An Effective Tensor Regression with Latent Sparse Regularization

Guannan Liang
Department of Computer Science and Engineering, University of Connecticut
Ko-Shin Chen
Department of Computer Science and Engineering, University of Connecticut
Tingyang Xu
Tencent AI Lab, Shenzhen, China
Jun Yan
Department of Statistics, University of Connecticut
Minghu Song
Department of Biomedical Engineering, University of Connecticut
Jinbo Bi
Department of Computer Science and Engineering, University of Connecticut

Abstract

As advances in data acquisition technologies, longitudinal analysis is facing challenges of exploring complex feature patterns from high-dimensional datasets as well as modeling potential temporal correlations and lagged effects. We propose a tensor-based model to analyze multi-dimensional data. It simultaneously discovers patterns in features and reveals past temporal points having impact on current outcomes. The model coefficient, a k-mode tensor, is decomposed into a summation of k tensors of the same dimension. To accomplish feature selection, we introduce the tensor 'latent F-1 norm' as a grouped penalty in our formulation. Meanwhile, the proposed model also takes into account within-subject correlations by involving a tensor-based quadratic inference function. We provide an asymptotic analysis of our model 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 result for a fixed sample size. Computational results on synthetic datasets, real-file fMRI and EEG problems demonstrate the superior performance of the proposed approach over existing techniques.

Keywords: longitudinal, quadratic inference function

1. Introduction

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 [1, 20] to scientific research [1, 15, 24], health-care and medicine [4, 7, 22].

One 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 correct predictions. There are several extended generalized linear models that can be applied to time-dependent data under different assumptions. Diggle et al. have provided a comprehensive overview of various models. For fitting marginal model, generalized estimating equation - GEE [13] and quadratic inference function - QIF [17] are common statistical approaches. They are generally more accurate than those of classic regression analysis that assumes independently and identically distributed (i.i.d.).

In 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 coefficients [5]. In addition, the inverse of the matrix R is essential that may cause poor estimation when R has high dimensions [18]. To overcome these disadvantages, Qu et al. suggested the QIF model 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 [8] and penalized QIF [2] are proposed.

In this work, we study the lagged effect of covariates on outcomes. In many studies, it is necessary and insightful to model simultaneously the correlation among outcomes and the lagged effects of covariates, which is the so-called Granger causality [9]. For example, Shen et al. pointed out evidences of brain diseases may appear in the functional magnetic resonance imaging (fMRI) of an early diagnosis before clear symptoms are identified. Recent graphical Granger models such as [1, 15] ignore the temporal correlations. Xu et al. have modeled such correlation through the

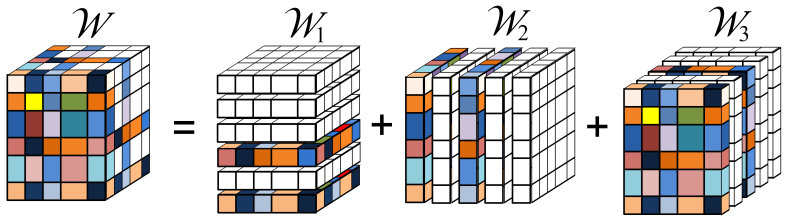

Figure 1: Case for $K = 3$. A 3-way tensor is decomposed into a summation of three 3-way tensors so that each part is sparse along a particular direction.

GEE method. But their model only applies to datasets with one spatial dimension. Our goal is to develop a new penalized QIF method in tensor setting to model the 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 [11, 28]. 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 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 is 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 1. 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 [3, 27] to efficiently solve the optimization problem. This approach then leads to K sub-problems 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.

The rest of this paper is organized as follows. We first briefly review the GEE and QIFs methods, and then introduce our proposed formulation: TensorQIF in the Method section, fol-

lowed by an Asymptotic Analysis section. An optimization algorithm for solving the formulation is depicted in the Algorithm section where we also prove convergence and the recovery of feature support. Experimental results are included and discussed in the Empirical Evaluation section, followed by a Conclusion section.

2. Method

2.1. Notations

We represent a K-way tensor as $\mathcal{A} \in \mathbb{R}^{d1 \times d2 \times ... d_K}$ which contains $N = \prod_{k=1}^{K} d_k$ elements. The inner product of two tensors $\mathcal{A}$ and $\mathcal{B}$ is given by $\langle \mathcal{A}, \mathcal{B} \rangle = \text{vect}(\mathcal{A})^\top \text{vect}(\mathcal{B})$. Here 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. $\mathcal{A}_{(k)}^{(j)} = \mathcal{A}(i_1, i_2, ..., i_k \equiv j, i_{k+1}, ...i_K)$. 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 $[\mathcal{A}_1, \mathcal{A}_2, ..., \mathcal{A}_m]$ 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.

2.2. Generalized Linear Models of a Tensor

Because our model is concerned with tensor regression and classification, 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 $(\mathcal{X}_i, y_i)_{1 \le i \le m}$ be a data set, where $\mathcal{X}_i \in \mathbb{R}^{d_1 \times d_2 \times ... \times d_K}$ 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 $\lambda \ge 0$ 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 [25] or Schatten norm [23] 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 $\mathbf{L_{F,1}}$ norm** defined by

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

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.

