IISE Transactions on Healthcare Systems Engineering

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uhse21

Robust coupled tensor decomposition and feature extraction for multimodal medical data

Meng Zhao, Mostafa Reisi Gahrooei & Nathan Gaw

To cite this article: Meng Zhao, Mostafa Reisi Gahrooei & Nathan Gaw (2022): Robust coupled tensor decomposition and feature extraction for multimodal medical data, IISE Transactions on Healthcare Systems Engineering, DOI: 10.1080/24725579.2022.2141929

To link to this article: https://doi.org/10.1080/24725579.2022.2141929

IISE TRANSACTIONS ON HEALTHCARE SYSTEMS ENGINEERING
https://doi.org/10.1080/24725579.2022.2141929

Robust coupled tensor decomposition and feature extraction for multimodal medical data

Meng Zhao[a], Mostafa Reisi Gahrooei[a], and Nathan Gaw[b]

[a]Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL, USA; [b]Department of Operational Sciences, Air Force Institute of Technology, Wright-Patterson Air Force Base, OH, USA

ABSTRACT

High-dimensional and multimodal data to describe various aspects of a patient's clinical condition have become increasingly abundant in the medical field across a variety of domains. For example, in neuroimaging applications, electroencephalography (EEG) and functional magnetic resonance imaging (fMRI) can be collected simultaneously (i.e., EEG-fMRI) to provide high spatial and temporal resolution of a patient's brain function. Additionally, in telemonitoring applications, a smartphone can be used to record various aspects of a patient's condition using its built-in microphone, accelerometer, touch screen, etc. Coupled CANDECOMP/PARAFAC decomposition (CCPD) is a powerful approach to simultaneously extract common structures and features from multiple tensors and can be applied to these high-dimensional, multi-modal data. However, the existing CCPD models are inadequate to handle outliers, which are highly present in both applications. For EEG-fMRI, outliers are common due to fluctuations in the electromagnetic field resulting from interference between the EEG electrodes and the fMRI machine. For telemonitoring, outliers can result from patients not properly following instructions while performing smartphone-guided exercises at home. This motivates us to propose a robust CCPD (RCCPD) method for robust feature extraction. The proposed method utilizes the Alternating Direction Method of Multipliers (ADMM) to minimize an objective function that simultaneously decomposes a pair of coupled tensors and isolates outliers. We compare the proposed RCCPD method with the classical CP decomposition, the coupled matrix-tensor/tensor-tensor factorization (CMTF/CTTF), and the tensor robust CP decomposition (TRCPD). Experiments on both synthetic and real-world data demonstrate that the proposed RCCPD effectively handles outliers and outperforms the benchmarks in terms of accuracy.

KEYWORDS

Coupled tensor decomposition; robust decomposition; alternating direction method of multipliers (ADMM); Parkinson's disease telemonitoring; EEG-fMRI

1. Introduction

In many real-world applications, high-dimensional (HD) data such as multi-channel waveforms, images, and videos are available and can naturally be represented by tensors, i.e., multi-dimensional or higher-order arrays. Tensors provide a unified framework to represent various data forms to design more generic data analysis techniques. For example, a vector can be considered as a one-dimensional (or first-order) tensor and a matrix as a two-dimensional (or second-order) tensor. Higher-order tensors are the ones with more than two modes. Traditional data analysis methods that are vector-based are becoming insufficient due to their limitations in capturing the inherent correlations and interactions within the tensor data. To capture the correlation structure of HD data, multi-linear algebra has been studied and applied in diverse domains, including systems monitoring and control (Khanzadeh et al., 2018; H. Yan et al., 2019; Miao et al., 2022), prognostics (Fang et al., 2019), and healthcare (He et al., 2019).

At the core of multi-linear algebra lies tensor decomposition, which is an essential tool to exploit the correlation structure of higher-dimensional data and decompose a tensor into more basic and interpretable components (Kolda & Bader, 2009). For example, CP decomposition, among others, is one of the most widely used methods, which factorizes a tensor into a sum of rank-one tensors (Kolda & Bader, 2009). These decomposition techniques, however, are not designed to handle the presence of outliers within a tensor and therefore may result in extracting factors and features that are biased and misleading. To address this issue, several robust tensor decomposition approaches have been developed to remove the influence of outliers. These techniques decompose a tensor into a summation of three tensors: a smooth low-rank tensor, a sparse tensor, and an error tensor (Anandkumar et al., 2016; Gu et al., 2014; Xue et al., 2017), among which the sparse tensor contains outliers. To estimate these tensors, Anandkumar et al. (2016) propose a non-convex iterative algorithm that alternates between low-rank CP decomposition through gradient ascent and hard thresholding of residuals. Xue et al. (2017) develop two robust low-rank tensor recovery algorithms: tensor orthonormal robust PCA (TORPCA) and tensor robust CP decomposition (TRCPD), using Tucker and CP decomposition respectively with l_p norm regularization.

CONTACT Mostafa Reisi Gahrooei mreisigahrooei@ufl.edu Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL, USA

These decomposition approaches, however, cannot be applied to situations where multiple sources of data (modalities), with potentially different structures (e.g., images versus waveform signals), are available. For instance, electroencephalography (EEG) and functional Magnetic Resonance Imaging (fMRI) data are often collected simultaneously for a more accurate analysis of brain activities. These two modes of data cannot be represented by a single tensor due to their differences in resolution, dimension, and scale. EEG signals (usually represented by a channel $\times$ time matrix) measure electric oscillations on the scalp surface with excellent temporal resolution but lower spatial resolution, while fMRI (usually represented by a 4 D tensor with 3 D spatial dimensions $\times$ time) helps analyze brain activities with higher spatial but poor temporal resolution.

Due to the complementary properties of the individual modalities (e.g., each tensor), coupled data analysis has the potential to improve the understanding of the underlying structures, and to extract more informative and comprehensive features, used for accurate systems modeling and decision making (Gaw et al., 2022). Coupled decomposition approaches have been proposed for this purpose to extract common information from multiple sources of data. For example, Acar et al. (2011) and Acar et al. (2014) formulate a coupled matrix and tensor factorization (CMTF) problem where a tensor and a matrix that share a common mode are decomposed together. Chatzichristos et al. (2018) present a coupled tensor-tensor decomposition model which is applied to coupled third-order fMRI tensor and fourth-order EEG tensor analysis. Similarly, Jonmohamadi et al. (2020) develop a coupled tensor-tensor decomposition model for extraction of common features in EEG-fMRI tensors. Coupled tensor decomposition techniques have benefited many applications, including signal processing (Sørensen & De Lathauwer, 2013), chemistry (Acar et al., 2014), and bioinformatics (Mosayebi & Hossein-Zadeh, 2020; Naskovska et al., 2017). Several algorithms have been developed to solve the coupled decomposition problems. Acar et al. (2011); Acar et al. (2014) develop an all-at-once optimization method named CMTF-OPT to solve the CCPD problem. Farias et al. (2016) demonstrate the detailed analysis of the Alternating Least Squares (ALS) method to solve the coupled CP decomposition problem. Naskovska and Haardt (2016) propose a coupled Semi-Algebraic CP decomposition via the simultaneous matrix diagonalizations (C-SECSI) framework, which is an extension of the SECSI framework (Roemer & Haardt, 2013). Their proposed model can efficiently decompose two tensors coupled in one or more modes under different noise variances.

Unfortunately, however, the available coupled tensor decomposition methods are not robust to the presence of outliers in one or multiple modalities, as they generally assume global Gaussian noise, which is not sufficient for most multimodality data sets, in practice. As an example, simultaneous EEG-fMRI data has been widely used to combine the best of both techniques and discern various aspects of functional networks across the brain (Bridwell & Calhoun, 2019; Dizaji & Soltanian-Zadeh, 2017; Soon et al., 2021). However, it is known that the simultaneous EEG-fMRI data is often contaminated by outliers for example due to magnetic field gradients and subjects' motion (Bullock et al., 2021). Another example is telemonitoring of Parkinson's Disease in which patients' data are collected over time by a smartphone application and used to predict the severity of the patient's condition. Specifically, the smartphone application asks patients to perform predesigned activities such as speaking and tapping (see Figure 1). Often the collected data contains outliers due to a lack of professional monitoring while patients perform the activities at home. Even with clear instructions on how to perform the activities, some patients still exhibit issues with performing each task properly on their own (which creates outliers in the data). Therefore, an approach that can integrate different modalities (speaking and tapping) and handle the presence of outliers is necessary.

To address the foregoing limitations and challenges, this paper proposes a robust coupled CP decomposition (RCCPD) framework to jointly analyze heterogeneous HD data that contains outliers. In this framework, an objective function that isolates the outliers and simultaneously decomposes tensors to obtain factor matrices and joint features is formulated. The ADMM algorithm is then used to minimize an augmented Lagrangian of the objective function.

The rest of this article is organized as follows: In Section 2, we briefly introduce multi-linear algebra and tensor notations used in the paper, and then we summarize the related work. Section 3 discusses the proposed RCCPD framework and the algorithmic approach to model estimation. We describe the simulation studies in Section 4. Two case studies related to Parkinson's Disease (PD) telemonitoring and functional brain analysis are then provided in Section 5. In the first case study, we evaluate the efficacy of our approach in predicting a PD severity index based on features collected from exercises performed by PD patients on smartphones; in the second case study, we evaluate the performance of the proposed method in extracting informative features from simultaneous EEG-fMRI data to perform a trial classification task. Section 6 provides some practical notes about the proposed method. Section 7 provides final remarks and concludes the article.

2. Tensor algebra and related work in data fusion

In this section, we first introduce notations and basic multilinear algebra used throughout this article. More details on tensor algebra can be found in (Kolda & Bader, 2009). Next, we present several studies related to tensor data analysis for handling multimodal data. A multimodal dataset is organized in such a way that data can be grouped into multiple perspectives, for which each perspective consists of at least one feature (Gaw et al., 2022).

2.1. Notations and tensor algebra

We denote scalars, vectors, and matrices by lowercase or capital letters (a or A), boldface lowercase letters ($\mathbf{a}$), and

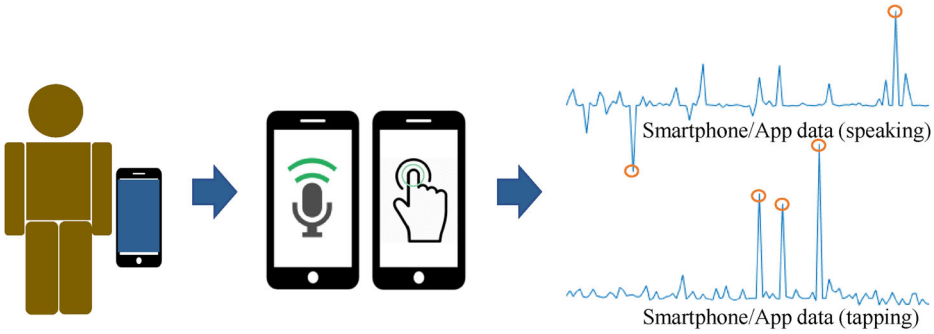


Figure 1. Telemonitoring workflow: first a patient performs exercises on a smartphone. The information from the exercises is then recorded by the smartphone, which can output signals from which features can be extracted and used for predictive modeling of disease conditions (shown to the right are example plots of one of the speaking and tapping features recorded over the course of a patient's monitoring period). Due to patients not being monitored at home, there is potential for outliers to occur (highlighted by the circles).

boldface capital letters ($\mathbf{A}$), respectively. Higher-order tensors are denoted by calligraphic letters. For example, an Nth-order tensor is denoted as $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$, where I_n $(n \in \{1, \ldots, N\})$ is the dimension of the nth mode of the tensor $\mathcal{X}$. Tensor fibers are defined by fixing all but one index of a tensor. The mode-n matricization of the tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ that reorders the elements of the N-way array into a matrix, is denoted by $\mathbf{X}_{(n)}$. This matricization is obtained by augmenting the nth mode fibers as the columns of the matrix. The element (i, j) of a matrix $\mathbf{A}$ is denoted by a_{ij} and the ith column vector of matrix $\mathbf{A}$ is denoted by $\mathbf{a}_i$. The symbols $\otimes$ and $\odot$ denote the Kronecker and Khatri-Rao products, respectively. $\mathbf{A} \otimes \mathbf{B}$, where $\mathbf{A} \in \mathbb{R}^{I \times J}$, $\mathbf{B} \in \mathbb{R}^{K \times L}$ is defined as:

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \cdots & a_{1J}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & & a_{2J}\mathbf{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{I1}\mathbf{B} & a_{I2}\mathbf{B} & \cdots & a_{IJ}\mathbf{B} \end{bmatrix},$$

and if $J = L$, $\mathbf{A} \odot \mathbf{B}$ is defined as $\mathbf{A} \odot \mathbf{B} = [\mathbf{a}_1 \otimes \mathbf{b}_1 \quad \mathbf{a}_2 \otimes \mathbf{b}_2 \quad \cdots \quad \mathbf{a}_J \otimes \mathbf{b}_J]$.

