Journal of Data Science 20 (2), 228–252
April 2022

DOI: 10.6339/22-JDS1048
Statistical Data Science

An Effective Tensor Regression with Latent Sparse Regularization

Ko-shin Chen[1], Tingyang Xu[2], Guannan Liang[1], Qianqian Tong[1], Minghu Song[3], and Jinbo Bi[1,*]

[1] *Department of Computer Science and Engineering, University of Connecticut, United States*
[2] *Tencent AI Lab, China*
[3] *Department of Biomedical Engineering, University of Connecticut, United States*

Abstract

As data acquisition technologies advance, longitudinal analysis is facing challenges of exploring complex feature patterns from high-dimensional data and modeling potential temporally lagged effects of features on a response. We propose a tensor-based model to analyze multidimensional data. It simultaneously discovers patterns in features and reveals whether features observed at past time points have impact on current outcomes. The model coefficient, a k-mode tensor, is decomposed into a summation of k tensors of the same dimension. We introduce a so-called latent F-1 norm that can be applied to the coefficient tensor to performed structured selection of features. Specifically, features will be selected along each mode of the tensor. The proposed model takes into account within-subject correlations by employing a tensor-based quadratic inference function. An asymptotic analysis shows that our model can identify true support when the sample size approaches to infinity. To solve the corresponding optimization problem, we develop a linearized block coordinate descent algorithm and prove its convergence for a fixed sample size. Computational results on synthetic datasets and real-life fMRI and EEG datasets demonstrate the superior performance of the proposed approach over existing techniques.

Keywords *longitudinal data; quadratic inference function; tensors*

1 Introduction

Nowadays, the advances in data acquisition technologies have collected ultra high dimensional data with complex structure in many disciplines and industrial societies (Donoho, 2000; Liu et al., 2016). Such datasets contain tensor data entries where each observed example is a high dimensional tensor. In a neuroscience study (Cong et al., 2015), researchers examine different electroencephalogram (EEG) recordings to distinguish patient trials (recordings) with successful working memory from those without. Particularly, EEG data is high-dimensional and complex, based on a time series of events sampled with high temporal resolution (i.e, millisecond level) and distributed spatially across multiple scalp locations (e.g., montages of 32 to 256 channels) (Figure 1(top-left and top-right)). A single EEG feature, such as the EEG signal amplitude in the α frequency band, can be extracted at different brain information processing stages and from various scalp locations (or EEG electrodes) (see Figure 1(bottom)). This single feature

[*] Corresponding author. A short version of this paper was presented at the Association for the Advancement of Artificial Intelligence in 2018. Email: jinbo.bi@uconn.com.

Received December 31, 2021; Accepted April 10, 2022

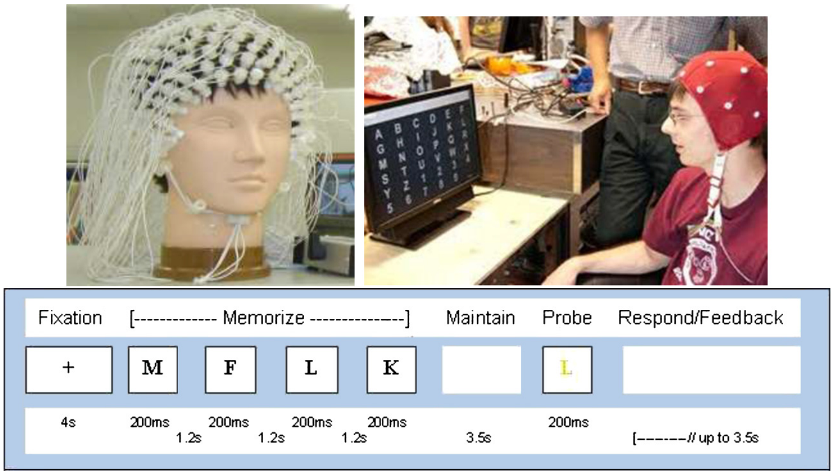


Figure 1: Illustration of EEG Brain Computer Interface (BCI) apparatus and working memory test: (top-left) EEG recording montage; (top-right) a BCI program called P300 speller; (bottom) a sample trial of Sternberg experiment depicting stages of information processing and time courses as extracted for EEG analysis based on memory span of four letter.

already forms a matrix with one dimension along the temporal line and the other along the spatial line. Then, when multiple EEG features are extracted from an EEG recording, these features altogether form a tensor. In other words, a single recording is represented by a tensor, or in other words a 3D array. Repeated measurement of functional magnetic resonance imaging (fMRI) can create tensors in even higher dimensions because an fMRI image itself is a high dimensional volume (Pereira et al., 2009). Classic statistical tools can hardly handle this kind of data efficiently. Standard machine learning methods also need to flatten an example represented by a tensor into a long vector before building a regression or classification model, thus losing the complex proximity structure.

Conventional tensor techniques comprise a set of tensor decomposition methods where a data tensor is decomposed into low-rank tensors (Hitchcock, 1927; Tucker, 1966). These techniques have proven to be useful in data mining (Acar and Yener, 2009), signal processing (De Lathauwer and Vandewalle, 2004) and computer vision (Vasilescu and Terzopoulos, 2002). Recently, a set of methods has focused on the convex optimization formulations for tensor decomposition. These methods added penalization of Schatten 1-norm on a sequence of unfolded matrices from the tensor to the objective function. Or they decomposed a tensor into a summation of several low rank tensors, leading to the so called *latent approach* (Tomioka et al., 2010).

In this paper we introduce a tensor-based quadratic inference function (TensorQIF) machine learning model that can be used to analyze longitudinal data and select features efficiently. Longitudinal data consists of repeated sample observations during a given time period. They appear in a variety of areas, from finance (Arnold et al., 2007; Sela and Simonoff, 2012) to scientific research such as health-care and medicine (Bi et al., 2013; Stappenbeck and Fromme, 2010).

A notable feature of longitudinal data is repeated-measurement within each subject. Thus observed responses are generally dependent and longitudinal correlation among different outcomes must be considered to obtain efficient predictions. There are several extended generalized linear models that can be applied to time-dependent data under different assumptions. P. Diggle et al. have provided a comprehensive overview of various models (Diggle et al., 2002). For fitting marginal model, generalized estimating equation — GEE (Liang and Zeger, 1986a) and quadratic inference function — QIF (Qu and Li, 2006) are common statistical approaches. They are generally more efficient than those of classic regression analysis that assumes working independently (ind).

In a GEE model, the correlation structure of outcomes is presumed and the so-called 'working' correlation matrix, R, is specified. However, in practice, the true correlation is often unknown. The GEE model with misspecified working correlation matrix will no longer result optimal estimation of the coefficients (Crowder, 1995). In addition, the inverse of the matrix R is essential that may cause poor estimation when R has high dimensions (Qu and Lindsay, 2003). To overcome these disadvantages, Qu et al. (2000) suggested the QIF method for which R^{-1} is approximated by a linear combination of several basis matrices. This method ensures that the estimator always exists and does not require any estimation for nuisance parameters associated with correlations. On the feature selection criteria, penalized GEE (Fu, 2003) and penalized QIF (Bai et al., 2009) are proposed.

In this work, we study the lagged effect of covariates on outcomes. It is necessary and insightful to model simultaneously the correlation among the outcomes and the lagged effects of covariates, as studied in Granger causality (Granger, 1980). For example, Shen et al. (2014) pointed out evidences of brain diseases may appear in the fMRI of an early diagnosis before clear symptoms are identified. Recent graphical Granger models (Arnold et al., 2007; Lozano et al., 2009) ignore the temporal correlations. The work in Xu et al. (2015) has modeled such correlation through the GEE method. But their model only applies to datasets with one spatial dimension. Our goal is to develop a new penalized QIF method in the tensor setting to make temporal prediction. Nowadays, tensor regressions have shown to be powerful in learning complex feature structures from multidimensional data. Many tensor techniques have been developed and applied to a broad range of applications (Hoff, 2015; Zhou et al., 2013). However when focusing on feature selections (e.g., sparse tensor decomposition), most of existing methods either assume i.i.d. samples, or assume correlated samples but do not model temporal additive effects.

We propose a new learning formulation that constructs a tensor-based predictive model as a function of covariates, not only from the current observation but also from multiple previous consecutive observations. Simultaneously the model determines the temporal contingency and the most influential features along each dimension of the tensor data. Given a data sample characterized by a tensor, the coefficients in our additive model also form a K-way tensor. To select features, we decompose the K-way coefficient tensor into a summation of K sparse K-way tensors as shown in Figure 2. These tensors each present sparsity along one direction and impose different block-wise least absolute shrinkage and selection operators (LASSO) to the components. We use linearized block coordinate descent algorithm via a proximal map (Xu and Yin, 2017) to efficiently solve the optimization problem. This approach then leads to K subproblems that share the same structure. We validate the effectiveness of the proposed method in simulations and in the analysis of real-life fMRI and EEG datasets.

2 Method

This section is dedicated to deriving the formulation of the proposed QIF where we first review the different generalized linear models.

Notations. We represent a K-way tensor as $\mathcal{A} \in \mathbb{R}^{d1 \times d2 \times \ldots d_K}$ which contains $N = \prod_{k=1}^{K} d_k$ elements. The inner product of two tensors $\mathcal{A}$ and $\mathcal{B}$ is defined by $\langle \mathcal{A}, \mathcal{B} \rangle = \text{vect}(\mathcal{A})^\top \text{vect}(\mathcal{B})$,

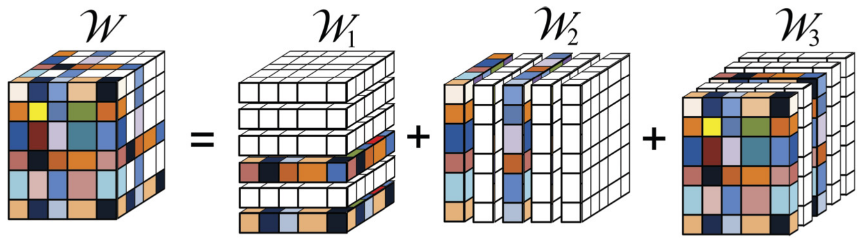


Figure 2: A 3-way tensor is decomposed into a summation of three 3-way tensors so that each part is sparse along a particular direction.

where vect$(\cdot)$ denotes the column-major vectorization of a tensor. The Frobenius norm of a tensor $\mathcal{A}$ is defined by $\|A\|_F = \sqrt{\langle \mathcal{A}, \mathcal{A} \rangle}$. The j-th sub-tensor of a tensor $\mathcal{A}$ along the mode-k can be obtained by fixing the k-th index as j, i.e. $\mathcal{A}_{(k)}^{(j)} = \mathcal{A}(i_1, i_2, \dots, i_k \equiv j, i_{k+1}, \dots i_K)$. Note that $\mathcal{A}_{(k)}^{(j)}$ is a $(K-1)$-way tensor. The mode-k fiber of $\mathcal{A}$ is a d_k dimensional vector which is obtained by fixing all index of $\mathcal{A}$ except the k-th one. The mode-k unfolding of $\mathcal{A}$ is a matrix $\mathbf{A}_{(k)} \in \mathbb{R}^{d_k \times N/d_k}$ formed by concatenating all the N/d_k mode-k fibers along its columns. The operator $[\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_m]$ creates a $(K+1)$-way tensor by concatenating m numbers of K-way tensors $\mathcal{A}_1$, $\mathcal{A}_2$, ..., $\mathcal{A}_m$ of the same dimension.

2.1 Generalized Linear Models with a Tensor

We first introduce a basic tensor formulation in which the objective function is written down into two parts: a loss function l and a regularizer. Let $(\mathcal{X}_i, y_i)_{1 \leqslant i \leqslant m}$ be a data set, where $\mathcal{X}_i \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_K}$ is a covariate tensor and $y_i \in \mathbb{R}$ (resp. $\{\pm 1\}$) for regression (resp. classification) is the corresponding outcome. We consider a linear model below:

$$\min_{\mathcal{W}} \sum_{i=1}^{m} l(y_i, \langle \mathcal{X}_i, \mathcal{W} \rangle) + \lambda \|\mathcal{W}\|_{(\cdot)}, \tag{1}$$

where $\lambda \geqslant 0$ is the regularization parameter, and $\|\cdot\|_{(\cdot)}$ is a certain tensor norm. Elements in the tensor $\mathcal{W}$ are the model coefficients to be fitted. In the study of low-rank tensor decompositions, overlapped/latent tensor trace norm (Wimalawarne et al., 2016) or Schatten norm (Tomioka and Suzuki, 2013) are widely applied in (1). Although these latent tensor norms facilitate the search for a low-rank tensor solution, they cannot enforce sparsity and thus unable to select the most relevant ones among features.