2.3. The Proposed QIF Formulation

Let $\mathcal{X}_t^{(i)} \in \mathbb{R}^{d_1 \times d_2 \times \ldots \times d_{K-1}}$ 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:

$$\mathcal{X}_{(i;t)} := [\mathcal{X}_t^{(i)}, \mathcal{X}_{t-1}^{(i)}, \mathcal{X}_{t-2}^{(i)}, ..., \mathcal{X}_{t-\tau}^{(i)}].$$

Suppose there are T total times of measurement for each subject i. In order to have enough previouse 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 $\mathcal{W} \in \mathbb{R}^{d_1 \times d_2 \times ... \times d_{K-1} \times d_K}$, where $d_K = \tau$. We denote $N := \prod_{k=1}^{K} d_k$ the number of elements in $\mathcal{W}$. However, training examples in (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 model 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

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

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

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

Here $\mathbf{R}(\alpha)$ 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 $\text{cov}(\mathbf{y}^{(i)})$ if $\mathbf{R}(\alpha)$ is the true correlation structure for $\mathbf{y}^{(i)}$ [13]. 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 $\boldsymbol{\mu}^{(i)} = (\mu_{\tau+1}^{(i)}, ..., \mu_{\tau+n}^{(i)})^\top$ which depends on $\mathcal{W}$ (see (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 $\mathbf{w} = \text{vect}(\mathcal{W})$ and $\left(\mathbf{D}^{(i)}\right)_{ab} = \partial(\boldsymbol{\mu}^{(i)})_a / \partial(\mathbf{w})_b$.

In a more advanced 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 (7), we formulate our optimization problem via the so-called 'extended score' by substituting (8) for $\mathbf{R}^{-1}(\alpha)$ in (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 $\{\mathbf{s}^{(i)}, \mathcal{X}_{(i)} = (\mathcal{X}_{(i;\tau+1)}, ..., \mathcal{X}_{(i;\tau+n)})\}$.

The vector $\mathbf{g}_m(\mathcal{W})$ in (9) is an $(N \cdot d)$-dimensional column vector. In fact, substituting (8) into (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 [17] and [19], 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 model [10] 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 (2). More precisely, we solve the following optimization problem:

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

where each $\mathcal{W}_k \in \mathbb{R}^{d_1\times d_2\times\ldots\times d_K}$ and the final coefficient tensor

$$\mathcal{W} = \sum_{k=1}^{K} \mathcal{W}_k. \tag{13}$$

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 (12):

$$\widetilde{Q}_m(\mathcal{W}) + \sum_{k=1}^{K}\left(\frac{\lambda_k}{m} \sum_{j=1}^{d_k} |||(\mathcal{W}_k)_{(k)}^{(j)}|||_F\right). \tag{14}$$

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 (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 $\Omega := \Omega_1 \times \Omega_2 \times ... \times \Omega_K$ 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 $\mathcal{W} \in \Omega$, $\|\mathbf{g}(\mathcal{W})\mathbf{g}(\mathcal{W})^\top\|_F \leq d_1(\mathcal{X}, \mathbf{s})$, $\|\|\nabla_{\mathbf{w}}\mathbf{g}(\mathcal{W})\|\|_F \leq d_2(\mathcal{X}, \mathbf{s})$ 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.

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* (14) *subject to* (13)*. Then as $m \to \infty$, we have*

$$\hat{\mathcal{W}}_m \to \mathcal{W}^* \quad \textit{in probability,} \tag{15}$$

$$\sqrt{m} \cdot vect\left(\hat{\mathcal{W}}_m - \mathcal{W}^*\right) \to \mathcal{N}(\mathbf{0}, (\mathbf{J}_0^\top \mathbf{C}_0^{-1} \mathbf{J}_0)^{-1}) \quad \textit{in distribution.} \tag{16}$$

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

The proof is given in Appendix.

4. Algorithm

In this section, we provide an algorithm to solve the optimization problem (12) followed by a convergence result. Since the sample size m is fixed throughout this section, we drop the subscript m in (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 (12). Denote the iterates at the r-th iteration by $\Phi^{(r)}$. At the point $\Phi = (\mathcal{W}_1, \cdots, \mathcal{W}_K)$, let

$$R(\Phi) := \sum_{k=1}^{K} \left(\lambda_k \sum_{j=1}^{d_k} |||(\mathcal{W}_k)_{(k)}^{(j)}|||_F \right). \tag{17}$$

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:

$$P_L(\Phi, \widetilde{\Phi}) := Q(\widetilde{\mathcal{W}}) + R(\Phi) + \frac{KL}{2} \sum_{k=1}^{K} |||\mathcal{W}_k - \widetilde{\mathcal{W}}_k|||_F^2 + \langle \sum_{k=1}^{K} \left(\mathcal{W}_k - \widetilde{\mathcal{W}}_k \right), \nabla_{\mathcal{W}} Q(\widetilde{\mathcal{W}}) \rangle \tag{18}$$

