NPAIRS/split halves, or might be testing sample sizes, in which case they can be based on selected increments in number of samples, e.g., subjects, again sampled from $\boldsymbol{X}$.

Given that our focus is interpretability using MTD, we note the following:

1. First, the uniqueness properties of the given decomposition should be studied for the given problem and setting.

2. The mechanism for hyper-parameter selection, including selection of the model order, should be clearly stated, and the most reproducible run—the best run—defined in step 3 below, should be used for each run that is used in the selection of hyper-parameters.

3. We described a number of approaches that can be used for best run selection, either using a measure on the consistency of the results, or using the minimum cost function value. Consistency metrics are more commonly used, and are important, however cost function value should also be taken into account as a solution with low variability but high bias as explained with the bias and variance dilemma will not be useful.

4. Finally, when selecting the best run, we should also make sure that a stationary point of the algorithm is reached. A large number of iterative MTD algorithms are (stochastic) gradient based as the evaluations of Hessian might be costly. In this case, a gradient-based stopping condition can be used as a criterion while making sure that the updates are stopped because that condition is satisfied, and not due to reaching the maximum number of iterations. Maximum number of iterations is typically used as a back-up given the fact that convergence might not be easy to achieve in certain scenarios.

In Figure 9, we summarize the main steps for the verification of reproducibility and replicability.

For replicability, we refer to our main focus in the paper where *different datasets* in the definition imply data from the same distribution, in this case, different samplings from the given dataset $\boldsymbol{X}$. When the reference is to a completely different dataset, e.g., coming from another study with the same overall goal such as identification of biomarkers for a certain disorder, then the measures of similarity of the solutions need to be modified. In this case, they are likely to be more general than the strict consistency measures we defined in this paper and are usually context dependent.

Computational reproducibility, and the effects of ending up in undesirable local optima are obviously important in supervised models like deep nets as well. In the case of MTD, since the solutions are directly interpretable, we can more easily demonstrate the problems in such cases. Since model match of which model order selection is an important component, this might also mean that identifying subspaces might help achieve a better match to the properties of observations, and helps enrich the simple linear/multi-linear mixing assumptions. More importantly, with good model match, the reproducibility considerations become easier. Obviously, a related important factor is the optimization landscape, which depends on multiple factors including order/rank, selected model, and nature and size of data. This is an area where further research, especially with reproducibility in mind, is needed. A good example is [76] where the properties of local optima are discussed in detail, also supporting the importance of using random initializations.

Given the growing emphasis on data-driven solutions across disciplines, interpretability and explainability of solutions have been receiving increased attention, and for MTD, this important aspect comes for free. Hence, we are hoping that our attempt at providing a reproducibility checklist for this important class of solutions in machine learning will help with their wider acceptability and applicability.

Acknowledgments: We would like to thank Simon Van Eyndhoven, Nico Vervliet, Dana Lahat, David Brie, Seung-Jun Kim, Florian Becker, and Lars Kai Hansen for their valuable feedback and pointers for references and work in the area.

References

[1] L. K. Hansen and L. Rieger, *Interpretability in Intelligent Systems – A New Concept?* Springer International Publishing, 2019, pp. 41–49.

[2] W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller, *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller, Eds. Springer International Publishing, 2019.

[3] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Review*, vol. 51, no. 3, pp. 455–500, Aug. 2009.

[4] E. Acar and B. Yener, "Unsupervised multiway data analysis: A literature survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 1, pp. 6–20, Jan. 2009.

[5] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. Academic Press, 2010.

[6] T. Adalı, C. Jutten, A. Yeredor, A. Cichocki, and E. Moreau, "Source separation and applications: Recent advances," *IEEE Signal Processing Magazine*, vol. 31, p. 3, 2014.

[7] T. Adalı, M. Anderson, and G.-S. Fu, "Diversity in independent component and vector analyses: Identifiability, algorithms, and applications in medical imaging," *IEEE Signal Proc. Mag.*, vol. 31, no. 3, pp. 18–33, May 2014.

[8] N. D. Sidiropoulos, L. D. Lathauwer, X. Fu, K. Huang, E. E. Papalexakis, and C. Faloutsos, "Tensor decomposition for signal processing and machine learning," *IEEE Transactions on Signal Processing*, vol. 65, no. 13, pp. 3551–3582, July 2017.