In this paper, we focus on sparsity and feature selection by imposing a regularization condition that forces to zero out an entire slice of the coefficient tensor. In other words, our model selects nonzero slices in each direction of the tensor $\mathcal{W}$. We hence introduce the latent $\mathrm{L}_{\mathrm{F},1}$ norm defined by

$$\|\mathcal{W}\|_{L_{F,1}} := \inf_{\sum_{k=1}^{K} \mathcal{W}_k = \mathcal{W}} \sum_{k=1}^{K} \left(\lambda_k \sum_{j=1}^{d_k} \|(\mathcal{W}_k)_{(k)}^{(j)}\|_F \right) \tag{2}$$

where λ_ks' are nonnegative constants. One can easily verify that Eq. (2) satisfies all required norm properties.

There are various of settings for the loss function l depending on the specific learning tasks. When the dataset is assumed to be i.i.d, the squared loss $l(y_i, \langle \mathcal{X}_i, \mathcal{W} \rangle) = (y_i - \langle \mathcal{X}_i, \mathcal{W} \rangle)^2$; for regression or the logistic loss $l(y_i, \langle \mathcal{X}_i, \mathcal{W} \rangle) = \log(1 + \exp(-y_i \langle \mathcal{X}_i, \mathcal{W} \rangle))$. For classification are two simple models usually applied. A more general family - generalized linear model (GLM) - has been used according to an exponential distribution assumption on the dependent variable. This family includes both the squared loss and logistic loss. To deal with correlated samples, GLM has been further extended from point estimation to variance estimation, which leads to more complicated formula, such as GEE or QIF. Between these two, QIF is more effective as discussed early on. In this paper, we will use the QIF setting to analyze additive effects in longitudinal datasets. The complete formula of l in our model will be given in the next section.

2.2 The Proposed QIF Formulation

Let $\mathcal{X}_t^{(i)} \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_{K-1}}$ be a $(K-1)$-way tensor which represents the covariate tensor measured for the subject i at time t. We denote $y_t^{(i)}$ the outcome of the subject i at time t. We assume that $y_t^{(i)}$ depends not only on the current record $\mathcal{X}_t^{(i)}$ but also on the previous τ records: $\mathcal{X}_{t-1}^{(i)}$, $\mathcal{X}_{t-2}^{(i)}$, ..., $\mathcal{X}_{t-\tau}^{(i)}$. Hence we may view a sample at a particular time t as a pair $(\mathcal{X}_{(i;t)}, y_t^{(i)})$, where $\mathcal{X}_{(i;t)}$ is a K-way tensor concatenating all considered records: $\mathcal{X}_{(i;t)} := [\mathcal{X}_t^{(i)}, \mathcal{X}_{t-1}^{(i)}, \mathcal{X}_{t-2}^{(i)}, \ldots, \mathcal{X}_{t-\tau}^{(i)}]$. Suppose there are T total times of measurement for each subject i. In order to have enough previous observations, the index t of $\mathcal{X}_{(i;t)}$ should start from $\tau + 1$ and there are $n := T - \tau$ training examples for each subject. In the graphical Granger model, the relation between $\mathcal{X}_{(i;t)}$ and $y_t^{(i)}$ is given by

$$y_t^{(i)} = \langle \mathcal{X}_{(i;t)}, \mathcal{W} \rangle \tag{3}$$

for some tensor coefficient $\mathcal{W} \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_{K-1} \times d_K}$, where $d_K = \tau$. We denote $N := \prod_{k=1}^{K} d_k$ the number of elements in $\mathcal{W}$. However, training examples in Eq. (3) are assumed to be i.i.d., which does not fit the intrinsic property of our dataset. In our case, the consecutive examples share overlapping records (e.g. $\mathcal{X}_{(i;t)}$ and $\mathcal{X}_{(i;t+1)}$ share $\tau - 1$ records: $\mathcal{X}_t^{(i)}$, $\mathcal{X}_{t-1}^{(i)}$, ..., $\mathcal{X}_{t-\tau+1}^{(i)}$) and outcomes $y_t^{(i)}$, $y_t^{(i-1)}$ are correlated. Hence in this paper, we adapt QIF method which together with GEE are members of GLM.

There are two essential ingredients in GLM: a link function and a variance function. The link function describes the relation between a linear predictor η and the mean (expectation) of an outcome y. The variance function tells how the variance of an outcome y depends on its mean. In our formulation, these can be expressed by

$$\mu_t^{(i)} := \mathbb{E}[y_t^{(i)}] = h^{-1}(\eta_t^{(i)}), \quad \mathrm{var}(y_t^{(i)}) = V(\mu_t^{(i)}), \tag{4}$$

where h is a link function determined according to a presumed distribution on y_t from the exponential family, V is a variance function, and

$$\eta_t^{(i)} = \langle \mathcal{X}_{(i;t)}, \mathcal{W} \rangle \tag{5}$$

is the linear predictor. Let $\mathbf{y}^{(i)} := (y_{\tau+1}^{(i)}, \ldots, y_{\tau+n}^{(i)})^T$ be an n-dimensional column vector. In GEE models, the covariance matrix $\mathbf{\Sigma}^{(i)}$ for $\boldsymbol{y}^{(i)}$ is modeled by

$$\mathbf{\Sigma}^{(i)} := \left(\mathbf{A}^{(i)}\right)^{1/2} \mathbf{R}(\alpha) \left(\mathbf{A}^{(i)}\right)^{1/2}, \tag{6}$$

where $\mathbf{R}(\alpha)$ is the 'working' correlation matrix, and $\mathbf{A}^{(i)}$ is an $n \times n$ diagonal matrix with $V(\mu_{\tau+j}^{(i)})$ as the j-th diagonal element. The matrix $\mathbf{\Sigma}^{(i)}$ will be equal to $\mathrm{cov}(\mathbf{y}^{(i)})$ if $\mathbf{R}(\alpha)$ is the true correlation structure for $\mathbf{y}^{(i)}$ (Liang and Zeger, 1986a). The model coefficients are then obtained by solving the score equation from the quasi-likelihood analysis. In our setting, it turns out to be

$$\sum_{i=1}^{m} \left(\mathbf{D}^{(i)}\right)^{T} \left(\mathbf{A}^{(i)}\right)^{-1/2} \mathbf{R}^{-1}(\alpha) \left(\mathbf{A}^{(i)}\right)^{-1/2} \mathbf{s}^{(i)} = \mathbf{0}. \tag{7}$$

Here $\mathbf{s}^{(i)} = \mathbf{y}^{(i)} - \boldsymbol{\mu}^{(i)})$, and $\boldsymbol{\mu}^{(i)} = (\boldsymbol{\mu}^{(i)}_{\tau+1}, \ldots, \boldsymbol{\mu}^{(i)}_{\tau+n})^{\top}$ which depends on $\mathcal{W}$ (see Eq. (4) and (5)). The $n \times N$ matrix $\mathbf{D}^{(i)}$ is given by $\mathbf{D}^{(i)} = \partial\boldsymbol{\mu}^{(i)}/\partial\mathbf{w}$ where $\mathbf{w} = \mathrm{vect}(\mathcal{W})$ and $\left(\mathbf{D}^{(i)}\right)_{ab} = \partial(\boldsymbol{\mu}^{(i)})_a/\partial(\mathbf{w})_b$.

In an alternative QIF method, the working correlation no longer needs to be pre-specified as in GEE, which can be very inaccurate. Rather, it directly models $\mathbf{R}^{-1}(\alpha)$ as

$$\mathbf{R}^{-1}(\alpha) = \sum_{j=1}^{d} a_j \mathbf{M}_j \tag{8}$$

where $\mathbf{M}_j$'s are known $n \times n$ matrices characterizing various basic correlation structures and a_j's are unknown parameters. For example, an AR-1 correlation can be expressed as $\mathbf{R}^{-1}(\alpha) = a_1\mathbf{M}_1 + a_2\mathbf{M_2} + a_3\mathbf{M_3}$, where $\mathbf{M}_1$ is an identity matrix, $\mathbf{M}_2$ satisfies $(\mathbf{M}_2)_{i,j} = 1$ if $|i - j| = 1$, $(\mathbf{M}_2)_{i,j} = 0$ if $|i - j| \neq 1$, and M_3 has 1 at $(i, j) = (1, 1), (n, n)$ and zeros at other positions. Instead of solving a_j's associated with Eq. (7), we formulate our optimization problem via the so-called 'extended score' by substituting Eq. (8) for $\mathbf{R}^{-1}(\alpha)$ in Eq. (7):

$$\begin{aligned} \mathbf{g}_m(\mathcal{W}) :=& \frac{1}{m}\sum_{i=1}^{m} \mathbf{g}^{(i)}(\mathcal{W}) \\ :=& \frac{1}{m}\sum_{i=1}^{m} \begin{pmatrix} \left(\mathbf{D}^{(i)}\right)^{\top} \left(\mathbf{A}^{(i)}\right)^{-1/2} \mathbf{M}_1 \left(\mathbf{A}^{(i)}\right)^{-1/2} \mathbf{s}^{(i)} \\ \vdots \\ \left(\mathbf{D}^{(i)}\right)^{\top} \left(\mathbf{A}^{(i)}\right)^{-1/2} \mathbf{M}_d \left(\mathbf{A}^{(i)}\right)^{-1/2} \mathbf{s}^{(i)} \end{pmatrix} \end{aligned} \tag{9}$$

We may view each $\mathbf{g}^{(i)}(\mathcal{W})$ as a random vector $\mathbf{g}(\mathcal{X}, \mathbf{s}, \mathcal{W})$ evaluated at the data $\{\mathbf{s}^{(i)}, \mathcal{X}_{(i)} = (\mathcal{X}_{(i;\tau+1)}, \ldots, \mathcal{X}_{(i;\tau+n)})\}$.

The vector $\mathbf{g}_m(\mathcal{W})$ is an $(N \cdot d)$-dimensional column vector. In fact, substituting Eq. (8) into Eq. (7) yields a linear combination of the row blocks of $\mathbf{g}_m(\mathcal{W})$. Since $\mathbf{g}_m(\mathcal{W})$ has a larger dimension than $\mathcal{W}$, we cannot estimate $\mathcal{W}$ by simply solving $\mathbf{g}_m(\mathcal{W}) = \mathbf{0}$. Adapting the idea of Qu and Li (2006) and Qu et al. (2000), we obtain $\mathcal{W}$ by minimizing the weighted length of $\mathbf{g}_m(\mathcal{W})$:

$$\min_{\mathcal{W}} Q_m(\mathcal{W}) := m\mathbf{g}_m(\mathcal{W})^{\top}\mathbf{C}_m^{-1}(\mathcal{W})\mathbf{g}_m(\mathcal{W}), \tag{10}$$

where

$$\mathbf{C}_m(\mathcal{W}) = \frac{1}{m}\sum_{i=1}^{m} \mathbf{g}^{(i)}(\mathcal{W})\mathbf{g}^{(i)}(\mathcal{W})^{\top} \tag{11}$$

which estimates the covariance matrix of $\mathbf{g}_m$. The use of $\mathbf{C}_m$ leads to an efficient method (Hansen, 1982) because the calculation of $\mathbf{C}_m$, a direct estimate of the covariance, allows us to omit the step of estimating a_j's.

In our tensorQIF model, the loss function $l(\mathcal{W}) = Q_m(\mathcal{W})$ and the regularization term is given by Eq. (2). More precisely, we solve the following optimization problem:

$$\min_{\mathcal{W}_1, \mathcal{W}_2, \ldots, \mathcal{W}_K} Q_m(\mathcal{W}) + \sum_{k=1}^{K} \left(\lambda_k \sum_{j=1}^{d_k} \|\!|(\mathcal{W}_k)^{(j)}_{(k)}\|\!|_F \right) \tag{12}$$

where each $\mathcal{W}_k \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_K}$ and the final coefficient tensor $\mathcal{W} = \sum_{k=1}^{K} \mathcal{W}_k$.