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

$$\frac{L}{2} |||\mathcal{W} - \widetilde{\mathcal{W}}|||_F^2 \leq \frac{KL}{2} \sum_{k=1}^{K} |||\mathcal{W}_k - \widetilde{\mathcal{W}}_k|||_F^2. \tag{19}$$

The inequality (19) and the Lipschitz continuity of $Q(\mathcal{W})$ indicate that for all $L \geq L_Q$,

$$F(\Phi) \leq P_L(\Phi, \widetilde{\Phi}) \quad \text{for all } \Phi \text{ and } \widetilde{\Phi}. \tag{20}$$

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

$$\min_{\Phi} \sum_{k=1}^{K} \left[\langle \nabla_{\mathcal{W}} Q^{(r)}, \mathcal{W}_k - \mathcal{W}_k^{(r)} \rangle + \frac{KL}{2} |||\mathcal{W}_k - \mathcal{W}_k^{(r)}|||_F^2 \right] + R(\Phi) \tag{21}$$

where $\nabla_{\mathcal{W}} Q^{(r)} = \nabla_{\mathcal{W}} Q(\mathcal{W}^{(r)})$. Since $R(\Phi)$ given in (17) is separable among $\mathcal{W}_k$'s, we can decompose the problem (21) into the following K separate subproblems:

$$\min_{\mathcal{W}_k} \left\{ \langle \nabla_{\mathcal{W}} Q^{(r)}, \mathcal{W}_k - \mathcal{W}_k^{(r)} \rangle + \frac{KL}{2} |||\mathcal{W}_k - \mathcal{W}_k^{(r)}|||_F^2 + \lambda_k \sum_{j=1}^{d_k} |||(\mathcal{W}_k)_{(k)}^{(j)}|||_F \right\} \tag{22}$$

for $k \in \{1, \ldots, K\}$. Since the subproblems share the same structure, we may fix k and solve (22) to find the best $\mathcal{W}_k$, which is equivalent to

$$\min_{\mathcal{W}_k} \frac{1}{2} \left\| \mathcal{W}_k - \left(\mathcal{W}_k^{(r)} - \frac{1}{KL} \nabla_{\mathcal{W}} Q^{(r)} \right) \right\|_F^2 + \frac{\lambda_k}{KL} \sum_{j=1}^{d_k} |||(\mathcal{W}_k)_{(k)}^{(j)}|||_F. \tag{23}$$

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

Input: $\mathcal{X}$, $\mathbf{y}$, L, λ_k

Output: $\hat{\Phi} = (\hat{\mathcal{W}}_1, \cdots, \hat{\mathcal{W}}_K)$

1. $r = 0$: compute $\tilde{L}$ and initialize $\mathcal{W}_k^{(0)}$ for $1 \leq k \leq K$.
2. Obtain $\Phi^{(r+1)} = (\mathcal{W}_1^{(r+1)}, \cdots, \mathcal{W}_K^{(r+1)})$ by solving (23) for each fixed $1 \leq k \leq K$.
3. $r = r + 1$.

Repeat 2 and 3 until convergence.

The problem (23) has a closed-form solution $\mathcal{W}_k^{(r+1)}$ where each of its sub-tensor is

$$(\mathcal{W}_k^{(r+1)})_{(k)}^{(j)} = \max\left(0, 1 - \frac{\lambda_k}{KL|||(\mathcal{P}^{(r)})_{(k)}^{(j)}|||_F}\right)(\mathcal{P}^{(r)})_{(k)}^{(j)}, \tag{24}$$

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

$$\nabla_{\mathcal{W}} Q^{(r)} + KL\left(\mathcal{W}_k^{(r+1)} - \mathcal{W}_k^{(r)}\right) + \lambda_k \mathcal{A}_k(\mathcal{W}_k^{(r)}) = 0 \tag{25}$$

for all $r \geq 1$ and $1 \leq k \leq K$. 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 [26] to find a proper approximation $\tilde{L} \geq L_Q$ and use $\tilde{L}$ as L in all of our computations. Algorithm 1 summarizes the steps for finding the optimal $\hat{\mathcal{W}}_k$.

4.2. Convergence Analysis

In this section, we prove that the sequence $\{\Phi^{(r)}\}_{r\geq 0}$ 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 [14], it has been shown that the function $Q(\mathcal{W})$ is not globally convex in general. Hence the standard convergence arguments such as in [3] 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 $\hat{\Phi} = (\hat{\mathcal{W}}_1, \ldots, \hat{\mathcal{W}}_K)$ be a global minimizer of $F(\Phi)$ and $\Omega = \Omega_1 \times \ldots \Omega_K$ 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

$$D(\Phi^{(0)}) := \sum_{k=1}^{K} \|\mathcal{W}_k^{(0)} - \hat{\mathcal{W}}_k\|_F^2 < \frac{1}{K}\left[\text{dist}(\partial\Pi(\Omega), \hat{\mathcal{W}})\right]^2. \tag{26}$$

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 $n \geq 1$,*

$$F(\Phi^{(n)}) - F(\hat{\Phi}) \leq \frac{KL\sum_{k=1}^{K} \|\mathcal{W}_k^{(0)} - \hat{\mathcal{W}}_k\|_F^2}{2n}. \tag{27}$$

To prove the theorem, we first show that if $\Phi^{(r)}$ satisfies (26) at the r-th iteration, then $\Phi^{(r+1)}$ also satisfies (26). This ensures that the entire sequence $\{\mathcal{W}(\Phi^{(n)})\}_{n\geq 0}$ generated by Algorithm 1 lies in $\Pi(\Omega)$ in which the function Q is convex. Thus the convex inequality is always valid and Theorem 2 is established. Details are provided in Appendix.