The CP decomposition of an Nth-order tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ factorizes the tensor into a sum of rank-one tensors. In general, the CP decomposition can be presented as:

$$\mathcal{X} \approx [\![\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \cdots, \mathbf{A}^{(N)}]\!] \equiv \sum_{r=1}^{R} \mathbf{a}_r^{(1)} \circ \mathbf{a}_r^{(2)} \circ \cdots \circ \mathbf{a}_r^{(N)},$$

where $\mathbf{A}^{(n)} \in \mathbb{R}^{I_n \times R}$ $(n \in \{1, \ldots, N\})$ denotes the nth factor matrix of the tensor $\mathcal{X}$, and R is a positive integer denoting the decomposition rank; $\mathbf{a}_r^{(n)} \in \mathbb{R}^{I_n}$ $(r \in \{1, \cdots, R\})$ denotes the rth column of the jth factor matrix. The symbol $\circ$ represents the outer product of vectors. It is often useful to assume that the columns of $\mathbf{A}^{(n)}$ are normalized to length one with the weights absorbed into the vector $\boldsymbol{\lambda} \in \mathbb{R}^R$ (Kolda & Bader, 2009) so that

$$\mathcal{X} \approx [\![\boldsymbol{\lambda}; \mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \cdots, \mathbf{A}^{(N)}]\!] \equiv \sum_{r=1}^{R} \lambda_r \mathbf{a}_r^{(1)} \circ \mathbf{a}_r^{(2)} \circ \cdots \circ \mathbf{a}_r^{(N)},$$

where $\lambda_r > 0$ is a singular value and $\|\mathbf{a}_r^{(n)}\|_2 = 1$. The mode-$n$ matricized version of the tensor $\mathcal{X}$ is expressed as:

$$\mathbf{X}_{(n)} \approx \mathbf{A}^{(n)} \boldsymbol{\Lambda} (\mathbf{A}^{(N)} \odot \cdots \odot \mathbf{A}^{(n+1)} \odot \mathbf{A}^{(n-1)} \odot \cdots \odot \mathbf{A}^{(1)})^\top,$$

where $\boldsymbol{\Lambda} = \mathrm{diag}(\boldsymbol{\lambda})$ is a diagonal matrix.

The Frobenius norm is denoted as $\|\cdot\|_F$. For a tensor $\mathcal{X}$, its Frobenius norm is written as $\|\mathcal{X}\|_F$, which can be calculated by $\|\mathcal{X}\|_F = \|\mathbf{X}_{(1)}\|_F$. The l_1 norm of a tensor $\mathcal{X}$ is denoted by $\|\mathcal{X}\|_1$ and is computed as the sum of the absolute value of its entries. The operation $\langle \mathbf{A}, \mathbf{B} \rangle$ represents the trace of the product of two matrices $\mathbf{A}^\top$ (transpose of $\mathbf{A}$) and $\mathbf{B}$, also denoted as $\mathrm{Tr}(\mathbf{A}^\top \mathbf{B})$.

2.2. Related work

Data fusion models, which aim to capture common variations in two or more datasets, have been developed for several decades (Gaw et al., 2022; Hotelling, 1992). Early research on data fusion mainly solved the problems of joint factorization of multiple matrices (Badea, 2008). Singh and Gordon (2008) propose a collective matrix factorization (CMF) to utilize correlations between different data and simultaneously factorize coupled matrices. Their proposed CMF model is formulated as follows: Given two matrices $\mathbf{X} \in \mathbb{R}^{I \times J}$ and $\mathbf{Y} \in \mathbb{R}^{I \times K}$ coupled in the first mode, the objective function is given as:

$$f(\mathbf{U}, \mathbf{V}, \mathbf{W}) = \|\mathbf{X} - \mathbf{U}\mathbf{V}^\top\|^2 + \|\mathbf{Y} - \mathbf{U}\mathbf{W}^\top\|^2, \qquad (1)$$

where $\mathbf{V} \in \mathbb{R}^{J \times R}$ and $\mathbf{W} \in \mathbb{R}^{K \times R}$ are specific factor matrices, $\mathbf{U} \in \mathbb{R}^{I \times R}$ is the shared factor matrix, and R is the rank. The above model can be extended to the factorization of matrices coupled in any mode. As an extension of problem (??), joint factorization of a third-order tensor $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$ coupled with a matrix $\mathbf{Y} \in \mathbb{R}^{I \times M}$ is proposed (Harshman & Lundy, 1994; Smilde et al., 2000). In this approach, a tensor and a matrix are decomposed simultaneously by minimizing the following objective function:

$$f(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{V}) = \|\mathcal{X} - [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]\|^2 + \|\mathbf{Y} - \mathbf{A}\mathbf{V}^\top\|^2, \qquad (2)$$

where $\mathbf{A} \in \mathbb{R}^{I \times R}$ is the shared factor matrix of $\mathcal{X}$ and $\mathbf{Y}$, $\mathbf{B} \in \mathbb{R}^{J \times R}$ and $\mathbf{C} \in \mathbb{R}^{K \times R}$ are factor matrices corresponding to the second and third modes of $\mathcal{X}$; and $\mathbf{V} \in \mathbb{R}^{M \times R}$ is a factor matrix of $\mathbf{Y}$. Other variations of CMTF have also been proposed (Acar et al., 2014). A more general form of CMTF factorizes an Nth order tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$

coupled with a matrix $\mathbf{Y} \in \mathbb{R}^{I_n \times M}$ in the nth mode ($n \in \{1, ..., N\}$) by minimizing an objective function as follows:

$$f(\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \cdots, \mathbf{A}^{(N)}, \mathbf{V}) = \|\mathcal{X} - [\![\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \cdots, \mathbf{A}^{(N)}]\!]\|^2 + \|\mathbf{Y} - \mathbf{A}^{(n)}\mathbf{V}^\top\|^2, \tag{3}$$

CMTF is also extended to coupled tensor and tensor factorization (Zhao et al., 2013), which simultaneously decomposes multiple coupled tensors into common and uncommon factor matrices. For example, if the matrix $\mathbf{Y} \in \mathbb{R}^{I_n \times M}$ is replaced by a higher-order tensor $\mathcal{Y} \in \mathbb{R}^{I_1 \times J_2 \times \cdots \times J_M}$ in (??), the coupled CP decomposition of tensors $\mathcal{X}$ and $\mathcal{Y}$ is obtained by minimizing the modified objective function as follows:

$$f(\mathbf{A}, \mathbf{V}^{(2)}, ..., \mathbf{V}^{(M)}) = \|\mathcal{X} - [\![\mathbf{A}^{(1)}, ..., \mathbf{A}^{(N)}]\!]\|^2 + \|\mathcal{Y} - [\![\mathbf{A}^{(1)}, \mathbf{V}^{(2)}, ..., \mathbf{V}^{(M)}]\!]\|^2. \tag{4}$$

Unfortunately, none of the proposed coupled decomposition methods can handle tensor data contaminated by arbitrary outliers. In the next section, we will propose a formulation of robust coupled tensor decomposition to extract robust factors and features.

3. Robust coupled CP decomposition framework

In this section, we introduce the robust coupled CP decomposition (RCCPD) model to jointly factorize tensors that are contaminated by arbitrary outliers. We decompose coupled tensors by extracting the outliers and fitting the CP decomposition model, simultaneously. We first consider a specific case of robust coupled tensor and matrix decomposition, where tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ and matrix $\mathbf{Y} \in \mathbb{R}^{I_1 \times J}$ are coupled in their first mode and both contain outliers. We decompose $\mathcal{X}$ as the sum of three tensors, i.e., $\mathcal{X} = \mathcal{L} + \mathcal{S} + \mathcal{E}$, where $\mathcal{L} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ is a smooth low-rank tensor without outliers, $\mathcal{S} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ is a sparse tensor that captures outliers, and $\mathcal{E} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ represents the tensor of errors. We consider similar decomposition of matrix $\mathbf{Y} = \mathbf{L} + \mathbf{S} + \mathbf{E}$, where $\mathbf{L} \in \mathbb{R}^{I_1 \times J}$, $\mathbf{S} \in \mathbb{R}^{I_1 \times J}$, and $\mathbf{E} \in \mathbb{R}^{I_1 \times J}$ are smooth, sparse, and error matrices, respectively. Assuming that $\mathcal{L}$ and $\mathbf{L}$ are coupled in their first mode and have the decomposition forms $\mathcal{L} = [\![\boldsymbol{\lambda}; \mathbf{A}^{(1)}, ..., \mathbf{A}^{(N)}]\!]$ and $\mathbf{L} = \mathbf{A}^{(1)}\mathbf{V}^\top$, the robust coupled CP decomposition of $\mathcal{X}$ and $\mathbf{Y}$ can be estimated by minimizing the following objective function $f(\mathbf{A}^{(1)}, ..., \mathbf{A}^{(N)}, \mathbf{V}, \mathcal{S}, \mathbf{S})$, denoted as f:

$$f = \frac{1}{2}\|\mathcal{X} - \mathcal{S} - [\![\boldsymbol{\lambda}; \mathbf{A}^{(1)}, ..., \mathbf{A}^{(N)}]\!]\|_F^2 + \alpha\|\mathcal{S}\|_1 + \frac{1}{2}\|\mathbf{Y} - \mathbf{S} - \mathbf{A}^{(1)}\mathbf{V}^T\|_F^2 + \beta\|\mathbf{S}\|_1, \tag{5}$$

where α and β are hyperparameters.

Next, we extend (5) to a robust coupled tensor-tensor decomposition form, by replacing matrix $\mathbf{Y} \in \mathbb{R}^{I_1 \times J}$ with a higher dimensional tensor $\mathcal{Y} \in \mathbb{R}^{I_1 \times J_2 \times \cdots \times J_M}$. Hereafter, we use subscript 1 for tensor $\mathcal{X}$ and subscript 2 for tensor $\mathcal{Y}$, i.e., $\mathcal{X} = \mathcal{L}_1 + \mathcal{S}_1 + \mathcal{E}_1$ and $\mathcal{Y} = \mathcal{L}_2 + \mathcal{S}_2 + \mathcal{E}_2$. Similar to the previous representation, $\mathcal{L}_2$, $\mathcal{S}_2$, and $\mathcal{E}_2$ represent a smooth tensor, a sparse tensor, and tensor of errors, respectively. Assuming (without loss of generality) that $\mathcal{L}_1$ and $\mathcal{L}_2$ are coupled in their first mode, Equation (5) can be extended as follows:

$$f = \frac{1}{2}\|\mathcal{X} - \mathcal{S}_1 - [\![\boldsymbol{\lambda}_1; \mathbf{A}^{(1)}, ..., \mathbf{A}^{(N)}]\!]\|_F^2 + \alpha\|\mathcal{S}_1\|_1 + \frac{1}{2}\|\mathcal{Y} - \mathcal{S}_2 - [\![\boldsymbol{\lambda}_2; \mathbf{A}^{(1)}, \mathbf{V}^{(2)}, ..., \mathbf{V}^{(M)}]\!]\|_F^2 + \beta\|\mathcal{S}_2\|_1, \tag{6}$$

where $\boldsymbol{\lambda}_1 \in \mathbb{R}^R$ and $\boldsymbol{\lambda}_2 \in \mathbb{R}^R$ are vectors that absorb weights obtained by normalizing factor matrices $\mathbf{A}^{(n)} \in \mathbb{R}^{I_n \times R}$ and $\mathbf{V}^{(m)} \in \mathbb{R}^{J_m \times R}$ $(n = 1, ..., N; m = 1, ..., M)$, respectively. Vectors $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$ are the joint robust features extracted from the data. These vectors identify how to linearly combine the factor matrices (basis vectors) to span the space of data. Therefore, for a fixed set of bases, these coefficients contain approximately all information within the data (analogous to coefficients of Fourier transformation). The first and third terms penalize the reconstruction error and the second and fourth terms enforce the sparsity of $\mathcal{S}_1$ and $\mathcal{S}_2$.

3.1. Alternating direction method of multipliers (ADMM) for model estimation

We aim to minimize the loss function in Equation (6) to estimate the factor matrices and sparse tensors. Specifically, we apply the alternating direction method of multipliers (ADMM) approach to solve this optimization problem. ADMM requires the differentiable and non-differentiable parts of the objective function to be separable (in terms of their variables), which is not the case in our original objective function. For example, $\mathcal{S}_1$ is in both the first (differentiable) and second (non-differentiable) terms of the objective function. To address this issue, we introduce two new auxiliary variables $\mathcal{F}$ and $\mathcal{M}$ and replace them with $\mathcal{S}_1$ and $\mathcal{S}_2$ in the second and fourth terms. This substitution requires adding two equality constraints $\mathcal{F} = \mathcal{S}_1$ and $\mathcal{M} = \mathcal{S}_2$ and produces the following formulation:

$$\begin{aligned} f = &\frac{1}{2}\|\mathcal{X} - \mathcal{S}_1 - [\![\boldsymbol{\lambda}_1; \mathbf{A}^{(1)}, ..., \mathbf{A}^{(N)}]\!]\|_F^2 + \alpha\|\mathcal{F}\|_1 \\ &+ \frac{1}{2}\|\mathcal{Y} - \mathcal{S}_2 - [\![\boldsymbol{\lambda}_2; \mathbf{A}^{(1)}, \mathbf{V}^{(2)}, ..., \mathbf{V}^{(M)}]\!]\|_F^2 + \beta\|\mathcal{M}\|_1, \\ &s.t. \quad \mathcal{F} = \mathcal{S}_1; \quad \mathcal{M} = \mathcal{S}_2. \end{aligned} \tag{7}$$

The corresponding augmented Lagrangian function $L_\rho(\mathbf{A}^{(1:N)}, \mathbf{V}^{(2:M)}, \mathcal{F}, \mathcal{S}_1, \mathcal{M}, \mathcal{S}_2, \mathcal{D}_1, \mathcal{D}_2)$ for problem (7), is constructed as:

$$\begin{aligned} L_\rho = &\frac{1}{2}\|\mathcal{X} - \mathcal{S}_1 - [\![\boldsymbol{\lambda}_1; \mathbf{A}^{(1)}, ..., \mathbf{A}^{(N)}]\!]\|_F^2 \\ &+ \frac{1}{2}\|\mathcal{Y} - \mathcal{S}_2 - [\![\boldsymbol{\lambda}_2; \mathbf{A}^{(1)}, \mathbf{V}^{(2)}, ..., \mathbf{V}^{(M)}]\!]\|_F^2 \\ &+ \alpha\|\mathcal{F}\|_1 + \beta\|\mathcal{M}\|_1 + \langle \mathcal{D}_1, \mathcal{F} - \mathcal{S}_1 \rangle \\ &+ \frac{\rho}{2}\|\mathcal{F} - \mathcal{S}_1\|_F^2 + \langle \mathcal{D}_2, \mathcal{M} - \mathcal{S}_2 \rangle + \frac{\rho}{2}\|\mathcal{M} - \mathcal{S}_2\|_F^2, \end{aligned} \tag{8}$$

where $\mathbf{A}^{(1:N)} = \{\mathbf{A}^{(i)}\}_{i=1}^{N}$, $\mathbf{V}^{(2:M)} = \{V^{(j)}\}_{j=2}^{M}$ are all factor matrices, $\mathcal{D}_1$ and $\mathcal{D}_2$ are the dual variables (or Lagrange multipliers) with the same dimensions as tensors $\mathcal{X}$ and $\mathcal{Y}$, respectively. Here, $\langle .,. \rangle$ indicates the inner product of two tensors and $\rho > 0$ is a penalty parameter.