[9] F. Cong, Q.-H. Lin, L.-D. Kuang, X.-F. Gong, P. Astikainen, and T. Ristaniemi, "Tensor decomposition of EEG signals: A brief review," *Journal of Neuroscience Methods*, vol. 248, pp. 59–69, June 2015.

[10] T. Adalı, M. A. B. S. Akhonda, and V. D. Calhoun, "ICA and IVA for data fusion: An overview and a new approach based on disjoint subspaces," *IEEE Sensors Letters*, vol. 3, no. 1, pp. 1–4, Jan. 2019.

[11] J. Pineau, P. Vincent-Lamarre, K. Sinha, V. Lariviere, A. Beygelzimer, F. d'Alche Buc, E. Fox, and H. Larochelle, "Improving reproducibility in machine learning research(a report from the NeurIPS 2019 reproducibility program)," *Journal of Machine Learning Research*, vol. 22, no. 164, pp. 1–20, 2021. [Online]. Available: http://jmlr.org/papers/v22/20-303.html

[12] S. C. Strother, J. Anderson, L. K. Hansen, U. Kjems, R. Kustra, J. Sidtis, S. Frutiger, S. Muley, S. LaConte, and D. Rottenberg, "The quantitative evaluation of functional neuroimaging experiments: The NPAIRS data analysis framework," *NeuroImage*, vol. 15, no. 4, pp. 747–771, April 2002.

[13] P. M. Rasmussen, L. K. Hansen, K. H. Madsen, N. W. Churchill, and S. C. Strother, "Model sparsity and brain pattern interpretation of classification models in neuroimaging," *Pattern Recognition*, vol. 45, no. 6, pp. 2085–2100, 2012.

[14] S. C. Strother, P. M. Rasmussen, N. W. Churchill, and L. K. Hansen, *Stability and Reproducibility in fMRI Analysis*. The MIT Press, 2014.

[15] S. V. Eyndhoven, N. Vervliet, L. D. Lathauwer, and S. V. Huffel, "Identifying stable components of matrix /tensor factorizations via low-rank approximation of inter-factorization similarity," in *2019 27th European Signal Processing Conference (EUSIPCO)*. IEEE, Sep. 2019.

[16] Q. Long, C. Jia, Z. Boukouvalas, B. Gabrielson, D. Emge, and T. Adali, "Consistent run selection for independent component analysis: Application to fMRI analysis," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, April 2018.

[17] A. H. Williams, T. H. Kim, F. Wang, S. Vyas, S. I. Ryu, K. V. Shenoy, M. Schnitzer, T. G. Kolda, and S. Ganguli, "Unsupervised discovery of demixed, low-dimensional neural dynamics across multiple timescales through tensor component analysis," *Neuron*, vol. 98, no. 6, pp. 1099–1115.e8, June 2018.

[18] National Academies of Sciences, Engineering, and Medicine, *Reproducibility and Replicability in Science*. Washington, DC: The National Academies Press, 2019. [Online]. Available: https://www.nap.edu/catalog/25303/reproducibility-and-replicability-in-science

[19] E. Acar, D. M. Dunlavy, T. G. Kolda, and M. Mørup, "Scalable tensor factorizations for incomplete data," *Chemometrics and Intelligent Laboratory Systems*, vol. 106, no. 1, pp. 41–56, Mar. 2011.

[20] I. Perros, E. E. Papalexakis, R. Vuduc, E. Searles, and J. Sun, "Temporal phenotyping of medically complex children via PARAFAC2 tensor factorization," *Journal of Biomedical Informatics*, vol. 93, p. 103125, May 2019.

[21] Y. Qi, P. Comon, and L.-H. Lim, "Uniqueness of nonnegative tensor approximations," *IEEE Transactions on Information Theory*, vol. 62, no. 4, pp. 2170–2183, April 2016.

[22] V. D. Calhoun and T. Adali, "Multisubject independent component analysis of fMRI: A decade of intrinsic networks, default mode, and neurodiagnostic discovery," *IEEE Reviews in Biomedical Engineering*, vol. 5, pp. 60–73, 2012.