3 Asymptotic Analysis

In this section we establish the asymptotic normality for our *TensorQIF* model as m approaches to infinity. We first rescale the objective function in Eq. (12):

$$\widetilde{Q}_m(\mathcal{W}) + \sum_{k=1}^{K} \left(\frac{\lambda_k}{m} \sum_{j=1}^{d_k} \|(\mathcal{W}_k)_{(k)}^{(j)}\|_F \right). \tag{13}$$

where $\widetilde{Q}_m = \mathbf{g}_m^\top \mathbf{C}_m^{-1} \mathbf{g}_m$. We require the following regularity conditions on the random vector $\mathbf{g}$ given after Eq. (9):

1. There exists a unique $\mathcal{W}^*$ that satisfies the mean zero model assumption, i.e. $\mathbb{E}[\mathbf{g}(\mathcal{W}^*)] = \mathbf{0}$.
2. The data $\{\mathcal{X}_{(i)}, \mathbf{s}^{(i)}\}'s$ are i.i.d. and the parameter space $\Omega := \Omega_1 \times \Omega_2 \times \cdots \times \Omega_K$ is compact.
3. $\mathcal{W}^*$ has a unique decomposition $\mathcal{W}^* = \sum_{k=1}^{K} \mathcal{W}_k^*$ such that for each k, $\mathcal{W}_k^*$ is an interior point of Ω_k.
4. Let $\mathbf{w} = \text{vect}(\mathcal{W})$. For all $\mathcal{W} \in \Omega$, $\|\mathbf{g}(\mathcal{W})\mathbf{g}(\mathcal{W})^\top\|_F \leqslant d_1(\mathcal{X}, \mathbf{s})$, $\|\nabla_{\mathbf{w}} \mathbf{g}(\mathcal{W})\|_F \leqslant d_2(\mathcal{X}, \mathbf{s})$ for some d_1, d_2 such that $\mathbb{E}[d_1(\mathcal{X}, \mathbf{s})]$ and $\mathbb{E}[d_2(\mathcal{X}, \mathbf{s})]$ are finite.

The proof the theorem is based on a uniform convergence result for stochastic functions. Using Lemma 2.4 in Newey and McFadden (1994), conditions 2, and 4, we obtain

Lemma 1. *Let* $\mathbf{C}_*(\mathcal{W}) = \mathbb{E}[\mathbf{g}(\mathcal{W})\mathbf{g}(\mathcal{W})^\top]$ *and* $\mathbf{J}_*(\mathcal{W}) = \mathbb{E}[\nabla_{\mathcal{W}} \mathbf{g}(\mathcal{W})]$. *Then we have*

$$\mathbf{C}_m(\mathcal{W}) \to \mathbf{C}_*(\mathcal{W}) \quad \textit{in probability} \tag{14}$$

and

$$\nabla_{\mathbf{w}} \mathbf{g}_m(\mathcal{W}) \to \mathbf{J}_*(\mathcal{W}) \quad \textit{in probability,} \tag{15}$$

uniformly for $\mathcal{W} \in \Omega$. *Moreover,* $\mathbf{C}_*(\mathcal{W})$ *and* $\mathbf{J}_*(\mathcal{W})$ *are uniformly continuous.*

Remark 1. By condition 1 and the weak law of large numbers, we have $\mathbf{g}_m(\mathcal{W}^*) \xrightarrow{p} \mathbf{0}$ as $m \to \infty$. The uniform convergence of the gradient in Eq. (15) then yields

$$\mathbf{g}_m(\mathcal{W}) \to \mathbb{E}[\mathbf{g}(\mathcal{W})] \quad \text{in probability} \tag{16}$$

uniformly for $\mathcal{W} \in \Omega$ and $\mathbb{E}[\mathbf{g}(\mathcal{W})]$ is continuous.

Under these regularity conditions, we have

Theorem 1. *Let* λ_k*'s be fixed constants and let* $\sum_{k=1}^{K} \hat{\mathcal{W}}_{k;m} := \hat{\mathcal{W}}_m$ *be the estimator obtained by minimizing Eq.* (13) *subject to* $\mathcal{W} = \sum_{k=1}^{K} \mathcal{W}_k$. *Then as* $m \to \infty$, *we have*

$$\hat{\mathcal{W}}_m \to \mathcal{W}^* \quad \textit{in probability,} \tag{17}$$

$$\sqrt{m} \cdot \textit{vect}\left(\hat{\mathcal{W}}_m - \mathcal{W}^*\right) \to \mathcal{N}(\mathbf{0}, (\mathbf{J}_0^\top \mathbf{C}_0^{-1} \mathbf{J}_0)^{-1}) \quad \textit{in distribution.} \tag{18}$$

where $\mathbf{C}_0 = \mathbf{C}_*(\mathcal{W}^*)$ *and* $\mathbf{J}_0 = \mathbf{J}_*(\mathcal{W}^*)$.

The proof can be found in Appendix.

4 Algorithm

In this section, we provide an algorithm to solve the optimization problem Eq. (12) followed by a convergence result. Since the sample size m is fixed throughout this section, we drop the subscript m in Eq. (12) and write Q_m as Q. We first give notations that will be used in our algorithm.

- $\Phi = (\mathcal{W}_1, \ldots, \mathcal{W}_K)$; $\mathcal{W}(\Phi) = \sum_{k=1}^{K} \mathcal{W}_k$.
- $F(\Phi) = Q(\mathcal{W}(\Phi)) + R(\Phi)$.
- $\Phi^{(r)} = (\mathcal{W}_1^{(r)}, \ldots, \mathcal{W}_K^{(r)})$; $\mathcal{W}^{(r)} = \mathcal{W}(\Phi^{(r)})$.

4.1 Optimization Algorithm

We develop a linearized block coordinate descent algorithm in the following iterative procedure to find optimal $\hat{\Phi}$ in Eq. (12). Denote the iterates at the r-th iteration by $\Phi^{(r)}$. At point $\Phi = (\mathcal{W}_1, \ldots, \mathcal{W}_K)$, let

$$R(\Phi) := \sum_{k=1}^{K} \left(\lambda_k \sum_{j=1}^{d_k} \| (\mathcal{W}_k)_{(k)}^{(j)} \|_F \right). \tag{19}$$

Assume $\nabla_{\mathcal{W}} Q(\mathcal{W})$ is Lipschitz continuous with Lipschitz modulus L_Q. The following $P_L(\Phi, \widetilde{\Phi})$ is a linearized proximal map for the non-smooth regularizer R:

$$P_L(\Phi, \widetilde{\Phi}) := Q(\widetilde{\mathcal{W}}) + R(\Phi) + \frac{KL}{2} \sum_{k=1}^{K} \| \mathcal{W}_k - \widetilde{\mathcal{W}}_k \|_F^2 + \left\langle \sum_{k=1}^{K} \left(\mathcal{W}_k - \widetilde{\mathcal{W}}_k \right), \nabla_{\mathcal{W}} Q(\widetilde{\mathcal{W}}) \right\rangle \tag{20}$$

where $L \geqslant L_Q$ is a fixed constant. Note that

$$\frac{L}{2} \| \mathcal{W} - \widetilde{\mathcal{W}} \|_F^2 \leqslant \frac{KL}{2} \sum_{k=1}^{K} \| \mathcal{W}_k - \widetilde{\mathcal{W}}_k \|_F^2. \tag{21}$$

The inequality (21) and the Lipschitz continuity of $Q(\mathcal{W})$ indicate that for all $L \geqslant L_Q$,

$$F(\Phi) \leqslant P_L(\Phi, \widetilde{\Phi}) \quad \text{for all } \Phi \text{ and } \widetilde{\Phi}. \tag{22}$$

At the r-th iteration, we update $\Phi^{(r+1)}$ by solving the following optimization problem

$$\min_{\Phi} \sum_{k=1}^{K} \left[\left\langle \nabla_{\mathcal{W}} Q^{(r)}, \mathcal{W}_k - \mathcal{W}_k^{(r)} \right\rangle + \frac{KL}{2} \| \mathcal{W}_k - \mathcal{W}_k^{(r)} \|_F^2 \right] + R(\Phi) \tag{23}$$

where $\nabla_{\mathcal{W}} Q^{(r)} = \nabla_{\mathcal{W}} Q(\mathcal{W}^{(r)})$. Since $R(\Phi)$ given in (19) is separable among $\mathcal{W}_k$'s, we can decompose the problem (23) into the following K separate subproblems:

$$\min_{\mathcal{W}_k} \left\langle \nabla_{\mathcal{W}} Q^{(r)}, \mathcal{W}_k - \mathcal{W}_k^{(r)} \right\rangle + \frac{KL}{2} \| \mathcal{W}_k - \mathcal{W}_k^{(r)} \|_F^2 + \lambda_k \sum_{j=1}^{d_k} \| (\mathcal{W}_k)_{(k)}^{(j)} \|_F \tag{24}$$

for $k \in \{1, \ldots, K\}$. Since the subproblems share the same structure, we may fix k and solve (24) to find the best $\mathcal{W}_k$, which is equivalent to

$$\min_{\mathcal{W}_k} \frac{1}{2} \left\| \mathcal{W}_k - \left(\mathcal{W}_k^{(r)} - \frac{1}{KL} \nabla_{\mathcal{W}} Q^{(r)} \right) \right\|_F^2 + \frac{\lambda_k}{KL} \sum_{j=1}^{d_k} \| (\mathcal{W}_k)_{(k)}^{(j)} \|_F. \tag{25}$$

The problem (25) has a closed-form solution $\mathcal{W}_k^{(r+1)}$ where each of its sub-tensor is

$$(\mathcal{W}_k^{(r+1)})_{(k)}^{(j)} = \max\left(0, 1 - \frac{\lambda_k}{KL|||(\mathcal{P}^{(r)})_{(k)}^{(j)}|||_F}\right)(\mathcal{P}^{(r)})_{(k)}^{(j)}, \tag{26}$$

and $\mathcal{P}^{(r)} := \mathcal{W}_k^{(r)} - \frac{1}{KL}\nabla_{\mathcal{W}} Q^{(r)}$. In fact, from optimality conditions, $\mathcal{W}_k^{(r+1)}$ satisfies

$$\nabla_{\mathcal{W}} Q^{(r)} + KL\left(\mathcal{W}_k^{(r+1)} - \mathcal{W}_k^{(r)}\right) + \lambda_k \mathcal{A}_k(\mathcal{W}_k^{(r)}) = 0 \tag{27}$$

for all $r \geqslant 1$ and $1 \leqslant k \leqslant K$. Here $\mathcal{A}_k(\mathcal{W})$ is a subgradient of $\sum_{j=1}^{d_k} |||(\mathcal{W})_{(k)}^{(j)}|||_F$. The calculation of the Lipschitz modulus L_Q can be computationally expensive. We therefore follow a similar argument in Xu et al. (2015) to find a proper approximation $\tilde{L} \geqslant L_Q$ and use $\tilde{L}$ as L in all of our computations. Algorithm 1 summarizes the steps for finding the optimal $\hat{\mathcal{W}}_k$.

Algorithm 1 Search for optimal $\hat{\Phi}$.

```
**Input:** $\mathcal{X}$, $\mathbf{y}$, L, $\lambda_k$
**Output:** $\hat{\Phi} = (\hat{\mathcal{W}}_1, \dots, \hat{\mathcal{W}}_K)$
1. $r = 0$: compute $\tilde{L}$ and initialize $\mathcal{W}_k^{(0)}$ for $1 \leqslant k \leqslant K$.
2. Obtain $\Phi^{(r+1)} = (\mathcal{W}_1^{(r+1)}, \dots, \mathcal{W}_K^{(r+1)})$ by solving (25) for each fixed $1 \leqslant k \leqslant K$.
3. $r = r + 1$.
Repeat 2 and 3 until convergence.
```

4.2 Convergence Analysis

In this section, we prove that the sequence $\{\Phi^{(r)}\}_{r\geqslant 0}$ generated by Algorithm 1 will converge to a global optimal solution $\hat{\Phi}$ with a convergence rate of $O(1/r)$ if the initial point $\Phi^{(0)}$ is located in a convex neighborhood of $\hat{\Phi}$. In Loader and Pilla (2007), it has been shown that the function $Q(\mathcal{W})$ is not globally convex in general. Hence the standard convergence arguments such as in Beck and Teboulle (2009) cannot be applied directly. Furthermore, with the latent approach $\mathcal{W} = \sum_{k=1}^K \mathcal{W}_k$, we have to carefully split or combine inequalities at certain points. All of these make the proof of the convergence nontrivial.