The ADMM algorithm iteratively updates the eight sets of variables by computing the partial (sub) derivative of the augmented Lagrangian function with respect to each variable. Specifically, to estimate $\mathbf{A}^{(n)}$, we fix all other variables and solve $\mathbf{A}^{(n)}$ by minimizing $L_\rho(\mathbf{A}^{(n)})$, which is:

$$\begin{aligned}\underset{\mathbf{A}^{(n)}}{\operatorname{argmin}} \frac{1}{2}\{\|\mathcal{X} - \mathcal{S}_1 - [\![\boldsymbol{\lambda}_1; \mathbf{A}^{(1)}, ..., \mathbf{A}^{(N)}]\!]\|_F^2 \\ + \|\mathcal{Y} - \mathcal{S}_2 - [\![\boldsymbol{\lambda}_2; \mathbf{A}^{(1)}, \mathbf{V}^{(2)}, ..., \mathbf{V}^{(M)}]\!]\|_F^2\}.\end{aligned} \tag{9}$$

We derive the solution for problem (9) as follows:

$$(\mathbf{A}^{(n)}\boldsymbol{\Lambda}_1) = \begin{cases} ((\mathcal{X} - \mathcal{S}_1)_{(1)}(\odot\mathbf{A}^{(-1)}) + (\mathcal{Y} - \mathcal{S}_2)_{(1)}(\odot\mathbf{V}^{(-1)}))\mathbf{Q}^\top(\mathbf{Q}\mathbf{Q}^\top)^{-1}, & n = 1 \\ (\mathcal{X} - \mathcal{S}_1)_{(n)}(\odot\mathbf{A}^{(-n)})(\boldsymbol{\Gamma}^{(n)})^\top(\boldsymbol{\Gamma}^{(n)}(\boldsymbol{\Gamma}^{(n)})^\top)^{-1}, & n \geq 1 \end{cases} \tag{10}$$

where $\boldsymbol{\Lambda}_1$ is a diagonal matrix whose (j, j)th element is the jth element of $\boldsymbol{\lambda}_1$; $\mathbf{Q} = \boldsymbol{\Omega}^{(1)} + \boldsymbol{\Gamma}^{(1)}$; $(\odot\mathbf{A}^{(-n)}) = \mathbf{A}^{(N)} \odot \cdots \odot \mathbf{A}^{(n+1)}\mathbf{A}^{(n-1)} \odot \cdots \odot \mathbf{A}^{(1)}$; $(\odot\mathbf{V}^{(-m)}) = \mathbf{V}^{(M)} \odot \cdots \odot \mathbf{V}^{(m+1)}\mathbf{V}^{(m-1)} \odot \cdots \odot \mathbf{A}^{(1)}$; $\boldsymbol{\Gamma}^{(n)} = (\mathbf{A}^{(1)\top}\mathbf{A}^{(1)}) * \cdots * (\mathbf{A}^{(n-1)\top}\mathbf{A}^{(n-1)})(\mathbf{A}^{(n+1)\top}\mathbf{A}^{(n+1)}) * \cdots * (\mathbf{A}^{(N)\top}\mathbf{A}^{(N)})$; $\boldsymbol{\Omega}^{(n)} = (\mathbf{A}^{(1)\top}\mathbf{A}^{(1)}) * \cdots * (\mathbf{V}^{(m-1)\top}\mathbf{V}^{(m-1)})\ (\mathbf{V}^{(m+1)\top}\mathbf{V}^{(m+1)}) * \cdots * (\mathbf{V}^{(M)\top}\mathbf{V}^{(M)})$. Here, $A * B$ denotes elementwise multiplication of two matrices with the same dimensions.

Next, assuming all other variables are known, we update $\mathbf{V}^{(m)}$ by minimizing $L_\rho(\mathbf{V}^{(m)})$:

$$\underset{\mathbf{V}^{(m)}}{\operatorname{argmin}} \left\{\frac{1}{2}\|\mathcal{Y} - \mathcal{S}_2^t - [\![\boldsymbol{\lambda}_2; \mathbf{A}^{(1)}, \mathbf{V}^{(2)}, ..., \mathbf{V}^{(M)}]\!]\|_F^2\right\}, \tag{11}$$

which results in,

$$(\mathbf{V}^{(m)}\boldsymbol{\Lambda}_2) = (\mathcal{Y} - \mathcal{S}_2)_{(m)}(\odot\mathbf{V}^{(-m)})(\boldsymbol{\Omega}^{(m)})^\top(\boldsymbol{\Omega}^{(m)}(\boldsymbol{\Omega}^{(m)})^\top)^{-1}, \tag{12}$$

where $\boldsymbol{\Lambda}_2$ is a diagonal matrix obtained from $\boldsymbol{\lambda}_2$.

Let us denote $\widetilde{\mathbf{A}^{(n)}} = \mathbf{A}^{(n)}\boldsymbol{\Lambda}_1$ and $\widetilde{\mathbf{V}^{(m)}} = \mathbf{V}^{(m)}\boldsymbol{\Lambda}_2$. The original factor matrices $\mathbf{A}^{(n)}$ and $\mathbf{V}^{(m)}$ are obtained by normalizing the columns of $\widetilde{\mathbf{A}^{(n)}}$ and $\widetilde{\mathbf{V}^{(m)}}$, respectively. Then we take the norms of their ith column as the ith element of vectors $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$.

Next, the variable $\mathcal{F}$ is updated by minimizing $L_\rho(\mathcal{F})$, the solution of $\mathcal{F}$ is:

$$\begin{aligned}\mathcal{F} &= \underset{\mathcal{F}}{\operatorname{argmin}} \left\{\alpha\|\mathcal{F}\|_1 + \langle\mathcal{D}_1, \mathcal{F} - \mathcal{S}_1\rangle + \frac{\rho}{2}\|\mathcal{F} - \mathcal{S}_1\|_F^2\right\} \\ &= \operatorname{sign}(\mathcal{S}_1 - \rho^{-1}\mathcal{D}_1) * (|\mathcal{S}_1 - \rho^{-1}\mathcal{D}_1| - \alpha\rho^{-1})_+,\end{aligned} \tag{13}$$

where $\operatorname{sign}(x) = 0$ *if* $x = 0$ and $\operatorname{sign}(x) = \frac{x}{|x|}$ *if* $x \neq 0$, and $(x)_+ = max\{0, x\}$, which is applied element-wise to a tensor. The notation $*$ is the element-wise tensor product, and scalar subtraction is for each element of a tensor.

Next, we update $\mathcal{S}_1$ by minimizing $L_\rho(\mathcal{S}_1)$ as follows:

$$\begin{aligned}\underset{\mathcal{S}_1}{\operatorname{argmin}} \Bigg\{ &\frac{1}{2}\|\mathcal{X} - \mathcal{S}_1 - [\![\boldsymbol{\lambda}_1; \mathbf{A}^{(1)}, ..., \mathbf{A}^{(N)}]\!]\|_F^2 + \langle\mathcal{D}_1, \mathcal{F} - \mathcal{S}_1\rangle \\ &+ \frac{\rho}{2}\|\mathcal{F} - \mathcal{S}_1\|_F^2\Bigg\},\end{aligned} \tag{14}$$

which results in,

$$\mathcal{S}_1 = (\mathcal{X} - [\![\boldsymbol{\lambda}_1; \mathbf{A}^{(1)}, ..., \mathbf{A}^{(N)}]\!] + \mathcal{D}_1 + \rho\mathcal{F})(1 + \rho)^{-1}. \tag{15}$$

By solving the problem $L_\rho(\mathcal{M})$, we can obtain the solution of $\mathcal{M}$ as follows:

$$\begin{aligned}\mathcal{M} &= \underset{\mathcal{M}}{\operatorname{argmin}} \left\{\beta\|\mathcal{M}\|_1 + \langle\mathcal{D}_2, \mathcal{M} - \mathcal{S}_2\rangle + \frac{\rho}{2}\|\mathcal{M} - \mathcal{S}_2\|_F^2\right\} \\ &= \operatorname{sign}(\mathcal{S}_2 - \rho^{-1}\mathcal{D}_2) * (|\mathcal{S}_2 - \rho^{-1}\mathcal{D}_2| - \beta\rho^{-1})_+.\end{aligned} \tag{16}$$

Derivations of soft-thresholding operations for updating $\mathcal{F}$ and $\mathcal{M}$ are shown in Appendix A.

Then, we update $\mathcal{S}_2$ by solving $L_\rho(\mathcal{S}_2)$ as follows:

$$\begin{aligned}\underset{\mathcal{S}_2}{\operatorname{argmin}} \Bigg\{ &\frac{1}{2}\|\mathcal{Y} - \mathcal{S}_2 - [\![\boldsymbol{\lambda}_2; \mathbf{A}^{(1)}, \mathbf{V}^{(2)}, ..., \mathbf{V}^{(M)}]\!]\|_F^2 \\ &+ \langle\mathcal{D}_2, \mathcal{M} - \mathcal{S}_2\rangle + \frac{\rho}{2}\|\mathcal{M} - \mathcal{S}_2\|_F^2\Bigg\},\end{aligned} \tag{17}$$

which results in,

$$\mathcal{S}_2 = (\mathcal{Y} - \mathbf{A}^{(1)}(\odot\mathbf{V}^{(-1)})^\top + \mathcal{D}_2 + \rho\mathcal{M})(1 + \rho)^{-1}. \tag{18}$$

Finally, we update the dual variables $\mathcal{D}_1$ and $\mathcal{D}_2$ using the following formulas:

$$\mathcal{D}_1 = \mathcal{D}_1 + \rho(\mathcal{F} - \mathcal{S}_1), \quad \mathcal{D}_2 = \mathcal{D}_2 + \rho(\mathcal{M} - \mathcal{S}_2). \tag{19}$$

In these update equations, the most updated estimates of the variables are used. The procedure for the proposed RCCPD model is summarized in Algorithm 1.

Algorithm 1. ADMM Solver for Robust Coupled CP Tensor Decomposition

Input: Coupled tensors $\mathcal{X}$ and $\mathcal{Y}$, hyperparameters α and β, rank R, and Scaling $k > 1$
Initialization:
$\mathbf{A}^{(n)}$, $\mathbf{V}$, $\mathcal{F} = \mathcal{S}_1 = \mathcal{Y}_1 = 0$; $\mathcal{M} = \mathcal{S}_2 = \mathcal{Y}_2 = 0$, $\rho > 0$
while not converged **do**
  for $n \in \{1, ..., N\}$ **do**
    Update $\widetilde{\mathbf{A}^{(n)}}$ based on Equation (10)
    for $j \in \{1, ..., R\}$ **do**
      Normalize columns of $\widetilde{\mathbf{A}^{(n)}}$, $\lambda_{1j} = \|\widetilde{\mathbf{A}^{(n)}}(:, j)\|$
      $\mathbf{A}^{(n)}(:, j) = \widetilde{\mathbf{A}^{(n)}}(:, j)\lambda_{1j}$
    end for
  end for
  for $m \in \{1, ..., M\}$ **do**
    $\widetilde{\mathbf{V}^{(m)}}$ based on Equation (12)
    for $j \in \{1, ..., R\}$ **do**

15: Normalize columns of $\widetilde{\mathbf{V}^{(m)}}$, $\lambda_{2j} = \|\widetilde{\mathbf{V}^{(m)}}(:,j)\|$
16: $\mathbf{V}^{(m)}(:,j) = \widetilde{\mathbf{V}^{(m)}}(:,j)\lambda_{2j}$
17: **end for**
18: **end for**
19: Update $\mathcal{F}$ based on Equation (13)
20: Update $\mathcal{S}_1$ based on Equation (15)
21: Update $\mathcal{M}$ based on Equation (16)
22: Update $\mathcal{S}_2$ based on Equation (18)
23: Update $\mathcal{D}_1, \mathcal{D}_2$ based on Equation (19)
24: $\rho = \rho \times k$
25: **end while**
26: **return** $\mathbf{A}^{(n)}, \mathbf{V}^{(m)}, \mathcal{S}_1, \mathcal{S}_2, \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2$

3.2. General settings and tuning parameter selection

We use the primal and dual stopping criteria of ADMM algorithms. More specifically, the primal criteria ensure the feasibility of the solution by evaluating $\frac{\|\mathcal{F}^{t+1}-\mathcal{S}_1^{t+1}\|_F}{\|\mathcal{S}_1^{t+1}\|_F}$ and $\frac{\|\mathcal{M}^{t+1}-\mathcal{S}_2^{t+1}\|_F}{\|\mathcal{S}_2^{t+1}\|_F}$. The dual criteria ensure the convergance of the algorithm by assessing $\frac{\rho*\|\mathcal{F}^{t+1}-\mathcal{F}^{t}\|_F}{\|\mathcal{F}^{t+1}\|_F}$ and $\frac{\rho*\|\mathcal{M}^{t+1}-\mathcal{M}^{t}\|_F}{\|\mathcal{M}^{t+1}\|_F}$. Here, $t+1$ denotes the current iteration and t denotes the previous iteration. We stop the algorithm if the above values are smaller than a threshold (δ) or a maximum number of iterations is reached. We have set $\delta = 10^{-6}$ and the maximum number of iterations to 1000. The initial value of ρ is set to 10^{-3}, which is geometrically increased by a constant $k = 1.2$ up to 10^8.