[23] M. P. van den Heuvel and H. E. H. Pol, "Exploring the brain network: A review on resting-state fMRI functional connectivity," *European Neuropsychopharmacology*, vol. 20, no. 8, pp. 519–534, Aug. 2010.

[24] M. J. McKeown, S. Makeig, G. G. Brown, T. P. Jung, S. S. Kindermann, A. J. Bell, and T. J. Sejnowski, "Analysis of fMRI data by blind separation into independent spatial components," *Human Brain Mapping*, vol. 6, no. 3, pp. 160–188, 1998.

[25] K. Friston, "SPM12," http://www.fil.ion.ucl.ac.uk/spm/software/spm12/, October 2014.

[26] K. R. Murphy, C. A. Stedmon, D. Graeber, and R. Bro, "Fluorescence spectroscopy and multi-way techniques. PARAFAC," *Analytical Methods*, vol. 5, no. 23, p. 6557, 2013.

[27] R. Bro, "PARAFAC. tutorial and applications," *Chemometrics and Intelligent Laboratory Systems*, vol. 38, no. 2, pp. 149–171, Oct. 1997.

[28] A. M. Dale and R. L. Buckner, "Selective averaging of rapidly presented individual trials using fMRI," *Human Brain Mapping*, vol. 5, no. 5, pp. 329–340, 1997.

[29] J. Rissanen, "Modeling by the shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.

[30] G. E. Schwarz, "Estimating the dimensions of a model," *Ann. Stats.*, vol. 6, no. 2, pp. 461–464, 1978.

[31] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 387–392, April 1985.

[32] Y.-O. Li, T. Adalı, and V. D. Calhoun, "Estimating the number of independent components for fMRI data," *Human Brain Mapping*, vol. 28, no. 11, pp. 1251–1266, Nov. 2007.

[33] G.-S. Fu, M. Anderson, and T. Adalı, "Likelihood estimators for dependent samples and their application to order detection," *Signal Processing, IEEE Transactions on*, vol. 62, no. 16, pp. 4237–4244, Aug 2014.

[34] L. K. Hansen, J. Larsen, F. Å. Nielsen, S. C. Strother, E. Rostrup, R. Savoy, N. Lange, J. Sidtis, C. Svarer, and O. B. Paulson, "Generalizable patterns in neuroimaging: How many principal components?" *NeuroImage*, vol. 9, no. 5, pp. 534–544, May 1999.

[35] J. Håstad, "Tensor rank is NP-complete," *Journal of Algorithms*, vol. 11, no. 4, pp. 644–654, Dec. 1990.

[36] R. Bro and H. A. L. Kiers, "A new efficient method for determining the number of components in PARAFAC models," *Journal of Chemometrics*, vol. 17, no. 5, pp. 274–286, 2003.

[37] M. Mørup and L. K. Hansen, "Automatic relevance determination for multi-way models," *Journal of Chemometrics*, vol. 23, pp. 352–363, 2009.

[38] M. Udell, C. Horn, R. Zadeh, and S. Boyd, *Generalized Low Rank Models*. Now Publishers, 2016, vol. 9, no. 1.

[39] R. Bro, K. Kjeldahl, A. K. Smilde, and H. A. L. Kiers, "Cross-validation of component models: A critical look at current methods," *Analytical and Bioanalytical Chemistry*, vol. 390, no. 5, pp. 1241–1251, Jan. 2008.

[40] A. B. Owen and P. O. Perry, "Bi-cross-validation of the SVD and the nonnegative matrix factorization," *The Annals of Applied Statistics*, vol. 3, no. 2, June 2009.

[41] R. B. Cattell, ""parallel proportional profiles" and other principles for determining the choice of factors by rotation," *Psychometrika*, vol. 9, no. 4, pp. 267–283, Dec 1944.

[42] R. A. Harshman, "Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-modal factor analysis," *UCLA Working Papers in Phonetics*, vol. 16, no. 1, p. 84, 1970.

[43] J. Mairal, F. Bach, and J. Ponce, "Task-driven dictionary learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 791–804, April 2012.