Let $\hat{\Phi} = (\hat{\mathcal{W}}_1, \dots, \hat{\mathcal{W}}_K)$ be a global minimizer of $F(\Phi)$ and $\Omega = \Omega_1 \times \dots \Omega_K$ is a neighborhood of $\hat{\Phi}$ such that $\Pi(\Omega) := \{\mathcal{W}(\Phi) : \Phi \in \Omega\}$ is convex and $Q(\mathcal{W})$ is convex in $\Pi(\Phi)$. Assume $\Phi^{(0)}$ satisfies

$$D(\Phi^{(0)}) := \sum_{k=1}^K |||\mathcal{W}_k^{(0)} - \hat{\mathcal{W}}_k|||_F^2 < \frac{1}{K}\left[\operatorname{dist}(\partial\Pi(\Omega), \hat{\mathcal{W}})\right]^2. \tag{28}$$

We first present a lemma which provides a key inequality in our proof of convergence.

Lemma 2. *Assume $\Phi^{(r)} = (\mathcal{W}_1^{(r)}, \dots, \mathcal{W}_K^{(r)})$ such that $\mathcal{W}(\Phi^{(r)}) \in \Pi(\Omega)$. Let $\Phi^{(r+1)} = (\mathcal{W}_1^{(r+1)}, \dots, \mathcal{W}_K^{(r+1)})$ be a minimizer of Eq. (25). Then for any $L \geqslant L_Q$ and for any $\Phi = (\mathcal{W}_1, \dots \mathcal{W}_K)$ such that $\mathcal{W}(\Phi) \in \Pi(\Omega)$, we have*

$$F(\Phi) - F(\Phi^{(r+1)}) \geqslant \frac{KL}{2}\sum_{k=1}^K |||\mathcal{W}_k^{(r+1)} - \mathcal{W}_k^{(r)}|||_F^2 + KL\sum_{k=1}^K \left\langle \mathcal{W}_k^{(r)} - \mathcal{W}_k, \mathcal{W}_k^{(r+1)} - \mathcal{W}_k^{(r)} \right\rangle. \tag{29}$$

Lemma 3. *Let $\hat{\mathcal{W}} = \mathcal{W}(\hat{\Phi})$. Suppose $\Phi^{(r)}$ satisfy*

$$D(\Phi) := \sum_{k=1}^{K} \|\mathcal{W}_k - \hat{\mathcal{W}}_k\|_F^2 < \frac{1}{K}\left[dist(\partial\Pi(\Omega), \hat{\mathcal{W}})\right]^2. \tag{30}$$

Then $\mathcal{W}(\Phi^{(r+1)})$ generated by (25) also satisfies (30).

Remark 2. Lemma 3 implies that all points in the sequence $\{\Phi^{(r)}\}_{r\geqslant 0}$ generated by Algorithm 1 satisfy (30) if the initial point $\Phi^{(0)}$ does. In particular, we have $\{\mathcal{W}(\Phi^{(r)})\}_{r\geqslant 0} \subset \Pi(\Omega)$. Thus we can apply Lemma 2 for all $r \geqslant 0$.

Then we have the following convergence result.

Theorem 2. *Let $\Phi^{(n)}$ be the tuple of tensors generated by Algorithm 1 at the n-th iteration. Then for any $n \geqslant 1$,*

$$F(\Phi^{(n)}) - F(\hat{\Phi}) \leqslant \frac{KL\sum_{k=1}^{K}\|\mathcal{W}_k^{(0)} - \hat{\mathcal{W}}_k\|_F^2}{2n}. \tag{31}$$

The proof can be found in Appendix.

Remark 3. Lemma 2 still holds if $\Phi^{(r)}$ is replaced by any $\widetilde{\Phi}^{(r)}$ such that $\mathcal{W}(\widetilde{\Phi}^{(r)}) \in \Pi(\Omega)$. Furthermore, from the proof of Lemma 3, we deduce that the minimizer of (25) generated by $\widetilde{\Phi}^{(r)}$ will satisfy (30) if $\widetilde{\Phi}^{(r)}$ does.

4.3 Group Support: Values of λ_k's and L

In this section we focus on the linear model in which each component of $\boldsymbol{\eta}^{(i)}$ is given by $\eta_t^{(i)} = \left\langle \mathcal{X}_{(i;t)}, \sum_{k=1}^{K}\mathcal{W}_k \right\rangle$ and the components of outcome $\mathbf{y}^{(i)}$ are of the form $y_t^{(i)} = \left\langle \mathcal{X}_t^{(i)}, \mathcal{W}^* \right\rangle + s_t^{(i)}$ for some true tensor coefficient $\mathcal{W}^* = \sum_{k=1}^{K}\mathcal{W}_k^*$, where $\tau \leqslant t \leqslant T$, and $\mathcal{W}_k^*s$ follow certain true patterns. Let $\mathcal{D} := \nabla_{\mathcal{W}} Q(\mathcal{W}^*)$. Motivated by the algorithm, we consider the following optimization problem for a fixed k:

$$\min_{\mathcal{W}_k} \frac{1}{2}\|\mathcal{W}_k - \mathcal{W}_k^* + \mathcal{D}\|_F^2 + \frac{\lambda_k}{KL}\sum_{j=1}^{d_k}\|(\mathcal{W}_k)_{(k)}^{(j)}\|_F. \tag{32}$$

Our goal is to estimate the group support for $\mathcal{W}_k^*$, i.e. obtain the subset $S_k^* \subset \{1, 2, \ldots, d_k\}$ such that $(\mathcal{W}_k^*)_{(k)}^{(j)} \neq 0$ if and only if $j \in S_k^*$. The Karush–Kuhn–Tucker (KKT) conditions for solutions of (32) immediately imply the following lemma.

Lemma 4 (KKT)**.** *Assume $\hat{\mathcal{W}}_k$ is a solution of (32). Then either*

$$(\hat{\mathcal{W}}_k)_{(k)}^{(j)} \neq 0 \quad \textit{and} \quad (\hat{\mathcal{W}}_k)_{(k)}^{(j)} - (\mathcal{W}_k^*)_{(k)}^{(j)} + \mathcal{D}_{(k)}^{(j)} = -\frac{\lambda_k}{KL}\frac{(\hat{\mathcal{W}}_k)_{(k)}^{(j)}}{\|(\hat{\mathcal{W}}_k)_{(k)}^{(j)}\|_F},$$

or

$$(\hat{\mathcal{W}}_k)_{(k)}^{(j)} = 0 \quad \textit{and} \quad \|(\mathcal{W}_k^*)_{(k)}^{(j)} + \mathcal{D}_{(k)}^{(j)}\|_F \leqslant \frac{\lambda_k}{KL}.$$

Lemma 4 then yield

Theorem 3. *Assume*

$$\frac{\lambda_k}{2} \geqslant \max_{1\leqslant j\leqslant d_k} |||\mathcal{D}^{(j)}_{(k)}|||_F. \tag{33}$$

Then (32) *has a solution* $\hat{\mathcal{W}}_k$ *such that*

$$\{j : (\hat{\mathcal{W}}_k)^{(j)}_{(k)} \neq 0\} := \hat{S}_k \subset S_k. \tag{34}$$

Furthermore, $\hat{S}_k = S^*_k$ *if* $\lambda_k < \frac{KL}{2}\min_{j\in S} |||(\mathcal{W}^*_k)^{(j)}_{(k)}|||_F$.

The proof can be found in Appendix.

5 Empirical Evaluation

In this section we present the results of both synthetic and real-life fMRI and EEG examples. We test the efficiency and effectiveness of the proposed method *TensorQIF* comparing to the state-of-the-art methods. The datasets containing continuous responses have a format as described in Section 2.2: $\{y_t^{(i)}, \mathcal{X}_t^{(i)} : 1 \leqslant i \leqslant m, 0 \leqslant t \leqslant T\}$. Here i denotes the subject id and t is a time point. For both synthetic and fMRI cases, each $\mathcal{X}_t^{(i)}$ is a matrix (i.e. a 2-way tensor).

5.1 Simulations

We examine the following methods: *TensorQIF*, Least Absolute Shrinkage and Selection Operator (LASSO), Graphical Granger Modeling (Lozano et al., 2009), GEE (Liang and Zeger, 1986a), and Kruskal (Zhou et al., 2013). The LASSO uses only the current record, the matrix $\mathcal{X}_t^{(i)}$, as the covariate to make a prediction on $y_t^{(i)}$, whereas the Granger and our TensorQIF have a tensor covariate. That is, they use $\mathcal{X}_{(i;t)}$ described in Section 2.2 as the input, which is a 3-way tensor formed by concatenating the current and several previous $\mathcal{X}_{(i;t)}$'s. In fact, the Granger model is equivalent to the LASSO with a tensor input. To show the importance of considering lagged effect and conduct a fair comparison between methods, we will demonstrate the results on both matrix and tensor inputs for GEE and Kruskal methods.

We consider the settings $(d_1, d_2, \tau + 1) = (2, 2, 3)$, $(3, 3, 3)$, and $(5, 5, 5)$ i.e. $\mathcal{X}_t^{(i)} \in \mathbb{R}^{2\times 2}$, $\mathbb{R}^{3\times 3}$, and $\mathbb{R}^{5\times 5}$. The tensor input $\mathcal{X}_{(i;t)} \in \mathbb{R}^{2\times 2\times 3}$, $\mathbb{R}^{3\times 3\times 3}$, and $\mathbb{R}^{5\times 5\times 5}$. Entries of $\mathcal{X}_t^{(i)}$ are generated by drawing from the normal distribution $N(0, 1)$ first and adding the uniform distribution $U(0, \sin(t))$. The number of time points is 10 and after concatenating the current and previous $\tau = 2$ records, we obtain $\mathcal{X}_{(i;t)}$ for $\tau + 1 \leqslant t \leqslant 10$. We assign the true latent tensor coefficients $\mathcal{W}_1$, $\mathcal{W}_2$, and $\mathcal{W}_3$ a non-zero pattern in the first feature along the directions 1, 2, and 3 respectively. The non-zero entries in $\mathcal{W}_k$s follow the distribution $c_k N(0, 1)$. Here we assign $\mathcal{W}_k$s different scales: $c_1 = 0.1$, $c_2 = 1.0$ and $c_3 = 0.01$. Finally, we set $\mathcal{W} = \mathcal{W}_1 + \mathcal{W}_2 + \mathcal{W}_3$. For each subject i, the outcome (observed) $y_t^{(i)}$ is calculated by $y_t^{(i)} = \langle \mathcal{X}_{(i;t)}, \mathcal{W}\rangle + s_t^{(i)}$, where the residual $\mathbf{s}^{(i)} \in \mathbb{R}^8$ is generated from the multivariate normal distribution of mean $\mathbf{0}$ and AR(1) correlation structure with $\sigma^2 = 4.0$ and $\alpha = 0.8$.

We generate 100 synthetic datasets each containing 1000 subjects and a test set containing 10000 subjects. Only the true coefficients $\mathcal{W}_1$, $\mathcal{W}_2$, and $\mathcal{W}_3$ are fixed across all datasets. In each fitting procedure, 80% of subjects form a training set and the remaining 20% are used for the validation that helps selecting hyper parameters in models. We examine the model performances in two error metrics on the test set: 1. the mean squared error (MSE) between the observed y and the predictive $\hat{y} = \langle \mathcal{X}, \hat{\mathcal{W}}\rangle$; and 2. the true root mean square error (RMSE) between true $\bar{y} = \langle \mathcal{X}, \mathcal{W}\rangle$ and $\hat{y}$.

Table 1: Simulation results from 100 replicates for dimensions $d_1 \times d_2 \times (\tau+1)$. The true correlation structure is AR(1). Reported are the average of MSE/RMSE.