Setting the hyperparameters α, β and the tensor rank R depends on the goal of the problem at hand. If the goal is to accurately decompose the coupled tensors (unsupervised model), then a set of hyperparameters that minimize the tensor reconstruction error is of interest. This is achieved by using Bayesian optimization (BO), which internally maintains a Gaussian process regression to train the model. At each search iteration in the BO (i.e., for each set of selected parameters), the ADMM method is executed and returns the reconstructed tensor and the reconstruction error. In this work, the built-in function "bayesopt" in MATLAB is applied with the appropriate ranges of the parameters to select the set of parameters with the smallest tensor reconstruction error. The reconstruction error is defined as:

$$TRS = \frac{\|\mathcal{X} - \hat{\mathcal{S}} - \hat{\mathcal{L}}\|_F}{\|\mathcal{X}\|_F}. \tag{20}$$

where $\hat{\mathcal{L}}$ represents the smooth reconstruction of tensor $\mathcal{X}$ based on the estimated factor matrices, e.g., $\hat{\mathcal{L}} = [\![\hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\mathbf{C}}]\!]$, and $\hat{\mathcal{S}}$ is the estimated sparse tensor of outliers.

The simulation studies in this paper follow the BO method for selecting the parameters, where the range of α, β and R are set as: $\alpha \in [0.05, 1]$, $\beta \in [0.05, 1]$ and $R \in \{2, 3, 4, 5, 6, 7, 8, 9, 10\}$ in the BO algorithm. The rank values larger than 10 resulted in singularity issues and are not considered.

In the cases where the extracted features are used for prediction purposes, the hyperparameters are selected to maximize the performances of the predictive models. For this purpose, the data is divided into the training and testing sets. Next, several models for various combinations of hyperparameters are constructed. The performance of these models are then evaluated on the testing data set to select the model with the highest prediction performance. Our case studies follow this approach and use a grid search over $\alpha \in [0.1, 1]$, $\beta \in [0.1, 1]$, and $R \in \{2, 3, 4, ..., 16\}$. Similar approaches are used for tuning the hyperparameters of the benchmarks. The hyperparameters of each benchmark are identified in the next section.

4. Performance evaluation using simulations

In this Section, we evaluate the effectiveness of the proposed method using two simulated experiments. In the first simulation, we consider a coupled tensor and matrix scenario. The second simulation evaluates the performance of RCCPD in decomposing two coupled higher-order tensors. In both simulations (and case studies), we compare the proposed method to the following three benchmarks.

(a) CPD: This benchmark is the basic CP decomposition model, which factorizes each tensor individually. The CP model for higher order tensors is a generalization of the matrix singular value decomposition (SVD) to higher order tensors. Please note that while CPD has a uniqueness property under the Kruskal condition for higher order tensors (Kolda & Bader, 2009), it does not produce unique decomposition for matrices. Particularly, it only achieves unique singular values. CPD is implemented in the Tensorlab 3.0 toolbox (Vervliet et al., 2016) in MATLAB. The tensor rank R is the only hyperparameter to be determined in this benchmark.

(b) CMTF/CTTF: This benchmark is the coupled matrix-tensor/tensor-tensor factorization, which decomposes coupled tensors (or matrices) simultaneously. It solves the objective functions in equation (1) (2) (3), and (4). CMTF is implemented in the CMTF toolbox in MATLAB (Acar et al., 2011); CTTF is implemented by setting the updates of the two outliers $\mathcal{S}_1$ and $\mathcal{S}_2$ in our proposed approach to zeroes. The tensor rank R is the only hyperparameter to be determined in this benchmark.

(c) TRCPD: This benchmark is the robust CP decomposition of a tensor (Xue et al., 2017) that decomposes a tensor into a low-rank tensor and a sparse tensor, which contains the outliers. TRCPD is implemented by following this work (Xue et al., 2017). This benchmark has two hyperparameters, i.e., the parameter λ for imposing the sparsity of the outlier tensor and the tensor rank R, to be determined. The parameter λ is tuned as suggested in their work.

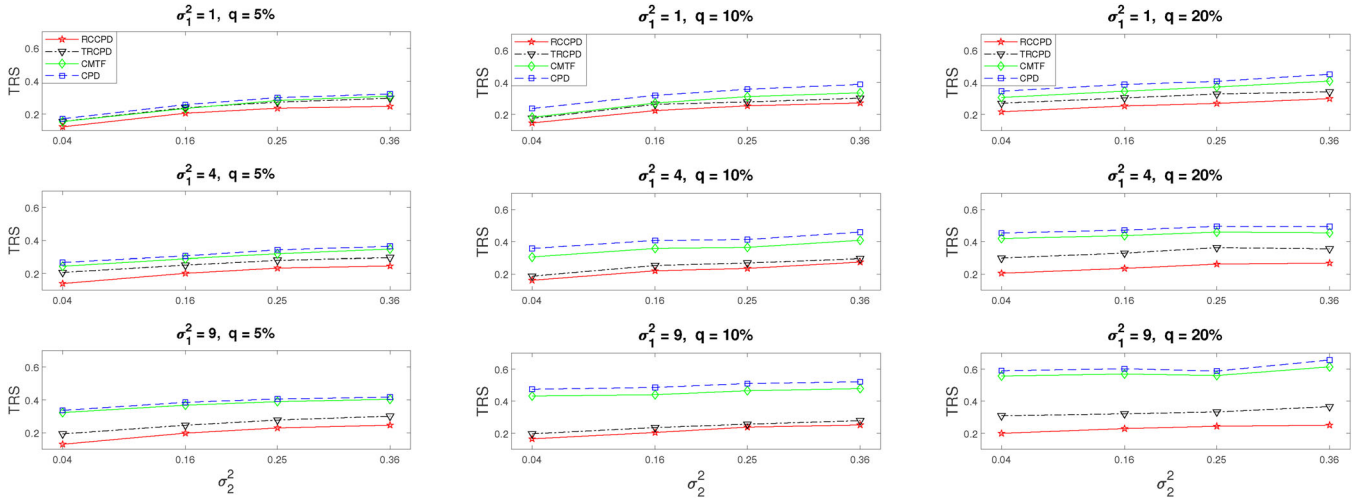


Figure 2. Comparison of the proposed method with benchmarks in terms of tensor reconstruction scores (TRS) for different σ_1^2, σ_2^2, and q.

In order to compare the proposed method to benchmarks, the above-defined tensor reconstruction score (TRS) is used.

4.1. Simulation I: coupled tensor and matrix decomposition

We first simulate a third-order tensor $\mathcal{L} \in \mathbb{R}^{I\times J\times K}$ and a matrix $\mathbf{L} \in \mathbb{R}^{I\times M}$ by generating factor matrices $\mathbf{A} \in \mathbb{R}^{I\times R}, \mathbf{B} \in \mathbb{R}^{J\times R}, \mathbf{C} \in \mathbb{R}^{K\times R}$, and $\mathbf{V} \in \mathbb{R}^{M\times R}$, whose entries are randomly drawn from a standard normal distribution. The factor matrices are used to construct a third-order tensor $\mathcal{L} = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$ coupled with a matrix $\mathbf{L} = \mathbf{A}\mathbf{V}^T$. Then, a sparse tensor $\mathcal{S} \in \mathbb{R}^{I\times J\times K}$ and a sparse matrix $\mathbf{S} \in \mathbb{R}^{I\times M}$ are generated as follows. First, we simulate all the entries from a normal distribution with mean zero and variance σ_1^2. Next, we randomly set $1-q$ of their entries to zero. These sparse tensors are added to the initial ones ($\mathcal{L}$ and $\mathbf{L}$) to represent q percent of outliers. Then, the small Gaussian noise $\mathcal{E} \in \mathbb{R}^{I\times J\times K}$ and $\mathbf{E} \in \mathbb{R}^{I\times M}$, with mean zero and variance σ_2^2 are separately added to the generated data. Finally, the tensors with outliers are denoted as $\mathcal{X} = \mathcal{L} + \mathcal{S} + \mathcal{E}$ and $\mathbf{Y} = \mathbf{L} + \mathbf{S} + \mathbf{E}$. Since the outlier ratio q, the outlier variance σ_1^2, and the noise variance σ_2^2 are important to the performance of models, different values of each of them are considered. In this simulation, the above parameters are set as follows: $I = J = K = 10$, $M = 20$, $R = 4$, $q \in \{5\%, 10\%, 20\%\}$, $\sigma_1^2 \in \{1, 4, 9\}$, and $\sigma_2^2 \in \{0.04, 0.16, 0.25, 0.36\}$.

Once the coupled data are generated, they are (jointly) factorized using the benchmarks CPD, CMTF, TRCPD, and the proposed RCCPD. Note that the benchmarks CPD and TRCPD decompose tensor $\mathcal{X}$ and matrix $\mathbf{Y}$ separately while CMTF and RCCPD factorize them simultaneously. Figure 2 demonstrates the average TRS (over 50 runs) achieved by each method at different values of σ_1^2, σ_2^2 and q. The reported results are computed based on the minimum TRS values of each method achieved by tuning their corresponding hyperparameters. For the proposed RCCPD, the average of α, β, and rank R are found by BO as 0.0512, 0.0531, and 9, respectively. For the benchmark CMTF, $R = 10$ is selected. The ranks for tensor $\mathcal{X}$ and matrix $\mathbf{Y}$ are selected to be 9 and 8 for the benchmark CPD, and 8 and 6 for the benchmark TRCPD. The parameter λ in TRCPD is set as 0.0509. As it is illustrated, the proposed RCCPD (in red line) outperforms all benchmarks at all levels of the outlier ratios and variances. In other words, the results demonstrate the limitations of the CPD and CMTF models in identifying outliers within tensors and the limitation of TRCPD in utilizing the information from multimodal data, which result in higher reconstruction errors of the tensor compared to the proposed method. Since our robust coupled tensor decomposition method is able to identify outliers, it generates a more accurate decomposition. For example, when $\sigma_1^2 = 9$, $q = 10\%$, and $\sigma_2^2 = 0.25$, the TRS of the proposed method is 0.2375, which is significantly smaller than the TRS achieved by TRCPD(0.2667), CPD(0.5104), and CMTF(0.4662). Please note that the higher the outlier variance and ratio are, the higher the TRS for CPD, CMTF, and TRCPD while the proposed RCCPD maintains similar performance.

4.2. Simulation II: coupled tensor and tensor decomposition

In simulation II, we generate two third-order tensors $\mathcal{L}_1 \in \mathbb{R}^{10\times 20\times 30}$ and $\mathcal{L}_2 \in \mathbb{R}^{10\times 20\times 10}$ using the same steps described in simulation I, i.e., generate factor matrices $\mathbf{A} \in \mathbb{R}^{10\times 4}, \mathbf{B} \in \mathbb{R}^{20\times 4}, \mathbf{C} \in \mathbb{R}^{30\times 4}$, $\mathbf{D} \in \mathbb{R}^{20\times 4}$ and $\mathbf{V} \in \mathbb{R}^{10\times 4}$, with entries from the standard normal distribution. The two tensors are constructed by $\mathcal{L}_1 = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$ and $\mathcal{L}_2 = [\![\mathbf{A}, \mathbf{D}, \mathbf{V}]\!]$. Similarly, two sparse tensors $\mathcal{S}_1 \in \mathbb{R}^{10\times 20\times 30}$, $\mathcal{S}_2 \in \mathbb{R}^{10\times 20\times 10}$ and two Gaussian noise $\mathcal{E}_1 \in \mathbb{R}^{10\times 20\times 30}$, $\mathcal{E}_2 \in \mathbb{R}^{10\times 20\times 10}$ are generated as before and added to $\mathcal{L}_1$ and $\mathcal{L}_2$, respectively. Finally, the two tensors $\mathcal{X}$ and $\mathcal{Y}$ with outliers are generated by $\mathcal{X} = \mathcal{L}_1 + \mathcal{S}_1 + \mathcal{E}_1$ and $\mathcal{Y} = \mathcal{L}_2 + \mathcal{S}_2 + \mathcal{L}_2$. The variances of added outliers and noise are set as $\sigma_1^2 \in \{1, 4, 9\}$ and $\sigma_2^2 \in \{0.04, 0.16, 0.25, 0.36\}$. The ratio of outliers is set to $q = 10\%$.

Table 1. Comparison of four methods in terms of averages and standard deviations of tensor reconstruction scores (TRS) of each tensor for different σ_1^2, σ_2^2 with the outlier ratio $q = 10\%$.

		$\mathcal{X}$				$\mathcal{Y}$			
σ_1^2	σ_2^2	RCCPD	TRCPD	CTTF	CPD	RCCPD	TRCPD	CTTF	CPD
1	0.04	0.091 0.01	0.104 0.01	0.183 0.02	0.187 0.02	0.089 0.01	0.107 0.01	0.179 0.01	0.183 0.02
	0.16	0.131 0.01	0.147 0.02	0.234 0.03	0.238 0.02	0.128 0.01	0.145 0.01	0.216 0.03	0.227 0.02
	0.25	0.148 0.01	0.155 0.02	0.275 0.03	0.278 0.02	0.144 0.01	0.152 0.02	0.238 0.04	0.257 0.02
	0.36	0.155 0.02	0.162 0.02	0.287 0.02	0.292 0.02	0.152 0.02	0.163 0.02	0.273 0.04	0.289 0.03
4	0.04	0.095 0.01	0.113 0.01	0.277 0.03	0.287 0.01	0.104 0.01	0.112 0.01	0.262 0.03	0.268 0.02
	0.16	0.129 0.01	0.138 0.01	0.321 0.04	0.323 0.02	0.126 0.01	0.137 0.01	0.307 0.03	0.318 0.02
	0.25	0.144 0.01	0.151 0.03	0.352 0.04	0.357 0.02	0.143 0.01	0.153 0.01	0.335 0.03	0.344 0.03
	0.36	0.146 0.01	0.162 0.02	0.369 0.03	0.375 0.03	0.147 0.02	0.165 0.02	0.358 0.04	0.362 0.03
9	0.04	0.114 0.02	0.129 0.02	0.386 0.02	0.392 0.03	0.119 0.01	0.128 0.01	0.370 0.04	0.375 0.03
	0.16	0.143 0.01	0.157 0.01	0.409 0.02	0.413 0.03	0.140 0.01	0.154 0.01	0.382 0.03	0.391 0.03
	0.25	0.147 0.01	0.168 0.02	0.425 0.04	0.429 0.04	0.148 0.02	0.167 0.02	0.408 0.03	0.417 0.04
	0.36	0.152 0.02	0.181 0.02	0.449 0.03	0.455 0.04	0.153 0.02	0.179 0.02	0.422 0.03	0.432 0.04

The experiments are repeated 50 times and the averages and standard deviations of tensor reconstruction scores (TRS) are computed for both tensors. For RCCPD, the average of α, β, and rank R are found by BO as 0.0542, 0.0568, and 8, respectively. The parameters for TRCPD are $R = 9$ (for both tensors $\mathcal{X}$ and $\mathcal{Y}$) and $\lambda = 0.0564$, the ranks for CTTF and CPD are set as 8 and 10 (for both tensors $\mathcal{X}$ and $\mathcal{Y}$), respectively. Results are reported in Table 1.