[44] G.-S. Fu, R. Phlypo, M. Anderson, and T. Adalı, "Complex independent component analysis using three types of diversity: Non-Gaussianity, nonwhiteness, and noncircularity," *IEEE Trans. Signal Processing*, vol. 63, no. 3, pp. 794–805, Feb. 2015.

[45] J. B. Kruskal, "Three-way arrays: rank and uniqueness of trilinear decompositions, with applications to arithmetic complexity and statistics," *Linear Algebra Appl.*, vol. 18, pp. 95–138, 1977.

[46] Y. Du, Z. Fu, J. Sui, S. Gao, Y. Xing, D. Lin, M. Salman, A. Abrol, M. A. Rahaman, J. Chen, L. E. Hong, P. Kochunov, E. A. Osuch, and V. D. Calhoun, "NeuroMark: An automated and adaptive ICA based pipeline to identify reproducible fMRI markers of brain disorders," *NeuroImage: Clinical*, vol. 28, p. 102375, 2020.

[47] A. Iraji, A. Faghiri, N. Lewis, Z. Fu, T. DeRamus, S. Qi, S. Rachakonda, Y. Du, and V. D. Calhoun, "Ultra-high-order ICA: an exploration of highly resolved data-driven representation of intrinsic connectivity networks (sparse ICNs)," in *Wavelets and Sparsity XVIII*, Y. M. Lu, M. Papadakis, and D. V. D. Ville, Eds. SPIE, Sep. 2019.

[48] M. Radek, M. Lamoš, R. Labounek, M. Bartoň, T. Slavíček, M. Mikl, I. Rektor, and M. Brázdil, "Multiway array decomposition of EEG spectrum: Implications of its stability for the exploration of large-scale brain networks," *Neural Computation*, vol. 29, no. 4, pp. 968–989, April 2017.

[49] A. Abou-Elseoud, T. Starck, J. Remes, J. Nikkinen, O. Tervonen, and V. Kiviniemi, "The effect of model order selection in group PICA," *Human Brain Mapping*, pp. 1207–1216, 2009.

[50] J. Himberg, A. Hyvärinen, and A. Esposito, "Validating the independent components of neuroimaging time-series via clustering and visualization," *NeuroImage*, vol. 22, pp. 1214–1222, 2004.

[51] M. Wernick, Y. Yang, J. Brankov, G. Yourganov, and S. Strother, "Machine learning in medical imaging," *IEEE Signal Processing Magazine*, vol. 27, no. 4, pp. 25–38, July 2010.

[52] R. Jin, K. Dontaraju, S.-J. Kim, S. Akhonda, and T. Adali, "Dictionary learning-based fMRI data analysis for capturing common and individual neural activation maps," *IEEE Journal of Selected Topics in Signal Processing*, pp. 1–1, 2020.

[53] E. Acar, C. Schenker, Y. Levin-Schwartz, V. D. Calhoun, and T. Adali, "Unraveling diagnostic biomarkers of schizophrenia through structure-revealing fusion of multi-modal neuroimaging data," *Frontiers in Neuroscience*, vol. 13, May 2019.

[54] Z. Boukouvalas, Y. Levin-Schwartz, V. D. Calhoun, and T. Adalı, "Sparsity and independence: Balancing two objectives in optimization for source separation with application to fMRI analysis," *Journal of the Franklin Institute*, vol. 355, no. 4, pp. 1873–1887, March 2018.

[55] G. Yourganov, X. Chen, A. S. Lukic, C. L. Grady, S. L. Small, M. N. Wernick, and S. C. Strother, "Dimensionality estimation for optimal detection of functional networks in BOLD fMRI data," *NeuroImage*, vol. 56, no. 2, pp. 531–543, May 2011.

[56] R. A. Harshman and W. S. D. Sarbo, *An application of PARAFAC to a small sample problem, demonstrating preprocessing, orthogonality constraints, and split-half diagnostic techniques*. Praeger, 1984, pp. 602–642.

[57] U. J. Wünsch, E. Acar, B. P. Koch, K. R. Murphy, P. Schmitt-Kopplin, and C. A. Stedmon, "The molecular fingerprint of fluorescent natural organic matter offers insight into biogeochemical sources and diagenetic state," *Analytical Chemistry*, vol. 90, no. 24, pp. 14 188–14 197, Nov. 2018.