Average MSE between observed y and the predictive $\hat{\mathbf{y}}$

	LASSO	Granger	Kruskal (rk2)		Kruskal (rk3)	
	matrix	tensor	matrix	tensor	matrix	tensor
$2 \times 2 \times 3$	8.120	3.886	8.119	3.892	8.119	3.885
$3 \times 3 \times 3$	9.994	3.974	10.03	5.332	9.993	4.367
$5 \times 5 \times 5$	27.56	4.048	27.65	5.144	27.60	4.370

	GEE (AR1)		GEE (ind)		TensorQIF (AR1)	TensorQIF (ind)
	matrix	tensor	matrix	tensor	tensor	tensor
$2 \times 2 \times 3$	8.166	3.880	8.119	3.885	**3.871**	3.884
$3 \times 3 \times 3$	10.11	3.973	9.993	3.983	**3.970**	3.979
$5 \times 5 \times 5$	27.58	4.033	27.58	4.078	**4.012**	4.035

Average RMSE between true $\bar{y}$ and the predictive $\hat{y}$ with tensor inputs

	Granger	Kruskal (rk2)	Kruskal (rk3)
$2 \times 2 \times 3$	0.090	0.116	0.089
$3 \times 3 \times 3$	0.125	1.231	0.611
$5 \times 5 \times 5$	0.253	1.076	0.613

	GEE (AR1)	GEE (ind)	TensorQIF (AR1)	TensorQIF (ind)
$2 \times 2 \times 3$	0.058	0.088	**0.054**	0.085
$3 \times 3 \times 3$	0.081	0.128	**0.077**	0.119
$5 \times 5 \times 5$	0.214	0.301	**0.196**	0.238

Since the Kruskal model focuses on the low rank decomposition for $\mathcal{W}$, we conduct the simulation by setting rank= 2 (rk2) and rank= 3 (rk3). Furthermore, to compare the results under miss specified correlation structures, we consider both AR(1) and independent (Id) correlation settings in GEE and TensorQIF. The average of predictive MSE or RMSE with the true synthetic model on the test set over 100 replications with different models. The settings are summarized in Table 1.

In Table 1, the proposed TensorQIF outperforms the other regression methods in terms of the average predicting accuracy (MSE) and the coefficient estimation (RMSE). Since the synthetic datasets are generated by using $\tau \geqslant 2$, i.e. the outcome depends on the current and previous τ records, we see that the models using matrix inputs (only current record) suffer larger errors. Granger and Kruskal models does not handle the within sample correlation, so they result higher mean MSE/RMSE even with tensor inputs. To further examine the importance of modeling correlation, we conduct the paired t-test on the 100 predictive MSE generated by each model fitting. We consider TensorQIF (AR1) v.s. Granger and TensorQIF (AR1) v.s. TensorQIF (Id). The p values are given in Table 2.

Table 2: p values of paired t-test with TenaorQIF (AR1): considering correlation or not.

	Granger	GEE (ind)	TensorQIF (ind)
$2 \times 2 \times 3$	4.08E-14	8.47E-13	8.04E-13
$3 \times 3 \times 3$	4.15E-14	5.16E-23	2.83E-18
$5 \times 5 \times 5$	1.71E-30	1.39E-48	1.67E-14

Table 3: p values of paired t-test between TensorQIF and GEE when both use correct correlation structure (AR1) and both use incorrect one (Id).

	TenaorQIF (AR1) v.s. GEE (AR1)	TensorQIF (ind) v.s. GEE (ind)
$2 \times 2 \times 3$	3.31E-08	1.07E-02
$3 \times 3 \times 3$	2.72E-09	1.04E-09
$5 \times 5 \times 5$	6.27E-23	2.00E-40

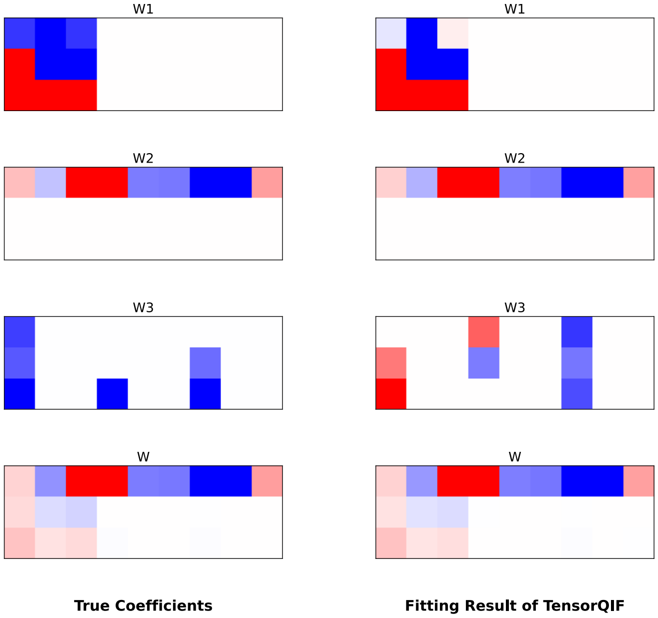


Figure 3: On synthetic data: the true coefficients and a TensorQIF fitting result.

The simulation also confirms that with the correct correlation structure (AR1), the fitting results of GEE (with an tensor input) and *TensorQIF* are more accurate. When both models are under mis-specified correlation structure (such as using the i.i.d. assumption), Table 1 shows that the proposed *TensorQIF* gives a lower average predictive MSE and more accurate coefficient estimations. We also conduct the paired t-test on the predictive MSE for model pairs in GEE (AR1), GEE (ind), *TensorQIF* (AR1), and *TensorQIF* (ind). The p values are given in Table 3. With larger models (more coefficients), these algorithms produce larger differences in MSE.

Figure 3 shows an example of *TensorQIF* (AR1) fitting result on a simulated dataset of $(d_1, d_2, \tau + 1) = (3, 3, 3)$, and $\lambda_1 = \lambda_2 = \lambda_3 = 350$. The white spaces represent zero coefficients;

red and blue colors represent positive and negative values respectively. We see that the proposed model captures the preassigned patterns in each of the three directions and recovers the true coefficient $\mathcal{W}$.

5.2 fMRI Data in Tensor

Functional magnetic resonance imaging (fMRI) is a functional neuroimaging procedure using MRI technology that measures brain activity by detecting associated changes in blood flow. The fMRI data used in the experiment were collected by the Alzheimer's Disease Neuroimaging Initiative (ADNI).[1] We cleaned up the fMRI data by filtering out the incomplete or low quality observations. After data cleaning, the data included 147 subjects diagnosed with mild cognitive impairment (MCI) from the year of 2009 to 2016. We used the participants' first fMRI scans as baseline and the other fMRI scans in 6th, 12th, 18th, and 24th months of the study. There were 67 brain areas and 4 properties (CV,SA,TA,TS) of the brain cortex[2] in our model. These properties were **CV**: Cortical Volume; **SA**: Surface Area; **TA**: Thickness Average; **TS**: Thickness Standard Deviation. Each example record naturally formed a 3-way tensor with one dimension for brain areas, one for signal property, and one along the temporal line. Our *TensorQIF* used the tensor directly without squashing dataset into a vector which may cause losing the proximity information. We aimed to predict the *mini-mental state examination* (MMSE) score quantified by a 30-point questionnaire, which is used extensively in clinical and research settings to measure cognitive impairment. At each time point, the MMSE score would be evaluated from participants' responses to the questionnaire.

We used 20% of subjects for testing, and set $\tau = 2$. The λ_1, λ_2, and λ_3 were tuned in a two-fold cross validation. In other words, the training records were further split into half: one used to build a model with a chosen parameter value from a range of 1 to 20 with a step size of 0.1; and the other used to test the resultant model. We chose the parameter values that gave the best two-fold cross validation performance.

The *TensorQIF* was able to select patterns along the three dimensions: among the features, among the brain areas, and among the different time points of month. The λ's were chosen as $\lambda_1 = 6$, $\lambda_2 = 20$, and $\lambda_3 = 24$. From Figure 4, we see that the structural damage of AD starting 6 months ago plays a major role in the current AD progression. Larger means and standard derivations of the thickness imply a higher risk of the AD. The proposed model selected 14 out of 68 brain areas that affect the MMSE score. According to the selected brain areas, signals in the Cuneus area, Transverse Temporal area in both sides, and the data at right Inferior Parietal area might be important in predicting the cognitive impairment together with a few other features.

5.3 EEG Data in Matrix

Human memory function can be assayed in real-time by EEG recording. In this section, we discuss the preliminary results we obtained on our (single trial) EEG data that were collected during Sternberg working memory tasks. In our study data, schizophrenia (SZ) patients went through three sessions of the Sternberg trials, and healthy normal (HN) members were only included in the first session. There were 90 trials in each session for each individual. However, very few patients participated all sessions and many trial records had missing values or significant

[1] http://adni.loni.usc.edu/

[2] http://adni.bitbucket.io/reference/ucsffresfr.html

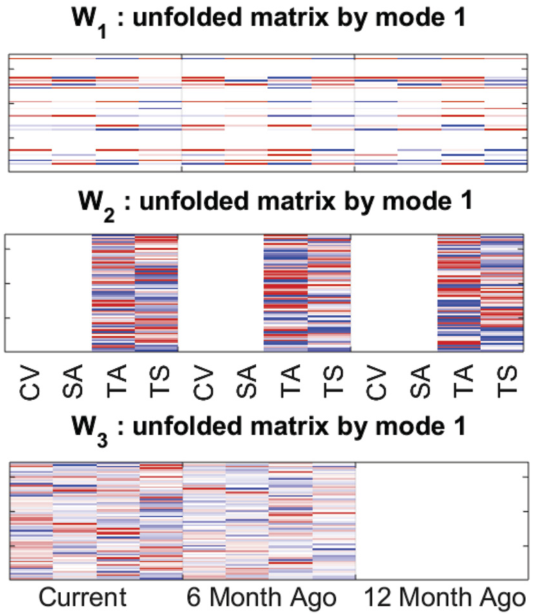


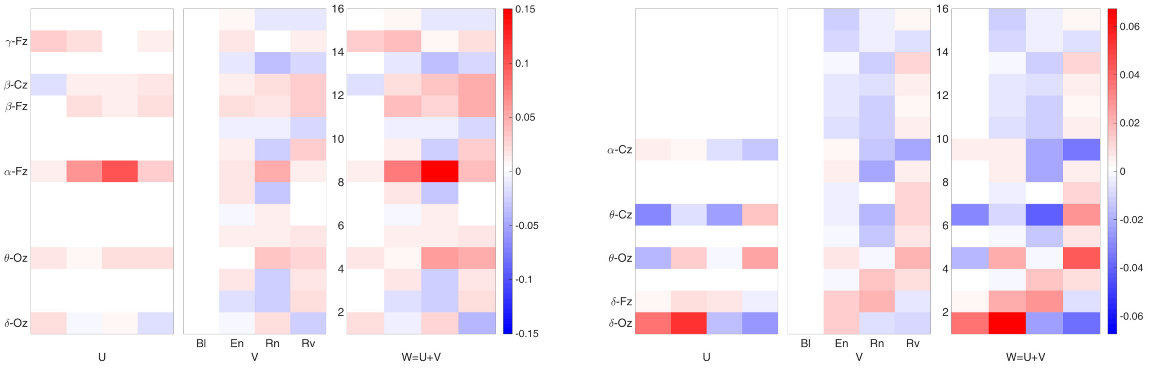
Figure 4: The columns, rows, and slices of the fMRI tensor selected by the TensorQIF for predicting MMSE score of participants.

Table 4: Comparison of AUC values (in percentage) between our approach and the GEE method on both healthy normal and schizophrenia data and for all different assumptions of correlation structures. (ind - independent sample-correlation structure.)