The results show the superior performance of our proposed method compared to benchmarks in terms of TRS corresponding to each tensor, under all settings of σ_1^2 and σ_2^2. For example, when $\sigma_1^2 = 9$ and $\sigma_2^2 = 0.16$, the TRS of tensor $\mathcal{X}$ achieved by RCCPD is 0.143, which is smaller than 0.157, 0.409 and 0.413 achieved by TRCPD, CTTF and CPD. The distinguished performance of RCCPD is due to its capability in isolating outliers from the initial data.

Finally, the average execution times of each method across 50 replications are reported in Table 2. This execution time is acquired for a given set of hyperparameters (as reported above) and the outlier ratio of $q = 10\%$, the outlier variance of $\sigma_1^2 = 4$, and the noise variance of $\sigma_2^2 = 0.04$. As it is reported in Table 2, all methods show a comparable running time when applied to decomposing two tensors. More specifically, TRCPD requires around 0.2856 seconds to perform robust decomposition of two tensors separately. The proposed approach requires 0.3103 seconds on average to simultaneously perform robust decomposition of both tensors. CTTF and CP require less computational time. This is expected as they are not robust and do not require estimating the tensor of outliers. It should also be noted that the proposed method and TRCPD have larger number of hyperparameters and may require more computational effort for tuning the algorithm. Nevertheless, this step is often performed offline with massive computational power and therefore it is not restrictive.

Table 2. Average execution time of each method applied to one specific simulated data.

	RCCPD	TRCPD	CTTF	CP
Time (s)	0.3103	0.2856	0.2018	0.0735

5. Case studies

In this Section, we evaluate the performance of the proposed method in two real-world case studies. In the first case study (i.e., PD telemonitoring) coupled features are extracted from the participants' voice and tapping data and used for health condition estimation. In this case study, we organize the data into two coupled second-order tensors and extract features using the proposed method and benchmarks. Next, we apply support vector regression (SVR) and random forest regression (RFR) to create the prediction models. The second case study analyzes coupled third-order EEG and second-order fMRI tensors to classify the trials of an oddball auditory experiment. First, each method is applied to extract features and then support vector machine (SVM) and random forest classifier (RFC) are used to construct the classification models. In both cases, the three tensor-based methods (CPD, CMTF, and TRCPD) and two general SVR/SVM and RFR/RFC models are implemented as benchmarks. In the case studies, we apply TRCPD to individual tensors as well as to coupled tensors. That is, we first remove outliers and then extract features using the coupled decomposition. The random forest model was trained in MATLAB using the "TreeBagger" function in the Statistics and Machine Learning Toolbox (setting the number of trees to 100). For the SVR/SVM model, we employed the LIBSVM toolbox in MATLAB developed by Chang and Lin (2011).

5.1. Case study I: Parkinson's disease telemonitoring

Telemonitoring is a form of mobile health that uses electronic devices to monitor patients remotely. In recent years, there has been a surge in smartphone usage for telemonitoring. In fact, 85% of American adults own a smartphone according to a 2021 Pew Research Survey (Pew Research Center, 2021). A smartphone is equipped with various sensors, such as an accelerometer, gyroscope, camera, and microphone. Using custom-designed apps, smartphones can collect abundant health data from the users. For this case study, we focus on smartphone-based telemonitoring of Parkinson's Disease (PD). PD affects 7-10 million people worldwide and is the second most common neurodegenerative disorder (after Alzheimer's Disease) (Parkinsons News Today, 2020). PD patients suffer from tremors, voice impairment, and movement disorders. Although there is currently

Table 3. Information of patient ID and numbers of records of each patient.

Patient ID	No. of records	Patient ID	No. of records	Patient ID	No. of records
1	470	12	395	25	144
2	423	13	315	27	137
4	451	15	250	28	148
5	354	16	289	29	213
6	236	17	252	30	181
7	377	18	144	33	163
8	277	19	279	34	162
9	337	20	228	35	148
10	198	22	157	36	160
11	250	23	263	37	165

no cure for PD, the progression of PD can be significantly reduced with effective treatment, for which timely monitoring and assessment is key. One common clinical score to measure PD severity is the Movement Disorder Society Unified Parkinson's Disease Rating Scale (MDS-UPDRS) (Goetz et al., 2008; Yoon & Gaw, 2021). MDS-UPDRS is obtained from a 65-question survey that is administered to patients at a specialized clinic. It is difficult to maintain the most up-to-date information on the patient's disease severity, as this would require the patient's physical presence in the clinic on a frequent basis. Most commonly, patients make clinical visits only every 4-6 months. This scenario is even more difficult for patients living in remote areas, which have limited availability for specialized care. Fortunately, there is a great opportunity to improve this situation with emerging smartphone-enabled telemonitoring technologies.

The ultimate goal of this research area is to build a model that can predict the MDS-UDPRS score using smartphone-collected activity data. The predicted score could then be used for clinical assessment to measure PD severity without requiring the patient's physical presence at a clinic. Having such a model would allow for significantly improved patient convenience, while also providing frequent monitoring and assessment of the disease. mPower is prominent among the smartphone apps created for telemonitoring of PD (Bot et al., 2016). This is largely due to mPower's broad use of many smartphone sensors (e.g., microphone, accelerometer, gyroscope, etc.) to collect patient data across different tasks. mPower leads the user across several pre-designed activities (such as speaking and tapping) to measure major symptoms of PD. In order to best utilize the data collected from these activities, a model is needed to connect the variation in activity data to a disease severity score.

There are some challenging issues in building the aforementioned predictive model. First, the data collected from the various smartphone tasks is high-dimensional, and it is difficult to condense the data into an accurate model for MDS-UPDRS prediction. Additionally, there is a time component associated with each of the patients that should be taken into account. Lastly, each patient is unique, and most smartphone tasks are not monitored by a clinician, so there is a high chance that outliers can drastically reduce the performance of a model trained on this data.

In this case study, we evaluate the performance of the proposed method in extracting features for predicting MDS-UDPRS. A subset of thirty patients' tapping and voice data collected from mPower is considered. These are the patients who have MDS-UPDRS scores for at least three months and completed tapping and speaking tasks at least once a day. Each of the thirty patients has 137 to 470 records. Each record includes 339 voice features and 43 tapping features. The voice features were extracted from the speaking time series data (see Tsanas et al. (2011); Tsanas (2012) and https://github.com/ThanasisTsanas/VoiceAnalysisToolbox) and characterize amplitude (shimmer variants), frequency (jitter variants), increased noise (signal-to-noise measures), etc. The tapping features measure tapping speed, inter-tap interval, position, fatigue, etc. (see Chaibub Neto et al. (2016) and https://github.com/Sage-Bionetworks/personalized_hypothesis_tests). Numbers of records collected from each patient are provided in Table 3.

To represent the data as two matrices (i.e., voice features $\times$ time and tapping features $\times$ time) coupled along their time mode, we first randomly sample (keeping the time order) 100 records of each of the 30 patients forty times. This process results in 1200 pairs of coupled matrices with sizes 100×339 and 100×43, for voice and tapping features, respectively. While, the re-sampling step is not necessary, it creates a large balanced set of samples from which features are extracted. In addition, it serves as a demonstration that the proposed method can handle a large number of samples that contain tensors with higher dimensions. Note that coupling the matrices along the time mode is reasonable since the two activities (tapping and speaking) are often performed within a few seconds of each other. Therefore, given that the time resolution of records is in the order of hours, one can assume the two activities within a record appeared simultaneously. Finally, we take the average of MDS-UPDRS values over the 100 records and assign it as an output to the coupled matrices. Since the measurements are of different orders of magnitude, data normalization is applied before evaluating the methods. Next, we apply the proposed and benchmark methods to extract features used for predicting the average MDS-UPDRS. Apart from the benchmarks used in Section 4 that extract features from data, we also implement two general regression models: support vector regression (SVR) and random forest regression (RFR) as benchmarks in this case study. For the four tensor-based methods RCCPD, TRCPD, CMTF, and CPD, we first split the data by a K-fold technique and use one of the folds to estimate the factor matrices of the decomposition. Next, given the factor matrices, we estimate the features (e.g., $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$) for each sample in the rest of the $K-1$ folds. These features are the inputs to the SVR and RFR models. In our proposed method, if a new pair of coupled matrices is denoted by $\mathbf{X}_{\text{new}}$ and $\mathbf{Y}_{\text{new}}$, their feature vectors $\boldsymbol{\lambda}_1^{\text{new}} = [\lambda_{11}^{\text{new}}, ..., \lambda_{1R}^{\text{new}}]^\top$ and $\boldsymbol{\lambda}_2^{\text{new}} = [\lambda_{21}^{\text{new}}, ..., \lambda_{2R}^{\text{new}}]^\top$ can be obtained by: $\boldsymbol{\lambda}_1^{\text{new}} = \underset{\lambda}{\text{argmin}} \|\text{vec}(\mathbf{X}_{\text{new}}) - ((\mathbf{A}_2 \odot \mathbf{A}_1)\boldsymbol{\lambda}_1)^\top\|_2^2$; $\boldsymbol{\lambda}_2^{\text{new}} = \underset{\lambda}{\text{argmin}} \|\text{vec}(\mathbf{Y}_{\text{new}}) - ((\mathbf{V}_2 \odot \mathbf{V}_1)\boldsymbol{\lambda}_2)^\top\|_2^2$. Here, $\mathbf{A}_i$ and $\mathbf{V}_i$ ($i = 1, 2$) are the estimated factor matrices. Finally, the coupled feature $\boldsymbol{\lambda}^{\text{new}} = [\boldsymbol{\lambda}_1^{\text{new}}, \boldsymbol{\lambda}_2^{\text{new}}] = [\lambda_{11}^{\text{new}}, ..., \lambda_{1R}^{\text{new}}, \lambda_{21}^{\text{new}}, ..., \lambda_{2R}^{\text{new}}]^\top$ is used as input to the SVR and RFR models.

These extracted features along with the corresponding MDS-UPDRS form a new data for model training. First, 80% of the data are selected randomly to train the SVR and

Table 4. Comparative regression results of RCCPD versus benchmarks (Key: RMSE-root mean squared error, MAE-mean absolute error, SCC-squared correlation coefficient).

Method	RMSE	MAE	SCC
RCCPD + SVR	**0.1342** 0.01	**0.1082** 0.02	**0.9681** 0.02
TRCPD + SVR	0.1564 0.03	0.1287 0.04	0.9552 0.02
TRCPD–Voice + SVR	0.1842 0.02	0.1428 0.02	0.9401 0.03
TRCPD–Tapping + SVR	0.1754 0.03	0.1247 0.02	0.9514 0.02
CMTF + SVR	0.2042 0.04	0.1533 0.03	0.9215 0.03
CP–Voice + SVR	0.2412 0.03	0.1465 0.04	0.9328 0.03
CP–Tapping + SVR	0.2301 0.04	0.1562 0.03	0.9413 0.02
SVR	0.2358 0.05	0.1772 0.04	0.9273 0.02
Method	RMSE	MAE	SCC
RCCPD + RFR	**0.1293** 0.02	**0.0742** 0.01	**0.9837** 0.02
TRCPD + RFR	0.1482 0.03	0.1233 0.03	0.9626 0.03
TRCPD–Voice + RFR	0.1885 0.01	0.1123 0.02	0.9545 0.01
TRCPD–Tapping + RFR	0.1654 0.03	0.1043 0.01	0.9569 0.02
CMTF + RFR	0.1986 0.02	0.1180 0.02	0.9475 0.03
CP–Voice + RFR	0.2246 0.04	0.1274 0.04	0.9468 0.02
CP–Tapping + RFR	0.2120 0.03	0.1076 0.03	0.9502 0.02
RFR	0.2393 0.03	0.1357 0.03	0.9487 0.03

RFR regression models and the remaining 20% of the data are used for model testing. For the general SVR and RFR benchmarks, we directly vectorize the coupled voice-tapping matrices and use these vectors as model inputs. The value of K is set to 40 and ρ is given as 0.01. When estimating the factor matrices, hyperparameters are selected as discussed in 3.2. Specifically, the parameter α and β for RCCPD are selected as 0.2 and 0.7, respectively. The penalty parameter in TRCPD is set to be 0.5. The corresponding ranks for RCCPD, TRCPD, CMTF and CPD are selected as: 3, 3, 4, and 4 for both tensors.

We perform 50 times of the above-described methods, and the average root mean squared error (RMSE), mean absolute error (MAE), and squared correlation coefficient (SCC) with their standard deviations are reported in Table 4. As it is reported, the performance of the proposed RCCPD method is superior to benchmark methods in terms of RMSE and MAE. For example, our proposed method results in an RMSE of 0.1342 when SVR is applied and 0.1293 when RFR is used, while the CMTF approach results in an RMSE of 0.2042 (SVR) and 0.1986 (RFR). Meanwhile, the RCCPD achieves the highest SCC, which is 0.9681 (SVR). The superiority of the proposed method is due to its capability in fuzing data and extracting features that are not corrupted by outliers. These benefits of RCCPD are translated into a better regression estimation compared to the benchmarks.