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

$$\eta_t^{(i)} = \langle \mathcal{X}_{(i;t)}, \sum_{k=1}^{K} \mathcal{W}_k \rangle$$

and the components of outcome $\mathbf{y}^{(i)}$ are of the form

$$y_t^{(i)} = \langle \mathcal{X}_t^{(i)}, \mathcal{W}^* \rangle + s_t^{(i)}$$

for some true tensor coefficient $\mathcal{W}^* = \sum_{k=1}^{K} \mathcal{W}_k^*$, where $\tau \leq t \leq T$, 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:

$$\min_{\mathcal{W}_k} \frac{1}{2}\|\mathcal{W}_k - \mathcal{W}_k^* + \mathcal{D}\|_F^2 + \frac{\lambda_k}{KL}\sum_{j=1}^{d_k} \|(\mathcal{W}_k)_{(k)}^{(j)}\|_F. \tag{28}$$

Our goal is to estimate the group support for $\mathcal{W}_k^*$, i.e. obtain the subset $S_k^* \subset \{1, 2, ..., d_k\}$ 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 (28) immediately imply the following lemma.

Lemma 1 (KKT). *Assume $\hat{\mathcal{W}}_k$ is a solution of* (28). *Then either*

$$(\hat{\mathcal{W}}_k)^{(j)}_{(k)} \neq 0 \quad \textit{and}$$

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

or

$$(\hat{\mathcal{W}}_k)^{(j)}_{(k)} = 0 \quad \textit{and}$$

$$|||(\mathcal{W}^*_k)^{(j)}_{(k)} + \mathcal{D}^{(j)}_{(k)}|||_F \leq \frac{\lambda_k}{KL}.$$

Lemma 1 then yield

Theorem 3. *Assume*

$$\frac{\lambda_k}{2} \geq \max_{1\leq j\leq d_k} |||\mathcal{D}^{(j)}_{(k)}|||_F. \tag{29}$$

Then (28) *has a solution $\hat{\mathcal{W}}_k$ such that*

$$\{j : (\hat{\mathcal{W}}_k)^{(j)}_{(k)} \neq 0\} := \hat{S}_k \subset S_k. \tag{30}$$

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

The proof is given in Appendix.

5. Empirical Evaluation

In this section we present the results of both synthetic and real-life fMRI 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 2.3: $\{y^{(i)}_t, \mathcal{X}^{(i)}_t : 1 \leq i \leq m, 0 \leq t \leq T\}$. Here i denotes the subject id and t is a time point. For both synthetic and fMRI cases, each $\mathcal{X}^{(i)}_t$ is a matrix (i.e. a 2-way tensor).

5.1. Simulation

We examine the following methods: *TensorQIF*, Least Absolute Shrinkage and Selection Operator (LASSO), Graphical Granger Modeling ([15]), GEE ([13]), and Kruskal ([28]). 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 2.3 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 models.

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 from the normal distribution $N(0,1)$ plus the uniform distribution $U(0, \sin(t))$. The number of time point is 10 and after concatenating the current and previous $\tau = 2$ records, we obtain $\mathcal{X}_{(i;t)}$ for $\tau + 1 \le t \le 10$. 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 $\mathcal{W} = \mathcal{W}_1 + \mathcal{W}_2 + \mathcal{W}_3$. And 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 contains 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 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 square error (MSE) between the observed y and the predictive $\hat{y} = \langle \mathcal{X}, \hat{\mathcal{W}} \rangle$; 2. The root mean square error (RMSE) between true $\bar{y} = \langle \mathcal{X}, \mathcal{W} \rangle$ and $\hat{y}$.

Since the Kruskal model focuses on the low rank decomposition for $\mathcal{W}$, we conduct the sim-

ulation 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/true y RMSE on the test set over 100 replications with different models and settings are summarized in Table 5.1.

In Table 5.1, 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 $\tau \geq 2$, 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.

The simulation also confirms that with the correct correlation structure (AR1), the fitting results of GEE (tensor input) and TensorQIF are more accurate. When both models are under miss specified correlation structure (Id), Table 5.1 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 (Id), TensorQIF (AR1), and TensorQIF (Id). The p values are given in Table 3. We see that for the larger coefficient size, the MSE differences between these settings are more significant.