	GEE				Our Approach			
Population	AR(1)	Exchangeable	Tri-diagonal	ind	AR(1)	exchangeable	Tri-diagonal	ind
Healthy Normal (HN)	54.1	52.2	55.5	57.3	55.1	54.9	55.0	68.0
Schizophrenia (SZ)	60.3	55.5	43.6	65.0	62.6	60.0	48.2	66.3

level of noise or outliers, for which we had to clean the data carefully. After data cleaning, there were 1,131 trials for 14 SZ in session 1, 761 trials for 9 SZ in session 2, and 1,191 trials for 14 SZ in session 3. Each patient had 74 to 94 trials, and 83 on average. The rate of incorrect responses for the SZ patients was 27.2%. There were 519 trials for 6 HN participants. Each participant had 82 to 90 trials, and 87 on average. The rate of incorrect responses for HN participants was 14.7%. Our study data contained a limited sample of subjects from the original parent study.

We validated the proposed approach by comparing it to the most relevant method, which was the GEE (Liang and Zeger, 1986b). We experimented with the different correlation structures including exchangeable, tri-diagonal, AR-1 and independent formula. The receiver operating characteristic (ROC) curves were used to evaluate the performance of each resultant classifier and the area under the ROC curve (AUC) was reported in Table 4 (Fawcett, 2006). We separated our analysis for SZ and HN with the hypothesis that SZ patients may use different mechanisms or brain functions to perform memory tasks from those of HN participants. We hence built classifiers to separate trials with correct responses from those with incorrect responses, respectively, for SZ and HN. We then compared the features selected for use in the SZ classifiers and HN classifiers.



(a) (b)

Figure 5: Columns and rows selected by the TensorQIF classifier for separating correct versus incorrect Sternberg trials of SZ patients (a) and HN participants (b). Red (blue) color indicates that the corresponding features were positive (negative) predictors of the incorrect response. Features with white color were not used in the classifier.

For each of the SZ and HN datasets, one third of the records were randomly chosen from every subject to form the test data and the rest of the records were used in training. The hyper-parameters λ_1 and λ_2 in our approach and GEE (one parameter) were tuned in a two-fold cross validation within the training data. In other words, the training records were further split in half: one used to build a classifier with a chosen parameter value from a range of 1 to 10 with a step size 0.1; and the other used to test the resultant classifier. We chose the parameter values that gave the best two-fold cross validation performance, which were $\lambda_1 = 5.9$ and $\lambda_2 = 10$ for SZ and $\lambda_1 = 2$ and $\lambda_2 = 3.1$ for HN.

Table 4 provides the AUC comparison results (shown in percentages) between the two methods and for different datasets and sample correlation assumptions. The results in Table 4 show that our approach outperformed the traditional GEE in almost all comparison scenarios in terms of classification accuracy. Most importantly, our approach was able to select along two dimensions: among the features and among the memory information processing stages. Traditional GEE did not have any shrinkage effect to select features. The advanced version of GEE used in our experiments implemented a ℓ_1 regularizer, so it could select among all 60 features. Because it did not use the spatio-temporal structure of the 60 features, it was unable to model along the different dimensions (locations versus temporal stages).

We noticed that both GEE and our approach performed the best when using independent sample-correlation assumption, which was naturally against our intuition because there were multiple trials from a single individual and these trials were expected to correlate. The equi-correlated (exchangeable) assumption assumed that the correlation among all trials was equal and indicated by a constant. Together with AR-1 and Tri-diagonal correlation structures, these assumptions were slightly worse than the independent correlation assumption. However, we also noticed that the trials were not labeled in sequence in our data so the algorithms would not be able to model and distinguish the correlations between consecutive trials from those of far-apart trials. (The trials that an individual performed in a short continuous timeframe may correlate more strongly than trials far apart.)

We include two figures to demonstrate the selected features and stages in the classifiers constructed by our approach. The selected features for SZ patients are shown in Figure 5(a). The selected features for HN participants are shown in Figure 5(b). An obvious observation is that the two populations selected quite different features but the most important information processing stages were the same. Some of the selected EEG features replicate those early reports, including upward modulation of γ in SZ patients and engagement of α during encoding and retention periods (Chen et al., 2014; Herrmann et al., 2004).

Based on our models, the two groups showed remarkably different patterns, with EEG activity in higher frequency bands during the encoding stage associated with incorrect trial responses in SZ (Figure 5(a)). However, these features were positive predictors of trial accuracy in healthy participants (Figure 5(b)), for whom engagement of low frequency activity was associated with incorrect responses. It appears that the SZ patients used more brain areas in the memory tasks than the HN participants. Frontal γ was previously identified as important for both SZ and HN subjects, but was not selected for HN participants in our new model, which may warrant further investigation. On the other hand, among the selected three stages of both groups, the features during the retention stage tended to receive the largest weights in magnitude on average. All these results will require careful examination in new studies to confirm the validity and replicate on independent samples.

5.4 EEG Data in Tensor

The clinical utility of the EEG method depends on the reliable determination of functionally and diagnostically relevant features. The proposed *TensorQIF* capable of modeling non-stationary signal has been explored as a way to synthesize large arrays of EEG data because the EEG record could be more precisely characterized by a 3-way tensor representing processing stages, spatial locations, and frequency bands as individual dimensions.

Participants of $n = 40$ SZ patients and $n = 20$ HN participants completed an EEG Sternberg task. EEG was analyzed to extract 5 frequency components (delta, theta, alpha, beta, gamma) at 4 processing stages (baseline, encoding, retention, retrieval) and 12 scalp sites representing central midline, and bi-lateral frontal and temporal regions. The proposed and comparing methods were applied to the resulting 240 features (forming a $5 \times 4 \times 12$ tensor) to classify correct (-1) vs. incorrect (+1) responses on a trial-by-trial basis. In this approach, the proposed method guided the respective selection of spectral frequency, temporal (processing stages), and spatial (electrode sites) dimensions most related to trial performance. The correlations among processing stages were also estimated by the proposed method. Separate models were constructed for SZ and HN samples for comparison of common and disparate feature patterns across the dimensions.

For each of the SZ and HN datasets, one fifth of the records were randomly chosen from every subject to form the test data and the rest of the records were used in training. The hyperparameters λ_1, λ_2, and λ_3 were tuned in a two-fold cross validation within the training data. We chose the parameter values that gave the best two-fold cross validation performance, which were $\lambda_1 = 7.5$, $\lambda_2 = 5.5$, $\lambda_3 = 7.4$ for SZ and $\lambda_1 = 3.3$, $\lambda_2 = 2.1$, $\lambda_3 = 3.1$ for HN.

As shown in Figure 6, in both groups, task performance is most dependent on encoding and retrieval stage activity, with higher encoding uniformly and lower retrieval activity generally associated with better task performance across electrode sites. This pattern appears most prominently in central alpha activity (Figure 6; blue border). This indicates the same findings as in Xu et al. (2015). Groups differed in two main ways: (1) centroparietal theta, beta, and gamma during encoding and retention have lower values in HN (Figure 6; red border), and (2)

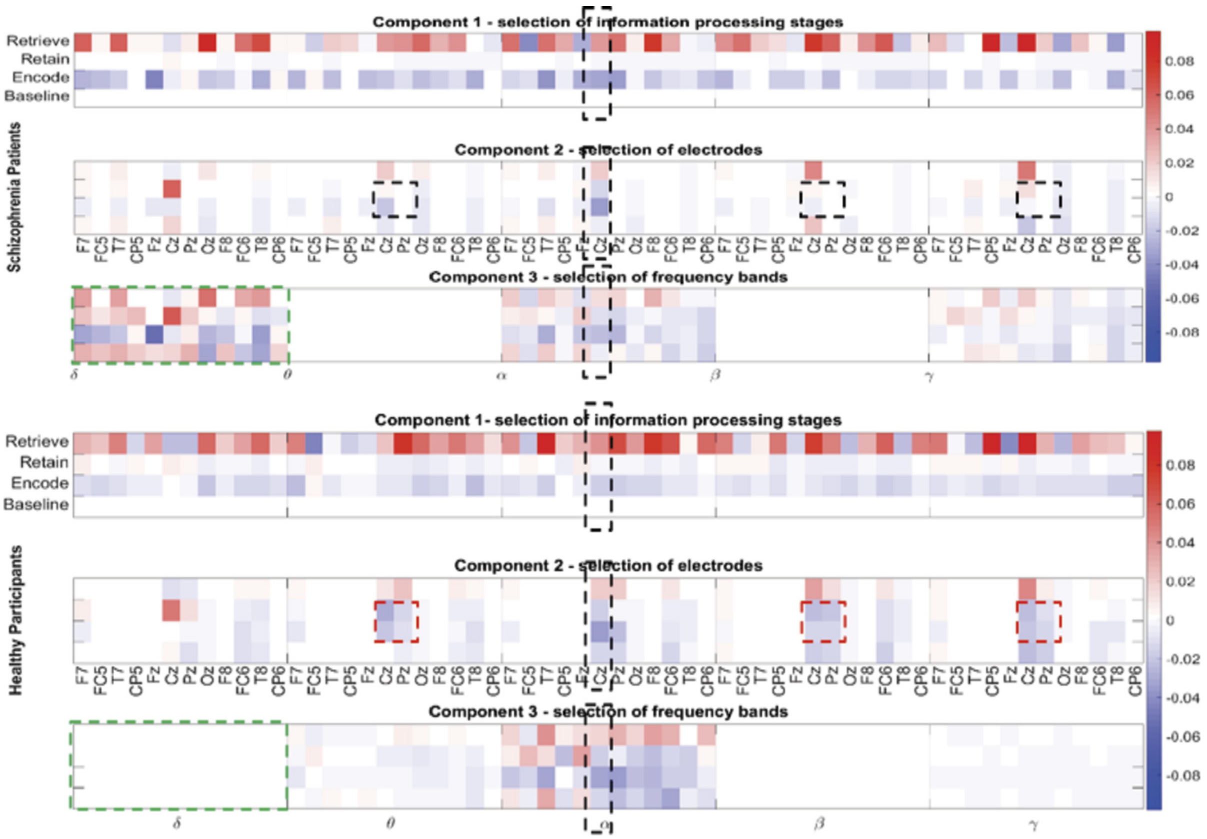


Figure 6: The columns, rows, and slices of the EEG tensor selected by the TensorQIF to predict the success of the memory tasks for SZ (top) and HN (bottom), respectively.

the delta activity across stages and electrodes (Figure 6; green border) was selected in SZ but no in HN. Here the experimental results give much clearer details of the working electrode sites and spectral frequencies comparing to the results in Johannesen et al. (2016). The proposed method outperform GEE and SVM solutions according to AUC values (HN: 55.5%; SZ: 58.8% versus the best AUC 53% from the other methods). This is because the proposed method enabled interpretation and summary across all dimensions, which is not possible for classifiers based on single vectors.

6 Conclusion

We have proposed a new learning formulation — *TensorQIF* — to analyze longitudinal data. It takes data matrices or tensors as inputs and make predictions. The proposed method can simultaneously determine the temporal contingency and the influential features from the observations of different modes without breaking into multiple models. The method creates a generalized linear model which has parameters forming a coefficient tensor. This coefficient tensor is computed by the summation of K component tensors so that each reflects the selection among a particular mode. Asymptotic analysis shows that the proposed formulation finds true coefficients when the sample size approaches to infinity. Moreover, the related optimization problem can be efficiently solved by a linearized block coordinate descent algorithm which has a sublinear convergence rate. The simulation results demonstrate the superior performance of the proposed method, and applications on real-life datasets show insightful discoveries. There can be a few future directions. For instance, we will try to formulate our *TensorQIF* into a more scalable learning model so large datasets can benefit from this approach. Model architectures such as deep neural networks might also be used in conjunction with our proposed regularization decomposition for broader utility. Other applications such as in single-cell sequencing data analysis might further show the power of our approach in understanding biological predictors.

Supplementary Material

The code and data can be found: https://doi.org/10.6084/m9.figshare.19166474.v1.

For data generation, we provide DataGenerator.py to generate synthetic data including training and test sets; For model fitting, we provide tensorQIF_model_Tensorflow_v2.py to run models and ReportGenerator.py to report on performance. For experiments comparisons, we have Granger_model.py, GEE_model.m, Kruskal_model.m.