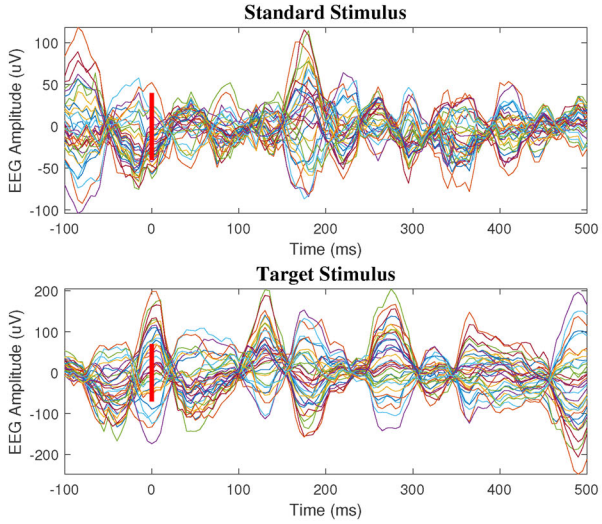
5.2. Case study II: neurosignal feature extraction and classification

Multimodal neurosignal fusion has become an important part of modern medicine to support clinical decision-making and diagnosis of various diseases. Neurosignals can span from neuroimaging to other data collection techniques, such as electroencephalography (EEG) and electrocorticography (ECoG). Because each neurosignal modality provides different but complementary information, there is an opportunity to enhance clinical decision support if the information can be fused in a way to identify patterns that are not discernable through a single image modality (Gaw et al., 2018). Previous studies have shown strong potential in the fusion of multimodal neurosignals across a variety of medical domains (Gaw et al., 2019; Hu et al., 2017; Liu et al., 2021). In particular, two of the most widely used modalities in clinical applications are functional Magnetic Resonance Imaging (fMRI) and EEG. EEG measures electrical activity in the brain through electrodes placed on the scalp and collects data at a high temporal resolution (on the magnitude of milliseconds). However, EEG alone does not have a high spatial resolution relative to neuroimaging techniques. fMRI can measure blood oxygenation levels at a higher spatial resolution (relative to EEG) but at the cost of a lower temporal resolution. Simultaneous EEG-fMRI data has been widely used to combine the best of both techniques and discern various aspects of functional networks across the brain (Bridwell & Calhoun, 2019; Dizaji & Soltanian-Zadeh, 2017; Soon et al., 2021). However, it is known that the simultaneous EEG-fMRI data is often contaminated by artifacts due to magnetic field gradients, subjects' motion, and the environment (Bullock et al., 2021).

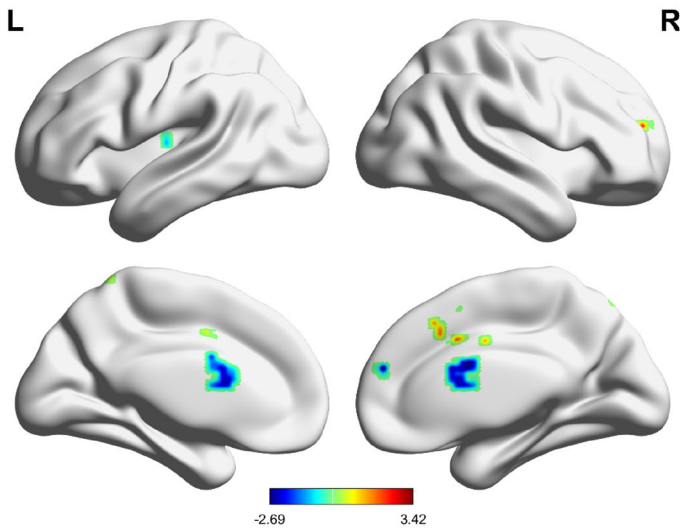
In this case study, we evaluate the performance of the proposed method in extracting robust features from simultaneously measured fMRI and EEG data for the purpose of brain activity classification. The data is obtained from a study by Walz et al. (2018) where 17 participants performed an auditory and a visual task in three runs. During each task, 375 stimuli were implemented with a 200 ms duration each and a 2-3 second inter-trial interval. A trial can be considered as a time window in which the brain receives stimuli and then gives responses. For the auditory task, a 390 Hz pure tone was considered as a standard stimulus while a broadband sound was the oddball/target stimulus. To test the proposed method, EEG and fMRI data from this auditory task are utilized to classify types of stimuli across trials. The data source is available on the OpenNeuro website (https://openneuro.org/datasets/ds000116/versions/00003).

We preprocess the EEG data with FieldTrip toolbox and the fMRI data with SPM12 (Ashburner et al., 2014) and DPABI (C.-G. Yan et al., 2016) in MATLAB R2020b. Details of processing steps are provided in Appendix B. Subject 4 is removed since its fMRI data are corrupted. Figure 3(a) demonstrates one processed EEG trial example under standard and target stimulus; Figure 3(b) shows the fMRI regions of interest (ROI) acquired in a second level analysis. In each trial, EEG data is represented as a tensor with modes of subjects × channels × time, and fMRI data is represented by a matrix with modes of subjects × voxels. Each EEG-fMRI trial is then labeled either by standard or target depending on the type of stimulus. For simplicity, we represent a standard trial by label 1 and a target trial by label 2. Eventually, each constructed coupled EEG-fMRI trial is denoted as $(\mathcal{X}_n, \mathbf{Y}_n, L_n)$, where $\mathcal{X}_n \in \mathbb{R}^{16\times 34\times 121}$, $\mathbf{Y}_n \in \mathbb{R}^{16\times 197}$ and $L_n \in \{1, 2\}$.

To test the effectiveness of the proposed RCCPD, 60 target EEG-fMRI trials and 60 standard EEG-fMRI trials are collected. For data analysis and model training, we first implement data processing (details are provided in Appendix B). Next, we apply the proposed and benchmark



(a) EEG trial samples



(b) The fMRI ROI demonstration (L=left, R=right)

Figure 3. A sample of (a) electroencephalography (EEG) and (b) functional magnetic resonance imaging (fMRI) ROI from the auditory task. For each trial, patients will either hear a standard stimulus (390 Hz pure tones) or a target stimulus (broadband sounds), while EEG and fMRI data are collected. The onset of the stimulus is represented by the red line in (a) at the 0 ms mark; (b) shows the location of regions of interest (ROI) of fMRI, i.e., activation in response to the trials, identified in the second-level analysis.

Table 5. Comparative classification results of RCCPD versus benchmarks.

Method	Accuracy	Precision	Recall	F1 score
RCCPD + SVM	**0.9102** 0.02	**0.9045** 0.03	**0.9121** 0.03	**0.9125** 0.02
TRCPD + SVM	0.8833 0.02	0.8915 0.02	0.8520 0.03	0.8810 0.02
TRCPD-EEG + SVM	0.8458 0.03	0.8460 0.04	0.8739 0.04	0.8480 0.03
TRCPD-fMRI + SVM	0.8583 0.03	0.8822 0.04	0.8385 0.04	0.8417 0.03
CMTF + SVM	0.8190 0.04	0.8365 0.04	0.8053 0.05	0.8123 0.02
CP-EEG + SVM	0.7358 0.05	0.7354 0.05	0.7492 0.04	0.7403 0.05
CP-fMRI + SVM	0.7875 0.07	0.8228 0.06	0.7717 0.08	0.7681 0.06
SVM-EEG	0.8396 0.06	0.8459 0.06	0.8452 0.07	0.8455 0.08
SVM-fMRI	0.8294 0.07	0.8850 0.07	0.8323 0.06	0.8254 0.08
Method	Accuracy	Precision	Recall	F1 score
RCCPD + RFC	**0.9208** 0.02	**0.9294** 0.02	**0.9100** 0.02	**0.9108** 0.02
TRCPD + RFC	0.8917 0.02	0.8959 0.02	0.8920 0.02	0.8885 0.02
TRCPD-EEG + RFC	0.8583 0.02	0.8478 0.03	0.8555 0.04	0.8349 0.03
TRCPD-fMRI + RFC	0.8750 0.03	0.8532 0.04	0.8568 0.03	0.8782 0.02
CMTF + RFC	0.8327 0.03	0.8459 0.04	0.8161 0.04	0.8267 0.02
CP-EEG + RFC	0.7500 0.04	0.7694 0.05	0.7672 0.06	0.7458 0.05
CP-fMRI + RFC	0.7833 0.06	0.8583 0.05	0.7287 0.09	0.7435 0.08
RFC-EEG	0.8550 0.06	0.8790 0.06	0.8693 0.07	0.8578 0.06
RFC-fMRI	0.8333 0.08	0.8667 0.05	0.8092 0.08	0.8080 0.10

methods to classify the trials. For the tensor-based methods RCCPD, TRCPD, CMTF, and CP, we first split the data into K folds and take one of the folds to estimate the factor matrices of decomposition. Next, we use the estimated factor matrices ($\mathbf{A}_i\,(i=1,2,3)$ and $\mathbf{V}_j\,(j=1,2)$) to calculate $\boldsymbol{\lambda}_i\,(i=1,2)$ for each sample in the rest of the K-1 folds. These features are the inputs of the support vector machine (SVM) and the Random Forest Classification (RFC) models. More specifically, denote a new pair of coupled tensor and matrix (EEG-fMRI) by $\mathcal{X}_{\text{new}}$ and $\mathbf{Y}_{\text{new}}$, then the feature vectors obtained by our proposed method (i.e., $\boldsymbol{\lambda}_1^{\text{new}} = [\lambda_{11}^{\text{new}},\ldots,\lambda_{1R}^{\text{new}}]^\top$ and $\boldsymbol{\lambda}_2^{\text{new}} = [\lambda_{21}^{\text{new}},\ldots,\lambda_{2R}^{\text{new}}]^\top$) can be calculated by: $\boldsymbol{\lambda}_1^{\text{new}} = argmin_\lambda \|\text{vec}(\mathcal{X}_{\text{new}}) - ((\mathbf{A}_3 \odot \mathbf{A}_2 \odot \mathbf{A}_1)\boldsymbol{\lambda}_1)^\top\|_2^2$ and $\boldsymbol{\lambda}_2^{\text{new}} = argmin_\lambda \|\text{vec}(\mathbf{Y}_{\text{new}}) - ((\mathbf{V}_2 \odot \mathbf{V}_1)\boldsymbol{\lambda}_2)^\top\|_2^2$. Finally, $\boldsymbol{\lambda}^{\text{new}} = [\boldsymbol{\lambda}_1^{\text{new}}, \boldsymbol{\lambda}_2^{\text{new}}] = [\lambda_{11}^{\text{new}},\ldots,\lambda_{1R}^{\text{new}},\lambda_{21}^{\text{new}},\ldots,\lambda_{2R}^{\text{new}}]^\top$ is used as inputs to the classification models.

These extracted features along with their corresponding label form a new data for model training. First, 80% of the samples are selected randomly for model training and the remaining 20% are used for model testing. In this case study, we also use the general SVM and RFC models as additional benchmarks. The inputs to SVM and RFC models are the vectorization of the tensor $\mathcal{X}_n \in \mathbb{R}^{16\times34\times121}$ and $\mathbf{Y}_n \in \mathbb{R}^{16\times197}$. The successful application of each method requires careful tuning of parameters. The number of folds K in the K-fold method is set to be 3. The penalty parameter ρ is set as 0.01. For the proposed RCCPD, the parameters are selected as $\alpha = 0.3$ and $\beta = 0.2$. The penalty parameter for TRCPD is set as 0.5. The rank for each tensor-based method is set to 16.

The average accuracy, precision, recall, and F1 score with their standard deviations of each method under 50 runs is reported in Table 5. It can be seen that the proposed RCCPD achieves the highest classification accuracy under both SVM and RFC models, which are 0.9102 and 0.9208, respectively. The results demonstrate that features with more discrimination ability are extracted by RCCPD, which is translated into a higher classification accuracy compared to the benchmarks. Note that the general SVM and RFC models, applied as baselines achieve better model accuracy than the CPD model. The inferior performance of CPD could be due its lack of robustness that may result in lower-quality features. This reasoning is verified given that the robust version of CPD (TRCPD) outperforms the plain SVM and RFC.

Since the hyperparameters α and β affect the performance of the proposed method, we investigate the performance sensitivity of our method with respect to α and β. We vary the parameters α and β from 0.1 to 1, respectively. Figure 4 shows the accuracy of the RFC model for various values of hyperparameters.

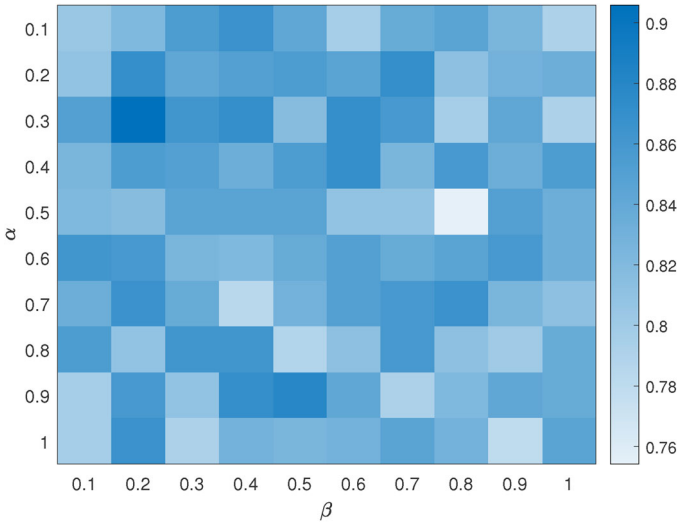


Figure 4. Average accuracy of the RFC model applied to coupled EEG-fMRI data for a range of α and β. The darker the color the higher the classification accuracy.