Figure 2 shows an example of TensorQIF (AR1) fitting result on one dataset with $(d_1, d_2, \tau + 1) = (3, 3, 3)$, 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

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

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 $\hat{y}$

	LASSO	Granger	Kruskal (rk2)		Kruskal (rk3)	
	matrix	tensor	matrix	tensor	matrix	tensor
$2 \times 2 \times 3$	8.1197	3.8862	8.1188	3.8916	8.1188	3.8853
$3 \times 3 \times 3$	9.9936	3.9744	10.025	5.3318	9.9930	4.3665
$5 \times 5 \times 5$	27.559	4.0476	27.647	5.1437	27.595	4.3695

	GEE (AR1)		GEE (Id)		TensorQIF (AR1)	TensorQIF (Id)
	matrix	tensor	matrix	tensor	tensor	tensor
$2 \times 2 \times 3$	8.1656	3.8802	8.1188	3.8853	**3.8707**	3.8842
$3 \times 3 \times 3$	10.113	3.9726	9.9930	3.9825	**3.9698**	3.9788
$5 \times 5 \times 5$	27.580	4.0326	27.578	4.0781	**4.0118**	4.0347

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

	Granger	Kruskal (rk2)	Kruskal (rk3)
$2 \times 2 \times 3$	0.0898	0.1158	0.0886
$3 \times 3 \times 3$	0.1253	1.2310	0.6107
$5 \times 5 \times 5$	0.2534	1.0764	0.6129

	GEE (AR1)	GEE (Id)	TensorQIF (AR1)	TensorQIF (Id)
$2 \times 2 \times 3$	0.0582	0.0882	**0.0544**	0.0849
$3 \times 3 \times 3$	0.0808	0.1282	**0.0773**	0.1193
$5 \times 5 \times 5$	0.2144	0.3006	**0.1958**	0.2379

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

	Granger	GEE (Id)	TensorQIF (Id)
$2 \times 2 \times 3$	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 (Id) v.s. GEE (Id)
$2 \times 2 \times 3$	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

Initiative (ADNI)[1]. We cleaned up the fMRI data by filtering out the incomplete or low quality observations. After data cleaning, the data includes 147 subjects diagnosed with mild cognitive impairment (MCI) from the year of 2009 to 2016. We use the participants' first fMRI scan as baseline and the other fMRI scans in 6th, 12th, 18th, and 24th months of the study. Here are 67 brain areas and 4 properties (CV,SA,TA,TS) of the brain cortex[2] in our model. These properties are **CV**: Cortical Volume; **SA**: Surface Area; **TA**: Thickness Average; **TS**: Thickness Standard Deviation. This record naturally form a 3-way tensor with one dimension for brain areas, one for property, and one along the temporal line. Our *TensorQIF* keeps such tensor form without squashing dataset into a vector which may cause losing the proximity. The outcome used in this experiment is 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' answers of the questionnaire.

We use 20% of subjects for testing. The lag variable is set to $\tau = 2$. The λ_1, λ_2, and λ_3 were

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

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

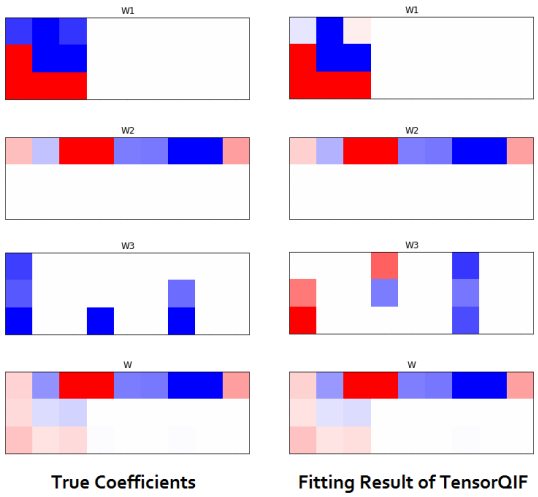

Figure 2: True coefficients and a TensorQIF fitting result.

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.

Our approach is able to select patterns along three dimensions: among the features, among the brain areas, and among the different lagged months. The λ's were chosen as $\lambda_1 = 6$, $\lambda_2 = 20$, and $\lambda_3 = 24$. In Figure 3, the structural damage of AD starting 6 months ago plays a major role in the development of the AD. Larger means and standard derivations of the thickness imply a higher risk of the AD. The proposed model selects 14 out of 68 brain areas that affect the MMSE score. According to the selections of the brain areas, the data at Cuneus area and Transverse Temporal area in both sides, and the data at right Inferior Parietal area, and so on might be important to predict the cognitive impairment.

5.3. EEG Data

Human memory function can be assayed in real-time by electroencephalographic (EEG) recording. However, the clinical utility of this method depends on the reliable determination of functionally and diagnostically relevant features. The proposed method approaches capable

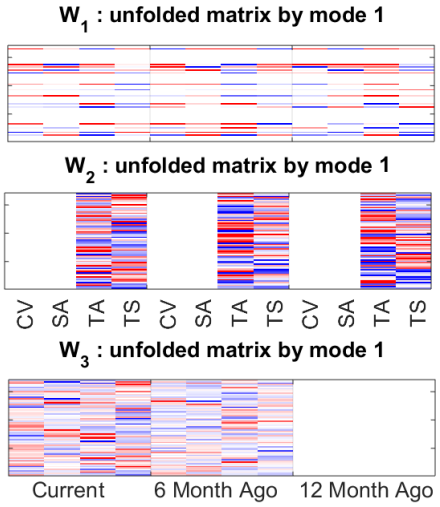

Figure 3: Columns, rows, and slices selected by the model for predicting MMSE score from participants' fMRI information.

of modeling non-stationary signal have 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.