Appendix

Proof of Theorem 1. Since $\hat{\mathcal{W}}_m$ is a minimizer, we have

$$\widetilde{Q}_m(\hat{\mathcal{W}}_m) + \sum_{k=1}^{K}\left(\frac{\lambda_k}{m}\sum_{j=1}^{d_k} \vert\vert\vert(\hat{\mathcal{W}}_{k;m})_{(k)}^{(j)}\vert\vert\vert_F\right) \leqslant \widetilde{Q}_m(\mathcal{W}^*) + \sum_{k=1}^{K}\left(\frac{\lambda_k}{m}\sum_{j=1}^{d_k} \vert\vert\vert(\mathcal{W}_k^*)_{(k)}^{(j)}\vert\vert\vert_F\right). \tag{35}$$

Note that

$$\begin{aligned} |\widetilde{Q}_m(\mathcal{W}^*)| &= \left|\mathbf{g}_m^\top(\mathcal{W}^*)\mathbf{C}_m^{-1}(\mathcal{W}^*)\mathbf{g}_m(\mathcal{W}^*)\right| \\ &\leqslant \left|\mathbf{g}_m^\top(\mathcal{W}^*)[\mathbf{C}_m^{-1}(\mathcal{W}^*) - \mathbf{C}_*^{-1}(\mathcal{W}^*)]\mathbf{g}_m(\mathcal{W}^*)\right| + \left|\mathbf{g}_m^\top(\mathcal{W}^*)\mathbf{C}_*^{-1}(\mathcal{W}^*)\mathbf{g}_m(\mathcal{W}^*)\right|. \end{aligned} \tag{36}$$

By Eq. (14), condition 1, and the weak law of large numbers, we deduce

$$\left|\widetilde{Q}_m(\mathcal{W}^*)\right| \to 0 \quad \text{in probability.} \tag{37}$$

Therefore from Eq. (35), we obtain

$$\left|\widetilde{Q}_m(\hat{\mathcal{W}}_m)\right| \to 0 \quad \text{in probability} \tag{38}$$

for fixed λ_k's. Using Eq. (14) and Eq. (16) we also have

$$|\widetilde{Q}_m(\hat{\mathcal{W}}_m) - \mathbb{E}[\mathbf{g}(\hat{\mathcal{W}}_m)]^\top \mathbf{C}_*(\hat{\mathcal{W}}_m)\mathbb{E}[\mathbf{g}(\hat{\mathcal{W}}_m)]| \to 0 \quad \text{in probability.} \tag{39}$$

Thus $\mathbb{E}[\mathbf{g}(\hat{\mathcal{W}}_m)] \to 0$ and Eq. (17) is followed by the uniqueness in condition 1 and the continuity of $\mathbb{E}[\mathbf{g}(\mathcal{W})]$ in Remark 1.

For m is large enough, we may assume the minimizer $\hat{\mathcal{W}}_m$ is an interior point which satisfies the Euler-Lagrange equation:

$$\nabla_{\mathbf{w}}\widetilde{Q}_m(\hat{\mathcal{W}}_m) + o(1) = 0. \tag{40}$$

Using the mean value theorem we obtain

$$\nabla_{\mathbf{w}}\widetilde{Q}_m(\mathcal{W}^*) + \nabla_{\mathbf{w}}^2\widetilde{Q}_m(\widetilde{\mathcal{W}}_m)\mathrm{vect}(\hat{\mathcal{W}}_m - \mathcal{W}^*) = o(1) \tag{41}$$

for some $\widetilde{\mathcal{W}}_m$ between $\hat{\mathcal{W}}_m$ and $\mathcal{W}^*$. Then we have

$$\sqrt{m}\cdot \mathrm{vect}(\hat{\mathcal{W}}_m - \mathcal{W}^*) = -\sqrt{m}[\nabla^2_{\mathbf{w}}\widetilde{Q}_m(\widetilde{\mathcal{W}}_m)]^{-1}[\nabla_{\mathbf{w}}\widetilde{Q}_m(\mathcal{W}^*) + o(1)]. \tag{42}$$

A direct calculation shows

$$\nabla_{\mathbf{w}}\widetilde{Q}_m = 2[\nabla_{\mathbf{w}}\mathbf{g}_m]^\top \mathbf{C}_m^{-1}\mathbf{g}_m + \mathbf{g}_m^\top[\nabla_{\mathbf{w}}\mathbf{C}_m^{-1}]\mathbf{g}_m, \tag{43}$$

and

$$\nabla^2_{\mathbf{w}}\widetilde{Q}_m = 2[\nabla_{\mathbf{w}}\mathbf{g}_m]^\top \mathbf{C}_m^{-1}\nabla_{\mathbf{w}}\mathbf{g}_m + \mathbf{R}_m, \tag{44}$$

where $\nabla_{\mathbf{w}}\mathbf{C}_m^{-1} = [\partial\mathbf{C}_m^{-1}/\partial(\mathbf{w})_1, \ldots, \partial\mathbf{C}_m^{-1}/\partial(\mathbf{w})_N]$ is a three dimensional array. And the second term of Eq. (43) is an N-dimensional column vector whose j-th component is given by $\mathbf{g}_m^\top[\partial\mathbf{C}_m^{-1}/\partial(\mathbf{w})_j]\mathbf{g}_m$. The formula of the $N \times N$ matrix $\mathbf{R}_m$ is

$$\mathbf{R}_m = 2\nabla_{\mathbf{w}}[\nabla_{\mathbf{w}}\mathbf{g}_m]^\top \mathbf{C}_m^{-1}\mathbf{g}_m + 4[\nabla_{\mathbf{w}}\mathbf{g}_m]^\top[\nabla_{\mathbf{w}}\mathbf{C}_m^{-1}]\mathbf{g}_m + \mathbf{g}_m^\top[\nabla^2_{\mathbf{w}}\mathbf{C}_m^{-1}]\mathbf{g}_m. \tag{45}$$

By the Central Limit Theorem,

$$\sqrt{m}\mathbf{g}_m(\mathcal{W}^*) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{C}_0) \quad \text{in distribution.} \tag{46}$$

In particular, we have $\mathbf{g}_m(\mathcal{W}^*) = O_p(m^{-1/2})$. Hence $\mathbf{g}_m^\top[\nabla_{\mathbf{w}}\mathbf{C}_m^{-1}]\mathbf{g}_m|_{\mathcal{W}=\mathcal{W}^*} \to o_p(1)$ and $\mathbf{R}_m(\widetilde{\mathcal{W}}_m) \to o_p(1)$. Applying Lemma 1 and Eq. (17) we deduce

$$[\nabla^2_{\mathbf{w}}\widetilde{Q}_m(\mathcal{W}_r)]^{-1} \to \frac{1}{2}(\mathbf{J}_0^\top\mathbf{C}_0^{-1}\mathbf{J}_0)^{-1} \quad \text{in probability,} \tag{47}$$

and

$$\nabla_{\mathbf{w}}\widetilde{Q}_m(\mathcal{W}^*) \to 2\mathbf{J}_0^\top\mathbf{C}_0^{-1}\mathbf{g}_m(\mathcal{W}^*) \quad \text{in probability.} \tag{48}$$

Combining Eq. (42) and Eq. (46)-(48) yields Eq. (18). □

Proof of Lemma 2. Since $\Phi^{(r+1)}$ is a minimizer, by (21), we have

$$F(\Phi) - F(\Phi^{(r+1)}) \geqslant F(\Phi) - P_L(\Phi^{(r+1)}, \Phi^{(r)}). \tag{49}$$

Using the convex property of $Q(\mathcal{W})$ in $\Pi(\Omega)$ and the assumption $\mathcal{W}(\Phi^{(r)}) \in \Pi(\Omega)$ we deduce that for all Φ satisfying $\mathcal{W}(\Phi) \in \Pi(\Omega)$,

$$Q(\mathcal{W}(\Phi)) \geqslant Q(\mathcal{W}^{(r)}) + \langle \sum_{k=1}^{K}\left(\mathcal{W}_k - \mathcal{W}_k^{(r)}\right), \nabla_{\mathcal{W}}Q^{(r)}\rangle. \tag{50}$$

Furthermore, since each part of R is globally convex, we have in general,

$$\sum_{j=1}^{d_k} |||(\mathcal{W}_k)_{(k)}^{(j)}|||_F \geqslant \sum_{j=1}^{d_k} |||(\mathcal{W}_k^{(r+1)})_{(k)}^{(j)}|||_F + \langle \mathcal{W}_k - \mathcal{W}_k^{(r+1)}, \mathcal{A}_k(\mathcal{W}_k^{(r)})\rangle. \tag{51}$$

for all $1 \leqslant k \leqslant K$. Combining (50) and (51) we obtain

$$\begin{aligned} F(\Phi) \geqslant\, & Q(\mathcal{W}^{(r)}) + \langle \sum_{k=1}^{K}\left(\mathcal{W}_k - \mathcal{W}_k^{(r)}\right), \nabla_{\mathcal{W}}Q^{(r)}\rangle + R(\Phi^{(r+1)}) \\ & + \sum_{k=1}^{K}\langle \mathcal{W}_k - \mathcal{W}_k^{(r+1)}, \lambda_k\mathcal{A}_k(\mathcal{W}_k^{(r)})\rangle. \end{aligned} \tag{52}$$

From (52) and the definition of $P_L(\Phi^{(r+1)}, \Phi^{(r)})$ we have

$$F(\Phi) - P_L(\Phi^{(r+1)}, \Phi^{(r)}) \geqslant -\frac{KL}{2}\sum_{k=1}^{K} |||\mathcal{W}_k^{(r+1)} - \mathcal{W}_k^{(r)}|||_F + \sum_{k=1}^{K}\langle \mathcal{W}_k - \mathcal{W}_k^{(r+1)}, \nabla_{\mathcal{W}} Q^{(r)} + \lambda_k \mathcal{A}_k(\mathcal{W}_k^{(r)})\rangle \tag{53}$$

By (27), the second term of (53) on the right hand side can be rewritten as

$$KL\sum_{k=1}^{K}\langle \mathcal{W}_k^{(r+1)} - \mathcal{W}_k, \mathcal{W}_k^{(r+1)} - \mathcal{W}_k^{(r)}\rangle \tag{54}$$

Note that for each $1 \leqslant k \leqslant K$,

$$-\frac{1}{2}|||\mathcal{W}_k^{(r+1)} - \mathcal{W}_k^{(r)}|||_F + \langle \mathcal{W}_k^{(r+1)} - \mathcal{W}_k, \mathcal{W}_k^{(r+1)} - \mathcal{W}_k^{(r)}\rangle = \frac{1}{2}|||\mathcal{W}_k^{(r+1)} - \mathcal{W}_k^{(r)}|||_F + \langle \mathcal{W}_k^{(r)} - \mathcal{W}_k, \mathcal{W}_k^{(r+1)} - \mathcal{W}_k^{(r)}\rangle. \tag{55}$$

The lemma then follows by (49), (53), and (55). □

Proof of Lemma 3. The condition (30) implies $|||\mathcal{W}(\Phi^{(r)}) - \hat{\mathcal{W}}|||_F < \mathrm{dist}(\partial\Pi(\Omega), \hat{\mathcal{W}})$, i.e. $\mathcal{W}(\Phi^{(r)}) \in \Pi(\Omega)$. Since $\hat{\Phi} \in \Omega$ is a global minimizer, applying Lemm 2 with $\Phi = \hat{\Phi}$ we deduce

$$0 \geqslant \sum_{k=1}^{K} |||\mathcal{W}_k^{(r+1)} - \mathcal{W}_k^{(r)}|||_F^2 + 2\sum_{k=1}^{K}\langle \mathcal{W}_k^{(r)} - \hat{\mathcal{W}}_k, \mathcal{W}_k^{(r+1)} - \mathcal{W}_k^{(r)}\rangle. \tag{56}$$

Using Pythagoras relation for each $1 \leqslant k \leqslant K$ we obtain

$$\sum_{k=1}^{K} |||\mathcal{W}_k^{(r+1)} - \hat{\mathcal{W}}_k|||_F^2 = \sum_{k=1}^{K} |||\mathcal{W}_k^{(r)} - \hat{\mathcal{W}}_k|||_F^2 + \sum_{k=1}^{K} |||\mathcal{W}_k^{(r)} - \mathcal{W}_k^{(r+1)}|||_F^2 + 2\sum_{k=1}^{K}\langle \mathcal{W}_k^{(r)} - \hat{\mathcal{W}}_k, \mathcal{W}_k^{(r+1)} - \mathcal{W}_k^{(r)}\rangle \leqslant \sum_{k=1}^{K} |||\mathcal{W}_k^{(r)} - \hat{\mathcal{W}}_k|||_F^2. \tag{57}$$