6. Discussion

This Section discusses a few important and practical notes related to various aspects of the paper. First, the focus of this paper is on extracting features from coupled tensors that are contaminated by outliers. Outliers are different from global noise. In tensor data, some of the observations within a tensor may not completely follow the underlying stochastic behavior (or correlation structure) of the tensor. This deviation could be due to various reasons such as sensor errors or differences in data sources. For example, in our second case study, not only could the source of outliers be due to patient motion but it could also be due to simple differences between a few subjects. Our approach isolates abnormal patterns and focuses on finding the underlying patterns that is shared among all observations. This may resemble the soft and flexible tensor decomposition approaches (Chatzichristos et al., 2022; Van Eyndhoven et al., 2017). However, while these methods relax the equality assumption of the shared modes to similarity and allow for finding common and uncommon patterns between the two tensors, they are prone to the impacts of outliers as their underlying optimization algorithm does not have an explicit mechanism for isolating outliers. Our approach can be integrated with these methods to allow explicit outlier separation. Similar to (Acar et al., 2014; 2017), our approach can also be extended to capture common and uncommon factors by imposing sparsity-inducing penalties on the feature vectors $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$. This approach may improve the results as it also alleviates the assumption that the factors of the common mode are fully shared between the two tensors. This extension is considered as a future work. Another note related to our second case study is that the preprocessing steps that we take are not necessary. Our proposed approach is generic as long as the data can be represented as coupled tensors. Our approach is an early fusion technique that extracts features from the data and uses them for model generation (as opposed to late fusion in which models are first built for each data source and then model decisions are fused (Gaw et al., 2022)).

7. Conclusions

Due to recent technological improvements in collecting multimodal high-dimensional medical data, including medical imaging and physiological signals, it has become increasingly important to develop statistical methods that can effectively fuze this data and extract informative features. Additional consideration should be used to handle outlying instances that may result from undesired interactions between data collection instruments (i.e., EEG and fMRI), or improper data collection (i.e., patients self-collecting data using telemonitoring devices). To address these challenges (high-dimensionality, multimodality, and presence of outliers), this paper proposes a robust coupled CP decomposition method (RCCPD) to fuze and reduce the dimension of different modes of data when they are contaminated with outliers. This decomposition approach results in the extraction of robust patterns and features. The extracted features can then be used for prediction (regression & classification) and decision-making purposes. A novel objective function to extract robust patterns and features is proposed. To minimize the objective function, an ADMM algorithm is developed for solving the Lagrangian of the objective function. Multiple simulation experiments and real case studies (i.e., EEG-fMRI and telemonitoring of PD patients) demonstrate the superiority of the proposed method in comparison to several benchmarks. The results of the case studies indicate the capacity of the proposed method in extracting features from multimodal medical data. For example, when the proposed method is used to extract features from the telemonitoring data, higher prediction accuracy of PD patient condition is achieved. In conclusion, the proposed model is a robust and powerful tool for feature extraction from multiple sources of potentially contaminated data. Future work may consider tensors with missing values and supervised robust feature extraction.

Disclosure statement

No potential conflict of interest was reported by the authors.

A consent and approval statement

The data used in this paper are publicly available.

Funding

This work has been supported by the University of Florida Foundation, AWD07148.

ORCID

Mostafa Reisi Gahrooei http://orcid.org/0000-0002-7633-9575

References

Acar, E., Levin-Schwartz, Y., Calhoun, V. D., & Adali, T. (2017). *ACMTF for fusion of multi-modal neuroimaging data and identification of biomarkers* [Paper presentation]. 2017 25th European Signal Processing Conference (EUSIPCO), (pp. 643–647). https://doi.org/10.23919/EUSIPCO.2017.8081286

Acar, E., Kolda, T. G., & Dunlavy, D. M. (2011). All-at-once optimization for coupled matrix and tensor factorizations. *arXiv preprint arXiv:1105.3422*

Acar, E., Papalexakis, E. E., Gürdeniz, G., Rasmussen, M. A., Lawaetz, A. J., Nilsson, M., & Bro, R. (2014). Structure-revealing data fusion. *BMC Bioinformatics*, *15*(1), 239. https://doi.org/10.1186/1471-2105-15-239

Anandkumar, A., Jain, P., Shi, Y., & Niranjan, U. N. (2016). Tensor vs. matrix methods: Robust tensor decomposition under block sparse perturbations. *Artificial Intelligence and Statistics*, *51*, 268–276.

Ashburner, J., Barnes, G., Chen, C.-C., Daunizeau, J., Flandin, G., & Friston, K. (2014). *Spm12 manual*. Wellcome Trust Centre for Neuroimaging.

Badea, L. (2008). Extracting gene expression profiles common to colon and pancreatic adenocarcinoma using simultaneous nonnegative matrix factorization. In *Biocomputing 2008* (pp. 267–278). World Scientific.

Bot, B. M., Suver, C., Neto, E. C., Kellen, M., Klein, A., Bare, C., Doerr, M., Pratap, A., Wilbanks, J., Dorsey, E. R., Friend, S. H., & Trister, A. D. (2016). The mpower study, Parkinson disease mobile data collected using researchkit. *Scientific Data*, *3*(1), 1–9. https://doi.org/10.1038/sdata.2016.11

Bridwell, D., & Calhoun, V. (2019). Fusing concurrent EEG and fMRI intrinsic networks. In *Magnetoencephalography: From signals to dynamic cortical networks*, 293–315. Springer.

Bullock, M., Jackson, G. D., & Abbott, D. F. (2021). Artifact reduction in simultaneous EEG-fMRI: A systematic review of methods and contemporary usage. *Frontiers in Neurology*, *12*, 193. https://doi.org/10.3389/fneur.2021.622719

Chaibub Neto, E., Bot, B. M., Perumal, T., Omberg, L., Guinney, J., & Kellen, M. (2016). Personalized hypothesis tests for detecting medication response in Parkinson disease patients using iphone sensor data. In *Biocomputing 2016: Proceedings of the Pacific Symposium* (pp. 273––284). .

Chang, C.-C., & Lin, C.-J. (2011). Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, *2*(3), 1–27. [Database] https://doi.org/10.1145/1961189.1961199

Chatzichristos, C., Davies, M., Escudero, J., Kofidis, E., & Theodoridis, S. (2018). *Fusion of EEG and fMRI via soft coupled tensor decompositions* [Paper presentation]. 2018 26th European Signal Processing Conference (EUSIPCO), (pp. 56–60). https://doi.org/10.23919/EUSIPCO.2018.8553077

Chatzichristos, C., Kofidis, E., Van Paesschen, W., De Lathauwer, L., Theodoridis, S., & Van Huffel, S. (2022). Early soft and flexible fusion of electroencephalography and functional magnetic resonance imaging via double coupled matrix tensor factorization for multisubject group analysis. *Human Brain Mapping*, *43*(4), 1231–1255. https://doi.org/10.1002/hbm.25717

Dizaji, A. S., & Soltanian-Zadeh, H. (2017). A change-point analysis method for single-trial study of simultaneous EEG-fMRI of auditory/visual oddball task. *Biorxiv*, 100487.

Fang, X., Paynabar, K., & Gebraeel, N. (2019). Image-based prognostics using penalized tensor regression. *Technometrics*, *61*(3), 369–384. https://doi.org/10.1080/00401706.2018.1527727

Farias, R. C., Cohen, J. E., & Comon, P. (2016). Exploring multimodal data fusion through joint decompositions with flexible couplings. *IEEE Transactions on Signal Processing*, *64*(18), 4830–4844. https://doi.org/10.1109/TSP.2016.2576425

Gaw, N., Hawkins-Daarud, A., Hu, L. S., Yoon, H., Wang, L., Xu, Y., Jackson, P. R., Singleton, K. W., Baxter, L. C., Eschbacher, J., Gonzales, A., Nespodzany, A., Smith, K., Nakaji, P., Mitchell, J. R., Wu, T., Swanson, K. R., & Li, J. (2019). Integration of machine learning and mechanistic models accurately predicts variation in cell density of glioblastoma using multiparametric MRI. *Scientific Reports*, *9*(1), 1–9. https://doi.org/10.1038/s41598-019-46296-4

Gaw, N., Schwedt, T. J., Chong, C. D., Wu, T., & Li, J. (2018). A clinical decision support system using multi-modality imaging data for disease diagnosis. *IISE Transactions on Healthcare Systems Engineering*, *8*(1), 36–46. https://doi.org/10.1080/24725579.2017.1403520

Gaw, N., Yousefi, S., & Reisi Gahrooei, M. (2022). Multimodal data fusion for systems improvement: A review. *IISE Transactions*, *54*(11), 1098–1116 .

Goetz, C. G., Tilley, B. C., Shaftman, S. R., Stebbins, G. T., Fahn, S., Martinez-Martin, P., Poewe, W., Sampaio, C., Stern, M. B., Dodel, R., Dubois, B., Holloway, R., Jankovic, J., Kulisevsky, J., Lang, A. E., Lees, A., Leurgans, S., LeWitt, P. A., Nyenhuis, D., … LaPelle, N. (2008). Movement disorder society-sponsored revision of the unified Parkinson's disease rating scale (mds-updrs): scale presentation and clinimetric testing results. *Movement Disorders*, *23*(15), 2129–2170. https://doi.org/10.1002/mds.22340

Gu, Q., Gui, H., & Han, J. (2014). Robust tensor decomposition with gross corruption. *Advances in Neural Information Processing Systems*, *27*, 1422–1430.

Harshman, R. A., & Lundy, M. E. (1994). Parafac: Parallel factor analysis. *Computational Statistics & Data Analysis*, *18*(1), 39–72. https://doi.org/10.1016/0167-9473(94)90132-5

He, H., Henderson, J., & Ho, J. C. (2019). *Distributed tensor decomposition for large scale health analytics* [Paper presentation]. The World Wide Web Conference, (pp. 659–669).

Hotelling, H. (1992). Relations between two sets of variates. In *Breakthroughs in statistics* (pp. 162–190). Springer.

Hu, L. S., Ning, S., Eschbacher, J. M., Baxter, L. C., Gaw, N., Ranjbar, S., Plasencia, J., Dueck, A. C., Peng, S., Smith, K. A., Nakaji, P., Karis, J. P., Quarles, C. C., Wu, T., Loftus, J. C., Jenkins, R. B., Sicotte, H., Kollmeyer, T. M., O'Neill, B. P., … Mitchell, J. R. (2017). Radiogenomics to characterize regional genetic heterogeneity in glioblastoma. *Neuro-oncology*, *19*(1), 128–137. https://doi.org/10.1093/neuonc/now135

Jonmohamadi, Y., Muthukumaraswamy, S., Chen, J., Roberts, J., Crawford, R., & Pandey, A. (2020). Extraction of common task features in EEG-fMRI data using coupled tensor-tensor decomposition. *Brain Topography*, *33*(5), 636–650.

Khanzadeh, M., Tian, W., Yadollahi, A., Doude, H. R., Tschopp, M. A., & Bian, L. (2018). Dual process monitoring of metal-based additive manufacturing using tensor decomposition of thermal image streams. *Additive Manufacturing*, *23*, 443–456.

Kolda, T. G., & Bader, B. W. (2009). Tensor decompositions and applications. *SIAM Review*, *51*(3), 455–500. https://doi.org/10.1137/07070111X

Liu, X., Chen, K., Weidman, D., Wu, T., Lure, F., & Li, J., for the Alzheimer's Disease Neuroimaging Initiative. (2021). A novel transfer learning model for predictive analytics using incomplete multimodality data. *IISE Transactions*, *53*(9), 1010–1022. https://doi.org/10.1080/24725854.2020.1798569

Miao, H., Wang, A., Li, B., & Shi, J. (2022). Structural tensor-on-tensor regression with interaction effects and its application to a hot rolling process. *Journal of Quality Technology*, *54*(5), 547–560.

Mosayebi, R., & Hossein-Zadeh, G.-A. (2020). Correlated coupled matrix tensor factorization method for simultaneous EEG-fMRI data fusion. *Biomedical Signal Processing and Control*, *62*, 102071. https://doi.org/10.1016/j.bspc.2020.102071

Naskovska, K., & Haardt, M. (2016). *Extension of the semi-algebraic framework for approximate CP decompositions via simultaneous matrix diagonalization to the efficient calculation of coupled CP decompositions* [Paper presentation]. 2016 50th Asilomar Conference on Signals, Systems and Computers, (pp. 1728–1732). https://doi.org/10.1109/ACSSC.2016.7869678

Naskovska, K., Korobkov, A. A., Haardt, M., & Haueisen, J. (2017). *Analysis of the photic driving effect via joint EEG and MEG data processing based on the coupled CP decomposition* [Paper presentation]. 2017 25th European Signal Processing Conference (EUSIPCO), (pp. 1285–1289). https://doi.org/10.23919/EUSIPCO.2017.8081415

Parkinsons News Today. (2020). *Parkinson's disease statistics*. Retrieved February 14, 2022, from https://parkinsonsnewstoday.com/parkinsons-disease-statistics/.

Pew Research Center. (2021). *Mobile fact sheet*. Retrieved February 14, 2022, from https://www.pewresearch.org/internet/fact-sheet/mobile/.

Roemer, F., & Haardt, M. (2013). A semi-algebraic framework for approximate CP decompositions via simultaneous matrix diagonalizations (SECSI). *Signal Processing*, *93*(9), 2722–2738. https://doi.org/10.1016/j.sigpro.2013.02.016

Singh, A. P., & Gordon, G. J. (2008). *Relational learning via collective matrix factorization* [Paper presentation]. Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (pp. 650–658). https://doi.org/10.1145/1401890.1401969

Smilde, A. K., Westerhuis, J. A., & Boque, R. (2000). Multiway multiblock component and covariates regression models. *Journal of Chemometrics*, *14*(3), 301–331. https://doi.org/10.1002/1099-128X(200005/06)14:3<301::AID-CEM594>3.0.CO;2-H

Soon, C. S., Vinogradova, K., Ong, J. L., Calhoun, V. D., Liu, T., Zhou, J. H., Ng, K. K., & Chee, M. W. L. (2021). Respiratory, cardiac, EEG, BOLD signals and functional connectivity over multiple microsleep episodes. *NeuroImage*, *237*, 118129. https://doi.org/10.1016/j.neuroimage.2021.118129

Sørensen, M., & De Lathauwer, L. (2013). *Coupled tensor decompositions for applications in array signal processing* [Paper presentation]. 2013 5th IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), (pp. 228–231).

Tsanas, A. (2012). *Accurate telemonitoring of Parkinson's disease symptom severity using nonlinear speech signal processing and statistical machine learning* [Unpublished doctoral dissertation]. Oxford University.