Schizophrenia (SZ, $n = 40$) patients and healthy control (HC, $n = 20$) 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 HC samples for comparison of common and disparate feature patterns across the dimensions.

For each of the SZ and HC datasets, 1/5 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 hyperpa-

rameters 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 HC.

As shown in Figure 4, 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 4; blue border). This indicates the same findings as in [26]. Groups differed in two main ways: (1) centroparietal theta, beta, and gamma during encoding and retention have lower values in HC (Figure 4; red border), and (2) the delta activity across stages and electrodes (Figure 4; green border) was selected in SZ but no in HC. Here the experimental results give much clearer details of the working electrode sites and spectral frequencies comparing to the results in [12]. The proposed method outperform GEE and SVM solutions according to AUC values (HC: 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 called *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 tensor coefficient is computed by the summation of K component tensors so that each reflects the selection among a particular mode. Asymptotic analysis shows the proposed formulation finds true coefficient 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 dataset provide insightful discoveries.

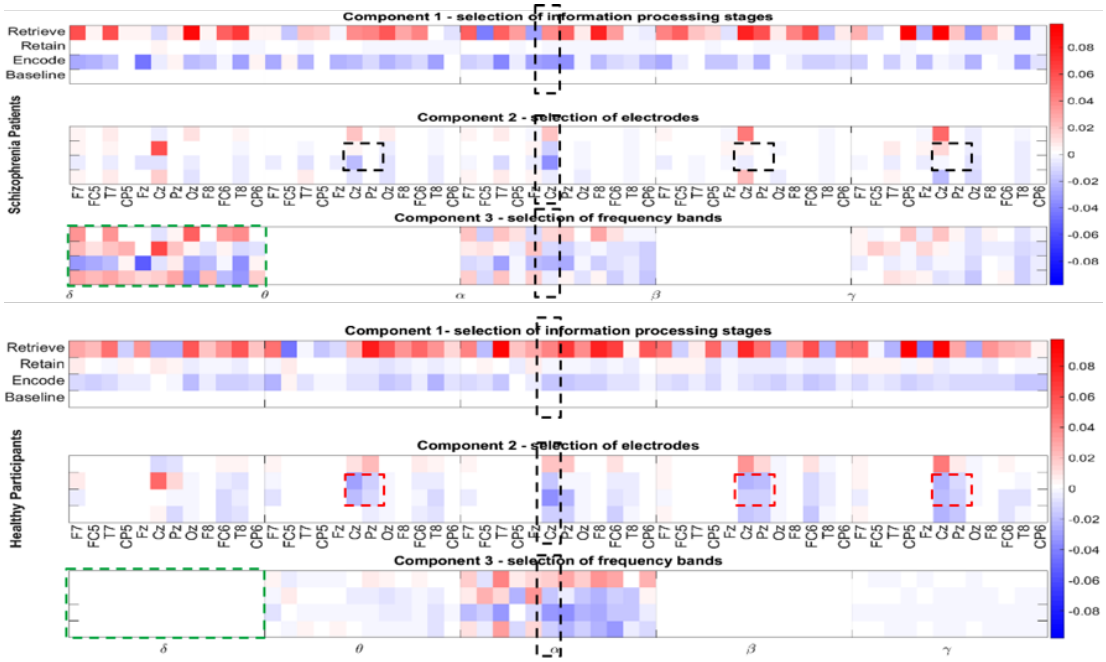

Figure 4: Columns, rows, and slices selected by the model to predict the succeeds of the memory tasks for SZ (top) and HN (bottom), respectively.

7. Appendix

7.1. Proof of Theorem 1

The proof the theorem is based on a uniform convergence result for stochastic functions. Using Lemma 2.4 in [16], conditions 2, and 4, we obtain

Lemma 2. *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*

$$\mathbf{C}_m(\mathcal{W}) \to \mathbf{C}_*(\mathcal{W}) \quad \textit{in probability} \tag{31}$$

and

$$\nabla_{\mathbf{w}}\mathbf{g}_m(\mathcal{W}) \to \mathbf{J}_*(\mathcal{W}) \quad \textit{in probability}, \tag{32}$$

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* (32) *then yield*

$$\mathbf{g}_m(\mathcal{W}) \to \mathbb{E}[\mathbf{g}(\mathcal{W})] \quad \textit{in probability} \tag{33}$$

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

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

$$\widetilde{Q}_m(\hat{\mathcal{W}}_m) + \sum_{k=1}^{K} \left(\frac{\lambda_k}{m} \sum_{j=1}^{d_k} \|\!| (\hat{\mathcal{W}}_{k;m})_{(k)}^{(j)} |\!\| _F \right) \leq \widetilde{Q}_m(\mathcal{W}^*) + \sum_{k=1}^{K} \left(\frac{\lambda_k}{m} \sum_{j=1}^{d_k} \|\!| (\mathcal{W}_k^*)_{(k)}^{(j)} |\!\| _F \right). \tag{34}$$