Here the last inequality comes from (56). Thus $\mathcal{W}(\Phi^{(r+1)})$ satisfies (30). □

Proof of Theorem 2. The condition (30) implies $|||\mathcal{W}(\Phi^{(r)}) - \hat{\mathcal{W}}|||_F < \mathrm{dist}(\partial\Pi(\Omega), \hat{\mathcal{W}})$, i.e. $\mathcal{W}(\Phi^{(r)}) \in \Pi(\Omega)$. Since $\hat{\Phi} \in \Omega$ is a global minimizer, applying Lemm 2 with $\Phi = \hat{\Phi}$ we deduce

$$0 \geqslant \sum_{k=1}^{K} |||\mathcal{W}_k^{(r+1)} - \mathcal{W}_k^{(r)}|||_F^2 + 2\sum_{k=1}^{K}\langle \mathcal{W}_k^{(r)} - \hat{\mathcal{W}}_k, \mathcal{W}_k^{(r+1)} - \mathcal{W}_k^{(r)}\rangle. \tag{58}$$

Using Pythagoras relation for each $1 \leqslant k \leqslant K$ we obtain

$$\begin{aligned}\sum_{k=1}^{K} |||\mathcal{W}_k^{(r+1)} - \hat{\mathcal{W}}_k|||_F^2 &= \sum_{k=1}^{K} |||\mathcal{W}_k^{(r)} - \hat{\mathcal{W}}_k|||_F^2 + \sum_{k=1}^{K} |||\mathcal{W}_k^{(r)} - \mathcal{W}_k^{(r+1)}|||_F^2 \\ &\quad + 2\sum_{k=1}^{K}\langle \mathcal{W}_k^{(r)} - \hat{\mathcal{W}}_k, \mathcal{W}_k^{(r+1)} - \mathcal{W}_k^{(r)}\rangle \\ &\leqslant \sum_{k=1}^{K} |||\mathcal{W}_k^{(r)} - \hat{\mathcal{W}}_k|||_F^2. \end{aligned} \tag{59}$$

Here the last inequality comes from (58). Thus $\mathcal{W}(\Phi^{(r+1)})$ satisfies (30). □

Proof of Theorem 3. For any tensor $\mathcal{W}$ and a set of indies S, we define $(\mathcal{W})_{(k)}^{S}$ by

$$((\mathcal{W})_{(k)}^{S})_{(k)}^{(j)} = \begin{cases} (\mathcal{W})_{(k)}^{(j)} & \text{if } j \in S \\ 0 & \text{otherwise.} \end{cases}$$

Let $\hat{\mathcal{W}}_k$ be a solution of the restricted version of (32):

$$\hat{\mathcal{W}}_k = \arg\min \left\{ \frac{1}{2} \left|\left|\left| (\mathcal{W}_k)_{(k)}^{S_k^*} - (\mathcal{W}_k^*)_{(k)}^{S_k^*} + \mathcal{D}_{(k)}^{S_k^*} \right|\right|\right|_F^2 + \frac{\lambda_k}{KL} \sum_{j \in S} |||(\mathcal{W}_k)_{(k)}^{(j)}|||_F \right\}.$$

Then $(\hat{\mathcal{W}}_k)_{(k)}^{(j)} = 0$ for $j \in S_k^{*c}$. From Lemma 4 and (33), $\hat{\mathcal{W}}_k$ is a solution of (32) and $(\hat{\mathcal{W}}_k)_{(k)}^{(j)}$ satisfies

$$(\hat{\mathcal{W}}_k)_{(k)}^{(j)} - (\mathcal{W}_k^*)_{(k)}^{(j)} + \mathcal{D}_{(k)}^{(j)} = -\frac{\lambda_k}{KL}(\mathcal{A})_{(k)}^{(j)}$$

for $j \in S_k^*$. Here $|||(\mathcal{A})_{(k)}^{(j)}|||_F \leqslant 1$ and

$$(\mathcal{A})_{(k)}^{(j)} = \frac{(\mathcal{W}_k)_{(k)}^{(j)}}{|||(\mathcal{W}_k)_{(k)}^{(j)}|||_F} \quad \text{if } (\mathcal{W}_k)_{(k)}^{(j)} \neq 0.$$

By the triangle inequality we have

$$|||(\hat{\mathcal{W}}_k)_{(k)}^{(j)}|||_F \geqslant \min_{j \in S_k^*} |||(\mathcal{W}_k^*)_{(k)}^{(j)}|||_F - \max_{j \in S_k^*} |||(\mathcal{U})_{(k)}^{(j)}|||_F$$

where

$$(\mathcal{U})_{(k)}^{(j)} = -\mathcal{D}_{(k)}^{(j)} - \frac{\lambda_k}{KL}(\mathcal{A})_{(k)}^{(j)}.$$

Using (33) we deduce

$$\max_{j \in S_k^*} |||(\mathcal{U})_{(k)}^{(j)}|||_F \leqslant \max_{j \in S_k^*} |||\mathcal{D}_{(k)}^{(j)}|||_F + \frac{\lambda_k}{KL} \leqslant \frac{2\lambda_k}{KL}.$$

Thus $|||(\hat{\mathcal{W}}_k)_{(k)}^{(j)}|||_F > 0$ if $\frac{2\lambda_k}{KL} < \min_{j \in S_K^*} |||(\mathcal{W}_k^*)_{(k)}^{(j)}|||_F$. □

Acknowledgements

The authors thank the anonymous reviewers and journal editors who provided valuable comments to improve the quality of this paper.

Funding

This work was partially supported by the National Science Foundation with a grant IIS-1718738 at the U.S., and by the National Institutes of Health with grants R01DA051922, K02DA043063, and R01MH119678 to J Bi.

References

Acar E, Yener B (2009). Unsupervised multiway data analysis: A literature survey. *IEEE Transactions on Knowledge and Data Engineering*, 21(1): 6–20.

Arnold A, Liu Y, Abe N (2007). Temporal causal modeling with graphical granger methods. In: *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD-07*, 66–75. ACM, New York, NY, USA.

Bai Y, Fung WK, Zhu ZY (2009). Penalized quadratic inference functions for single-index models with longitudinal data. *Journal of Multivariate Analysis*, 100(1): 152–161.

Beck T, Teboulle M (2009). A fast iterative shrinkage thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1): 83–202.

Bi J, Sun J, Wu Y, Tennen H, Armeli S (2013). A machine learning approach to college drinking prediction and risk factor identification. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 4(4): 1–24.

Chen CMA, Stanford AD, Mao X, Abi-Dargham A, Shungu DC, Lisanby SH, et al. (2014). Gaba level, gamma oscillation, and working memory performance in schizophrenia. *NeuroImage. Clinical*, 4: 531–539.

Cong F, Lin QH, Kuang LD, Gong XF, Astikainen P, Ristaniemi T (2015). Tensor decomposition of EEG signals: A brief review. *Journal of Neuroscience Methods*, 248: 59–69.

Crowder M (1995). On the use of a working correlation matrix in using generalised linear models for repeated measures. *Biometrika*, 82(2): 407–410.

De Lathauwer L, Vandewalle J (2004). Dimensionality reduction in higher-order signal processing and rank-$(r_1, r_2, \ldots, r_n)$ reduction in multilinear algebra. *Linear Algebra and its Applications*, 391: 31–55.

Diggle P, Heagerty P, Liang KY, Zeger S (2002). *Analysis of Longitudinal Data.* Oxford University Press.

Donoho DL (2000). High-dimensional data analysis: The curses and blessings of dimensionality. *AMS math challenges lecture*, 1(2000): 32.

Fawcett T (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8): 861–874.

Fu WJ (2003). Penalized estimating equations. *Biometrics*, 59(1): 126–132.

Granger C (1980). Testing for causality: A personal viewpoint. *Journal of Economic Dynamics and Control*, 2(1): 329–352.

Hansen LP (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4): 1029–1054.

Herrmann CS, Senkowski D, Rottger S (2004). Phase-locking and amplitude modulations of EEG alpha: Two measures reflect different cognitive processes in a working memory task. *Experimental Psychology*, 51(4): 311.

Hitchcock FL (1927). The expression of a tensor or a polyadic as a sum of products. *Journal of Mathematics and Physics*, 6(1–4): 164–189.

Hoff PD (2015). Multilinear tensor regression for longitudinal relational data. *The Annals of Applied Statistics*, 9(3): 1169.

Johannesen JK, Bi J, Jiang R, Kenney JG, Chen CMA (2016). Machine learning identification of EEG features predicting working memory performance in schizophrenia and healthy adults. *Neuropsychiatric Electrophysiology*, 2(1): 3.

Liang KY, Zeger SL (1986a). Longitudinal data analysis using generalised estimating equations. *Biometrika*, 73(1): 13–22.

Liang KY, Zeger SL (1986b). Longitudinal data-analysis using generalized linear-models. *Biometrika*, 73(1): 13–22.

Liu S, Maljovec D, Wang B, Bremer PT, Pascucci V (2016). Visualizing high-dimensional data: Advances in the past decade. *IEEE transactions on visualization and computer graphics*, 23(3): 1249–1268.

Loader C, Pilla RS (2007). Iteratively reweighted generalized least squares for estimation and testing with correlated data: An inference function framework. *Journal of Computational and Graphical Statistics*, 16(4): 925–945.

Lozano AC, Abe N, Liu Y, Rosset S (2009). Grouped graphical granger modeling methods for temporal causal modeling. In: *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD-09*, 577–586. ACM, New York, NY, USA.

Newey WK, McFadden D (1994). Chapter 36 Large sample estimation and hypothesis testing. *Handbook of Econometrics*, 4: 2111–2245.

Pereira F, Mitchell T, Botvinick M (2009). Machine learning classifiers and fMRI: A tutorial overview. *Neuroimage*, 45(1): S199–S209.

Qu A, Li R (2006). Quadratic inference functions for varying-coefficient models with longitudinal data. *Biometrics*, 62(2): 379–391.

Qu A, Lindsay BG (2003). Building adaptive estimating equations when inverse of covariance estimation is difficult. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1): 127–142.

Qu A, Lindsay BG, Li B (2000). Improving generalised estimating equations using quadratic inference functions. *Biometrika*, 87(4): 823–836.

Sela RJ, Simonoff JS (2012). RE-EM trees: A data mining approach for longitudinal and clustered data. *Machine Learning*, 86(2): 169–207.

Shen L, Thompson PM, Potkin SG, Bertram L, Farrer LA, Foroud TM, et al. (2014). Genetic analysis of quantitative phenotypes in AD and MCI: Imaging, cognition and biomarkers. *Brain Imaging and Behavior*, 8(2): 183–207.

Stappenbeck CA, Fromme K (2010). A longitudinal investigation of heavy drinking and physical dating violence in men and women. *Addictive Behaviors*, 35(5): 479–485.

Tomioka R, Hayashi K, Kashima H (2010). Estimation of low-rank tensors via convex optimization. arXiv preprint: https://arxiv.org/abs/1010.0789.

Tomioka R, Suzuki T (2013). Convex tensor decomposition via structured schatten norm regularization. In: *Advances in Neural Information Processing Systems 26* (C Burges, L Bottou, M Welling, Z Ghahramani, K Weinberger, eds.), 1331–1339.

Tucker LR (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3): 279–311.

Vasilescu MAO, Terzopoulos D (2002). Multilinear analysis of image ensembles: Tensorfaces. In: *European Conference on Computer Vision*, 447–460. Springer.

Wimalawarne K, Tomioka R, Sugiyama M (2016). Theoretical and experimental analyses of tensor-based regression and classification. *Neural Computation*, 28(4): 686–715.

Xu T, Sun J, Bi J (2015). Longitudinal lasso: Jointly learning features and temporal contingency for outcome prediction. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD-15*, 1345–1354. ACM, New York, NY, USA.

Xu Y, Yin W (2017). A globally convergent algorithm for nonconvex optimization based on block coordinate update. *Journal of Scientific Computing*, 72(2): 700–734.

Zhou H, Li L, Zhu H (2013). Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association*, 108(502): 540–552.