[58] B. W. Bader, M. W. Berry, and M. Browne, "Discussion tracking in enron email using PARAFAC," in *Survey of Text Mining: Clustering, Classification, and Retrieval, Second Edition*, M. W. Berry and M. Castellanos, Eds. Springer, 2007, pp. 147–162.

[59] Z. Bai, P. Walker, A. Tschiffely, F. Wang, and I. Davidson, "Unsupervised network discovery for brain imaging data," in *Knowledge Discovery and Data Mining*, 2017, pp. 55–64.

[60] M. Morante, Y. Kopsinis, S. Theodoridis, and A. Protopapas, "Information assisted dictionary learning for fMRI data analysis," *IEEE Access*, vol. 8, pp. 90 052–90 068, 2020.

[61] S. Wu, A. Joseph, A. S. Hammonds, S. E. Celniker, B. Yu, and E. Frise, "Stability-driven nonnegative matrix factorization to interpret spatial gene expression and build local gene networks," *Proceedings of the National Academy of Sciences*, vol. 113, no. 16, pp. 4290–4295, Apr. 2016.

[62] A. Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Trans. Neural Networks*, vol. 10, no. 3, pp. 626–634, May 1999.

[63] X.-L. Li and T. Adalı, "Independent component analysis by entropy bound minimization," *IEEE Trans. Signal Processing*, vol. 58, no. 10, pp. 5151–5164, Oct. 2010.

[64] S. Ma, N. M. Correa, X.-L. Li, T. Eichele, V. D. Calhoun, and T. Adali, "Automatic identification of functional clusters in fMRI data using spatial dependence," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 12, pp. 3406–3417, Dec. 2011, the main reference for our ICASSO modification.

[65] W. Du, S. Ma, G.-S. Fu, V. D. Calhoun, and T. Adali, "A novel approach for assessing reliability of ICA for FMRI analysis," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, May 2014.

[66] A. Bell and T. Sejnowski, "An information maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, Nov. 1995.

[67] 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 Mapping*, vol. 40, no. 2, pp. 489–504, Feb 2019.

[68] C. J. Aine, H. J. Bockholt, J. R. Bustillo, J. M. Cañive, A. Caprihan, C. Gasparovic, F. M. Hanlon, J. M. Houck, R. E. Jung, J. Lauriello, J. Liu, A. R. Mayer, N. I. Perrone-Bizzozero, S. Posse, J. M. Stephen, J. A. Turner, V. P. Clark, and V. D. Calhoun, "Multimodal neuroimaging in schizophrenia: Description and dissemination," *Neuroinformatics*, vol. 15, no. 4, pp. 343–364, Aug. 2017.

[69] V. D. Calhoun, T. Adalı, J. J. Pekar, and G. D. Pearlson, "A method for making group inferences from functional MRI data using independent component analysis." *Human Brain Mapping*, vol. 14, no. 1, pp. 140–151, Nov. 2001.

[70] Q. Long, S. Bhinge, V. D. Calhoun, and T. Adali, "Independent vector analysis for common subspace analysis: Application to multi-subject fMRI data yields meaningful subgroups of schizophrenia," *NeuroImage*, vol. 216, p. 116872, Aug. 2020.

[71] E. Allen *et al.*, "A baseline for the multivariate comparison of resting state networks," *Frontiers in Systems Neuroscience*, vol. 5, p. 12, 2011.

[72] M. Jafri, G. D. Pearlson, M. Stevens, and V. D. Calhoun, "A method for functional network connectivity among spatially independent resting-state components in schizophrenia," *NeuroImage*, vol. 39, pp. 1666–1681, 2008.

[73] Q.-H. Zou, C.-Z. Zhu, Y. Yang, X.-N. Zuo, X.-Y. Long, Q.-J. Cao, Y.-F. Wang, and Y.-F. Zang, "An improved approach to detection of amplitude of low-frequency fluctuation (ALFF) for resting-state fMRI: Fractional ALFF," *Journal of Neuroscience Methods*, vol. 172, no. 1, pp. 137–141, July 2008.