Tsanas, A., Little, M. A., McSharry, P. E., & Ramig, L. O. (2011). Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity. *Journal of the Royal Society, Interface*, *8*(59), 842–855.

Van Eyndhoven, S., Hunyadi, B., De Lathauwer, L., & Van Huffel, S. (2017). *Flexible fusion of electroencephalography and functional magnetic resonance imaging: Revealing neural-hemodynamic coupling through structured matrix-tensor factorization* [Paper presentation]. 2017 25th European Signal Processing Conference (EUSIPCO), (pp. 26–30). https://doi.org/10.23919/EUSIPCO.2017.8081162

Vervliet, N., Debals, O., & De Lathauwer, L. (2016). Tensorlab 3.0—numerical optimization strategies for large-scale constrained and coupled matrix/tensor factorization. *2016 50th Asilomar Conference on Signals, Systems and Computers*, (pp. 1733–1738). https://doi.org/10.1109/ACSSC.2016.7869679.

Walz, J. M., Goldman, R. I., Muraskin, J., Conroy, B., Brown, T. R., & Sajda, P. (2018). *Auditory and Visual Oddball EEG-fMRI*. OpenNeuro.

Xue, N., Papamakarios, G., Bahri, M., Panagakis, Y., & Zafeiriou, S. (2017). *Robust low-rank tensor modelling using tucker and CP decomposition* [Paper presentation]. 2017 25th European Signal Processing Conference (EUSIPCO), (pp. 1185–1189). https://doi.org/10.23919/EUSIPCO.2017.8081395

Yan, C.-G., Wang, X.-D., Zuo, X.-N., & Zang, Y.-F. (2016). Dpabi: data processing & analysis for (resting-state) brain imaging. *Neuroinformatics*, *14*(3), 339–351. https://doi.org/10.1007/s12021-016-9299-4

Yan, H., Paynabar, K., & Pacella, M. (2019). Structured point cloud data analysis via regularized tensor regression for process modeling and optimization. *Technometrics*, *61*(3), 385–395. https://doi.org/10.1080/00401706.2018.1529628

Yoon, H., & Gaw, N. (2021). A novel multi-task linear mixed model for smartphone-based telemonitoring. *Expert Systems with Applications*, *164*, 113809. https://doi.org/10.1016/j.eswa.2020.113809

Zhao, Q., Caiafa, C. F., Mandic, D. P., Chao, Z. C., Nagasaka, Y., Fujii, N., Zhang, L., & Cichocki, A. (2013). Higher order partial least squares (HOPLS): a generalized multilinear regression method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(7), 1660–1673. https://doi.org/10.1109/TPAMI.2012.254

Appendix A. Derivations of Equation 13 and Equation 16

In this appendix, derivations of Equation 13 and Equation 16 are provided.

(i) The update of variable $\mathcal{F}$ is derived by minimizing the following function:

$$\begin{aligned}\mathcal{F}^{t+1} &= \underset{\mathcal{F}}{argmin}\, L_\rho(\mathcal{F}; \mathcal{S}_1^t, \mathcal{D}_1^t)\\ &= \underset{\mathcal{F}}{argmin}\left\{\alpha\|\mathcal{F}\|_1 + \langle \mathcal{D}_1^t, \mathcal{F}-\mathcal{S}_1^t\rangle + \frac{\rho}{2}\|\mathcal{F}-\mathcal{S}_1^t\|_F^2\right\}\\ &= \underset{\mathcal{F}}{argmin}\left\{\alpha\rho^{-1}\|\mathcal{F}\|_1 + \frac{1}{2}\|\mathcal{F}-\mathcal{S}_1^t+\rho^{-1}\mathcal{D}_1^t\|_F^2\right\}\end{aligned} \tag{A1}$$

The optimal condition for the above formula is: $0 \in \alpha\rho^{-1}\partial\|\mathcal{F}\|_1 + \mathcal{F} - (\mathcal{S}_1^t - \rho^{-1}\mathcal{D}_1^t)$.

For the case $\mathcal{F} \neq 0$, we have $\partial\|\mathcal{F}\|_1 = \text{sign}(\mathcal{F})$ and the optimum $\mathcal{F}^*$ is: $\mathcal{F}^* = \mathcal{S}_1^t - \rho^{-1}\mathcal{D}_1^t - \alpha\rho^{-1}\text{sign}(\mathcal{F}^*)$. For the case $\mathcal{F} < 0$, $\mathcal{S}_1^t - \rho^{-1}\mathcal{D}_1^t < -\alpha\rho^{-1}$ and equivalently for $\mathcal{F} > 0$, we have $\mathcal{S}_1^t - \rho^{-1}\mathcal{D}_1^t > \alpha\rho^{-1}$. Thus, $|\mathcal{S}_1^t - \rho^{-1}\mathcal{D}_1^t| > \alpha\rho^{-1}, \text{sign}(\mathcal{F}^*) = \text{sign}(\mathcal{S}_1^t - \rho^{-1}\mathcal{D}_1^t)$, then we have $\mathcal{F}^* = \mathcal{S}_1^t - \rho^{-1}\mathcal{D}_1^t) - \alpha\rho^{-1}\text{sign}(\mathcal{S}_1^t - \rho^{-1}\mathcal{D}_1^t)$. In the case where $\mathcal{F} = 0$, the optimal condition is $|\mathcal{S}_1^t - \rho^{-1}\mathcal{D}_1^t)| \leq \alpha\rho^{-1}$. Putting the above cases together, we have the update of $\mathcal{F}$ in a soft-threshold form as follows:

$$\begin{aligned}&\text{Prox}_{\alpha\rho^{-1}}(\mathcal{S}_1^t - \rho^{-1}\mathcal{D}_1^t)\\ &= \begin{cases}\mathcal{S}_1 - \rho^{-1}\mathcal{D}_1^t - \alpha\rho^{-1}, & \text{if} \quad \mathcal{S}_1 - \rho^{-1}\mathcal{D}_1^t > \alpha\rho^{-1}\\ 0, & \text{if} \quad |\mathcal{S}_1 - \rho^{-1}\mathcal{D}_1^t| \leq \alpha\rho^{-1}\\ \mathcal{S}_1 - \rho^{-1}\mathcal{D}_1^t + \alpha\rho^{-1}, & \text{if} \quad \mathcal{S}_1 - \rho^{-1}\mathcal{D}_1^t < -\alpha\rho^{-1}\end{cases}\end{aligned} \tag{A2}$$

Finally, $\mathcal{F}^{t+1} = \text{Prox}_{\alpha\rho^{-1}}(\mathcal{S}_1^t - \rho^{-1}\mathcal{D}_1^t) = \text{sign}(\mathcal{S}_1^t - \rho^{-1}\mathcal{D}_1^t) * (|\mathcal{S}_1^t - \rho^{-1}\mathcal{D}_1^t| - \alpha\rho^{-1}, 0)_+$.

The update of variable $\mathcal{M}$ is derived by minimizing the function below:

$$\begin{aligned}\mathcal{M}^{(t+1)} &= \underset{\mathcal{M}}{argmin}\, L_\rho(\mathcal{M}; \mathcal{S}_2^t, \mathcal{D}_2^t)\\ &= \underset{\mathcal{M}}{argmin}\left\{\beta\|\mathcal{M}\|_1 + \langle \mathcal{D}_2^t, \mathcal{M}-\mathcal{S}_2^t\rangle + \frac{\rho}{2}\|\mathcal{M}-\mathcal{S}_2^t\|_F^2\right\}\\ &= \underset{\mathcal{M}}{argmin}\left\{\beta\rho^{-1}\|\mathcal{M}\|_1 + \frac{1}{2}\|\mathcal{M}-\mathcal{S}_2^t+\rho^{-1}\mathcal{D}_2^t\|_F^2\right\}\end{aligned} \tag{A3}$$

Similarly, the optimal condition for equation A3 is: $0 \in \beta\rho^{-1}\partial\|\mathcal{M}\|_1 + \mathcal{M} - (\mathcal{S}_2^t - \rho^{-1}\mathcal{D}_2^t)$.

When $\mathcal{M} \neq 0$, we have $\partial\|\mathcal{M}\|_1 = \text{sign}(\mathcal{M})$ and the optimum $\mathcal{M}^*$ is obtained as: $\mathcal{M}^* = \mathcal{S}_2^t - \rho^{-1}\mathcal{D}_2^t - \beta\rho^{-1}\text{sign}(\mathcal{M}^*)$. For $\mathcal{M} < 0$, $\mathcal{S}_2^t - \rho^{-1}\mathcal{D}_2^t < -\beta\rho^{-1}$ and equivalently if $\mathcal{M} > 0$ we have $\mathcal{S}_2^t - \rho^{-1}\mathcal{D}_2^t > \beta\rho^{-1}$. Thus, $|\mathcal{S}_2^t - \rho^{-1}\mathcal{D}_2^t| > \beta\rho^{-1}\text{sign}(\mathcal{M}^*) = \text{sign}(\mathcal{S}_2^t - \rho^{-1}\mathcal{D}_2^t)$, then we have $\mathcal{M}^* = \mathcal{S}_2^t - \rho^{-1}\mathcal{D}_2^t - \beta\rho^{-1}\text{sign}(\mathcal{S}_2^t - \rho^{-1}\mathcal{D}_2^t)$. In the case where $\mathcal{M} = 0$, the optimal condition is $|\mathcal{S}_2^t - \rho^{-1}\mathcal{D}_2^t| \leq \beta\rho^{-1}$. Putting the above cases together, we have:

$$\begin{aligned}&\text{Prox}_{\beta\rho^{-1}}(\mathcal{S}_2^t - \rho^{-1}\mathcal{D}_2^t)\\ &= \begin{cases}\mathcal{S}_2 - \rho^{-1}\mathcal{D}_2^t - \beta\rho^{-1}, & \text{if} \quad \mathcal{S}_2 - \rho^{-1}\mathcal{D}_2^t > \beta\rho^{-1}\\ 0, & \text{if} \quad |\mathcal{S}_2 - \rho^{-1}\mathcal{D}_2^t| \leq \beta\rho^{-1}\\ \mathcal{S}_2 - \rho^{-1}\mathcal{D}_2^t + \beta\rho^{-1}, & \text{if} \quad \mathcal{S}_2 - \rho^{-1}\mathcal{D}_2^t < -\beta\rho^{-1}\end{cases}\end{aligned} \tag{A4}$$

Finally, $\mathcal{M}^{t+1} = \text{Prox}_{\beta\rho^{-1}}(\mathcal{S}_2^t - \rho^{-1}\mathcal{D}_2^t) = \text{sign}(\mathcal{S}_2^t - \rho^{-1}\mathcal{D}_2^t) * (|\mathcal{S}_2^t - \rho^{-1}\mathcal{D}_2^t| - \beta\rho^{-1}, 0)_+$.

Appendix B. EEG-fMRI data processing for Section 5.2

Details of EEG-fMRI data preprocessing and fMRI data extraction are demonstrated in this appendix. For most processing procedures, we referred the work from Ashburner et al. (2014).

B.1. EEG data

The EEG data is recorded by a custom-built MR-compatible EEG system, using 49 bipolar electrode pairs. In this article, the re-referenced EEG data with 34 channels provided by Walz et al. (2018) is used. This version of EEG data is originally sampled at 1,000 Hz, which is then down-sampled to 200 Hz in our work. This down-sampling step is not necessary and is merely performed to smoothen the data and reduce the noise and the dimensions of the corresponding tensors. A function named $'ft_{preproc}resample'$ in Field Trip toolbox is used for data down-sampling. Next, the band-pass filter with cutoff frequencies of 1 Hz and 100 Hz is used to remove direct current drift and high frequency artifacts not related to neuronal oscillations. Finally, we split EEG trials with a time window which begins 100 ms before the stimulus onset and ends 500 ms after the stimulus onset. We did not perform further preprocessing steps to remove the ballistocardiac artifact or other noise.

B.2. fMRI data

The fMRI data is collected by a 3 T Philips Achieva MRI scanner, which results in 170 volumes in each run. Its repetition time (TR) is 2 s, the number of slices is 34 with no slice gap and the resolution is $3 \times 3 \times 4$ mm. We preprocess the fMRI images in SPM12 toolbox by performing the following steps: slice timing, realignment, co-registration, segmentation, normalization, and smoothing.

(i) Slice timing. Due to the nature of the fMRI acquisition principles, slices cannot be simultaneously obtained, in other words, they are temporally misaligned from each other. To correct differences in acquisition time, slice timing preprocessing is necessary. This procedure generates a file with prefix $'a'$.

(ii) Realignment. This step is used to align the time-series of 3D BOLD volumes from the same subject to remove the influence caused by head motions. In the auditory task, total 510 fMRI scans are acquired from each subject in three runs. These scans are realigned to the average of these 510 scans and the mean scan is generated in this step for co-registration.

(iii) Co-registration. Since all of the fMRI scans have been aligned to the mean scan in the realignment step, the T1 weighted anatomical scan needs to be transformed to match their orientation as well. In this step, the mean fMRI scan is stationary, and the T1 anatomical scan is moved to match it, meanwhile, a re-sliced T1 weighted scan is created.

(iv) Segmentation. In this procedure, a deformation transformation is estimated to map data into MNI 152 template space. A forward deformation field is generated in this step.

(v) Normalization. This procedure consists of two components: estimation and writing. In the estimation part, a deformation is estimated through deforming the MNI template to match each single fMRI scan. The voxel size in this step is $3 \times 3 \times 4$ mm. The writing part applies previously estimated warps to series of images and then generate a file with prefix $'w'$.

(vi) Smoothing. After normalization, all fMRI images are then smoothed to suppress noise via a Gaussian kernel with $6 \times 6 \times 6$ mm full width at half maximum (FWHM). This step creates a smoothed image file with prefix $'s'$. Finally, the smoothed image data is normalized by a Z-score method in this work.

Once the preprocessing steps are completed, the first level analysis and the second level analysis are implemented successively in SPM12 toolbox and DPABI toolbox, separately. The ROIs obtained after the second level analysis are shown in Figure 3. Finally, we use the extracted ROI voxel data for model estimation.