Note that

$$\begin{aligned} |\widetilde{Q}_m(\mathcal{W}^*)| &= \left| \mathbf{g}_m^\top(\mathcal{W}^*) \mathbf{C}_m^{-1}(\mathcal{W}^*) \mathbf{g}_m(\mathcal{W}^*) \right| \\ &\leq \left| \mathbf{g}_m^\top(\mathcal{W}^*) [\mathbf{C}_m^{-1}(\mathcal{W}^*) - \mathbf{C}_*^{-1}(\mathcal{W}^*)] \mathbf{g}_m(\mathcal{W}^*) \right| + \left| \mathbf{g}_m^\top(\mathcal{W}^*) \mathbf{C}_*^{-1}(\mathcal{W}^*) \mathbf{g}_m(\mathcal{W}^*) \right|. \end{aligned} \tag{35}$$

By (31), condition 1, and the weak law of large numbers, we deduce

$$\left| \widetilde{Q}_m(\mathcal{W}^*) \right| \to 0 \quad \text{in probability.} \tag{36}$$

Therefore from (34), we obtain

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

for fixed λ_k's. Using (31) and (33) we also have

$$|\widetilde{Q}_m(\hat{\mathcal{W}}_m) - \mathbb{E}[\mathbf{g}(\hat{\mathcal{W}}_m)]^\top \mathbf{C}_*(\hat{\mathcal{W}}_m) \mathbb{E}[\mathbf{g}(\hat{\mathcal{W}}_m)]| \to 0 \quad \text{in probability.} \tag{38}$$

Thus $\mathbb{E}[\mathbf{g}(\hat{\mathcal{W}}_m)] \to 0$ and (15) 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:

$$\nabla_{\mathbf{w}} \widetilde{Q}_m(\hat{\mathcal{W}}_m) + o(1) = 0. \tag{39}$$

Using the mean value theorem we obtain

$$\nabla_{\mathbf{w}} \widetilde{Q}_m(\mathcal{W}^*) + \nabla_{\mathbf{w}}^2 \widetilde{Q}_m(\widetilde{\mathcal{W}}_m) \text{vect}(\hat{\mathcal{W}}_m - \mathcal{W}^*) = o(1) \tag{40}$$

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

$$\sqrt{m} \cdot \text{vect}(\hat{\mathcal{W}}_m - \mathcal{W}^*) = -\sqrt{m}[\nabla^2_{\mathbf{w}} \widetilde{Q}_m(\widetilde{\mathcal{W}}_m)]^{-1}[\nabla_{\mathbf{w}} \widetilde{Q}_m(\mathcal{W}^*) + o(1)]. \tag{41}$$

A direct calculation shows

$$\nabla_{\mathbf{w}} \widetilde{Q}_m = 2[\nabla_{\mathbf{w}} \mathbf{g}_m]^\top \mathbf{C}_m^{-1} \mathbf{g}_m + \mathbf{g}_m^\top [\nabla_{\mathbf{w}} \mathbf{C}_m^{-1}] \mathbf{g}_m, \tag{42}$$

and

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

Here $\nabla_{\mathbf{w}} \mathbf{C}_m^{-1} = [\partial \mathbf{C}_m^{-1}/\partial(\mathbf{w})_1, ..., \partial \mathbf{C}_m^{-1}/\partial(\mathbf{w})_N]$ is a three dimensional array. And the second term of (42) 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

$$\mathbf{R}_m = 2\nabla_{\mathbf{w}} [\nabla_{\mathbf{w}} \mathbf{g}_m]^\top \mathbf{C}_m^{-1} \mathbf{g}_m + 4[\nabla_{\mathbf{w}} \mathbf{g}_m]^\top [\nabla_{\mathbf{w}} \mathbf{C}_m^{-1}] \mathbf{g}_m + \mathbf{g}_m^\top [\nabla^2_{\mathbf{w}} \mathbf{C}_m^{-1}] \mathbf{g}_m. \tag{44}$$

By the Central Limit Theorem,

$$\sqrt{m} \mathbf{g}_m(\mathcal{W}^*) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{C}_0) \quad \text{in distribution.} \tag{45}$$

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 2 and (15) we deduce

$$[\nabla^2_{\mathbf{w}} \widetilde{Q}_m(\mathcal{W}_r)]^{-1} \to \frac{1}{2}(\mathbf{J}_0^\top \mathbf{C}_0^{-1} \mathbf{J}_0)^{-1} \quad \text{in probability,} \tag{46}$$

and

$$\nabla_{\mathbf{w}} \widetilde{Q}_m(\mathcal{W}^*) \to 2\mathbf{J}_0^\top \mathbf{C}_0^{-1} \mathbf{g}_m(\mathcal{W}^*) \quad \text{in probability.} \tag{47}$$