[74] E. Acar, D. M. Dunlavy, and T. G. Kolda, "A scalable optimization approach for fitting canonical tensor decompositions," *Journal of Chemometrics*, vol. 25, pp. 67–86, Feb. 2011.

[75] B. W. Bader, T. G. Kolda *et al.*, "Matlab tensor toolbox version 3.1," Jun. 2019. [Online]. Available: https://www.tensortoolbox.org

[76] M. Wang and L. Li, "Learning from binary multiway data: Probabilistic tensor decomposition and its statistical optimality," *Journal of Machine Learning Research*, vol. 21, no. 154, pp. 1–38, 2020. [Online]. Available: http://jmlr.org/papers/v21/18-766.html

9 Authors

Tülay Adali received the Ph.D. degree in Electrical Engineering from North Carolina State University, and joined the faculty at UMBC, Baltimore, MD, the same year. She is currently a Distinguished University Professor in the Department of Computer Science and Electrical Engineering at UMBC. She has been active within the IEEE and is the past VP for Technical Directions for the IEEE SPS and is currently the Chair-Elect of IEEE Brain. Prof. Adali is a Fellow of the IEEE and the AIMBE, a Fulbright Scholar, and an IEEE SPS Distinguished Lecturer. She is the recipient of a Humboldt Research Award, an IEEE SPS Best Paper Award, the University System of Maryland Regents' Award for Research, and an NSF CAREER Award. Her current research interests are in the areas of statistical signal processing, machine learning, and their applications with emphasis on applications in medical image analysis and fusion.

Furkan Kantar received the B.S. degree in electrical engineering from Yildiz Technical University, Istanbul, Turkey (2015). He is currently pursuing his Ph.D. at Machine Learning for Signal Processing Lab in University of Maryland Baltimore County. His current research interests include neuroimaging analysis, matrix and tensor decomposition, blind source separation, statistical signal processing and machine learning.

Mohammad Abu Baker Siddique Akhonda received the B.S. degree in electronics and communication engineering from Khulna University of Engineering and Technology, Bangladesh, in 2013 and the M.S. degree in electrical engineering from the University of Maryland Baltimore County, USA, in 2019. From 2013 to 2016, he worked as a software engineer at Samsung Research and Development Institute. He is currently pursuing his Ph.D. at the Machine Learning for Signal Processing Lab in the University of Maryland Baltimore County. His research interests include joint blind source separation, multimodal and multiset data fusion, common and distinct subspace analysis, and joint model order selection problems.

Stephen Strother is a member, Rotman Research Institute, Baycrest, and Professor of Medical Biophysics at University of Toronto, Canada. He received his PhD in electrical engineering in 1986 from McGill University, Montreal,

Canada. Following a fellowship at MSK-Cancer Center, New York, USA, from 1989 he was a neuroimaging physicist, and Assistant Professor of Radiology, University of Minnesota. In 2004 he moved to Toronto. His research interests include neuro-informatics and data science for neuroimaging and big clinical data sets using statistical and machine learning techniques, applying these techniques in cognitive neuroscience and brain disease, and translating this work to non-academic settings.

Vince D. Calhoun received his BS in electrical engineering from the University of Kansas, MS in information systems and MA in biomedical engineering from Johns Hopkins, and a PhD in electrical engineering from the University of Maryland Baltimore County. He directs TReNDS with appointments at Georgia State, Georgia Tech and Emory. He is the author of 900+ peer-reviewed papers. He develops flexible methods to analyze neuroimaging data. He is a fellow of IEEE, AAAS, AIBME, ACNP, OHBM, and ISMSM. He serves on the IEEE BISP TC, the IEEE Data Science Initiative Steering Committee and the IEEE Brain TC.

Evrim Acar is a Chief Research Scientist at Simula Metropolitan Center for Digital Engineering (Oslo, Norway). She holds an M.S. and a Ph.D. in Computer Science from Rensselaer Polytechnic Institute (Troy, NY) and a B.S. in Computer Engineering from Bogazici University (Istanbul, Turkey). Prior to joining Simula, Evrim was a faculty member at the University of Copenhagen (Denmark), and a postdoctoral researcher at Sandia National Labs (Livermore, CA). Her research focuses on data mining methods, in particular, matrix/tensor factorizations, and their applications in diverse disciplines.