Combining (41) and (45)-(47) yields (16). □

7.2. Proof of Theorem 2

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

Lemma 3. *Assume* $\Phi^{(r)} = (\mathcal{W}_1^{(r)}, \cdots, \mathcal{W}_K^{(r)})$ *such that* $\mathcal{W}(\Phi^{(r)}) \in \Pi(\Omega)$. *Let* $\Phi^{(r+1)} = (\mathcal{W}_1^{(r+1)}, \ldots, \mathcal{W}_K^{(r+1)})$ *be a minimizer of* (23). *Then for any* $L \geq L_Q$ *and for any* $\Phi = (\mathcal{W}_1, \ldots \mathcal{W}_K)$ *such that* $\mathcal{W}(\Phi) \in \Pi(\Omega)$, *we have*

$$F(\Phi) - F(\Phi^{(r+1)}) \geq \frac{KL}{2} \sum_{k=1}^K \|\mathcal{W}_k^{(r+1)} - \mathcal{W}_k^{(r)}\|_F^2 + KL \sum_{k=1}^K \langle \mathcal{W}_k^{(r)} - \mathcal{W}_k, \mathcal{W}_k^{(r+1)} - \mathcal{W}_k^{(r)} \rangle. \tag{48}$$

Proof. Since $\Phi^{(r+1)}$ is a minimizer, by (19), we have

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

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 $\mathcal{W}(\Phi) \in \Pi(\Omega)$,

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

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

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

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

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

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

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

By (25), 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 $1 \leq k \leq K$,

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

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

Lemma 4. *Let $\hat{\mathcal{W}} = \mathcal{W}(\hat{\Phi})$. Suppose $\Phi^{(r)}$ satisfy*

$$D(\Phi) := \sum_{k=1}^{K} \|\mathcal{W}_k - \hat{\mathcal{W}}_k\|_F^2 < \frac{1}{K}\left[dist(\partial\Pi(\Omega), \hat{\mathcal{W}})\right]^2. \tag{56}$$

Then $\mathcal{W}(\Phi^{(r+1)})$ generated by (23) *also satisfies* (56).

Proof. The condition (56) implies $\|\mathcal{W}(\Phi^{(r)}) - \hat{\mathcal{W}}\|_F < \text{dist}(\partial\Pi(\Omega), \hat{\mathcal{W}})$, i.e. $\mathcal{W}(\Phi^{(r)}) \in \Pi(\Omega)$. Since $\hat{\Phi} \in \Omega$ is a global minimizer, applying Lemm 3 with $\Phi = \hat{\Phi}$ we deduce

$$0 \geq \sum_{k=1}^{K} \|\mathcal{W}_k^{(r+1)} - \mathcal{W}_k^{(r)}\|_F^2 + 2\sum_{k=1}^{K} \langle \mathcal{W}_k^{(r)} - \hat{\mathcal{W}}_k, \mathcal{W}_k^{(r+1)} - \mathcal{W}_k^{(r)} \rangle. \tag{57}$$

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

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

Here the last inequality comes from (57). Thus $\mathcal{W}(\Phi^{(r+1)})$ satisfies (56). □

Remark 2. *Lemma 4 implies that all points in the sequence $\{\Phi^{(r)}\}_{r\geq 0}$ generated by Algorithm 1 satisfy* (56) *if the initial point $\Phi^{(0)}$ does. In particular, we have $\{\mathcal{W}(\Phi^{(r)})\}_{r\geq 0} \subset \Pi(\Omega)$. Thus we can apply Lemma 3 for all $r \geq 0$.*

Proof of the Theorem. From Remark 2, we apply Lemma 3 with $\Phi = \hat{\Phi}$ for all $0 \leq r \leq n-1$:

$$\begin{aligned}
\frac{2}{KL}\left[F(\hat{\Phi}) - F(\Phi^{(r+1)})\right] \geq& \sum_{k=1}^{K} \|\mathcal{W}_k^{(r+1)} - \mathcal{W}_k^{(r)}\|_F^2 + 2\sum_{k=1}^{K} \langle \mathcal{W}_k^{(r)} - \hat{\mathcal{W}}_k, \mathcal{W}_k^{(r+1)} - \mathcal{W}_k^{(r)} \rangle \\
=& \sum_{k=1}^{K} \|\hat{\mathcal{W}}_k - \mathcal{W}_k^{(r+1)}\|_F^2 - \sum_{k=1}^{K} \|\hat{\mathcal{W}}_k - \mathcal{W}_k^{(r)}\|_F^2.
\end{aligned} \tag{59}$$

Summing (59) over r we have

$$\frac{2}{KL}\left[nF(\hat{\Phi}) - \sum_{r=0}^{n-1} F(\Phi^{(r)})\right] \geq \sum_{k=1}^{K} \|\hat{\mathcal{W}}_k - \mathcal{W}_k^{(n)}\|_F^2 - \sum_{k=1}^{K} \|\hat{\mathcal{W}}_k - \mathcal{W}_k^{(0)}\|_F^2. \tag{60}$$

Using Lemma 3 again with $\Phi = \Phi^{(r)}$ we have for all $0 \leq r \leq n-1$,

$$\frac{2r}{KL}\left[F(\Phi^{(r)}) - F(\Phi^{(r+1)})\right] \geq r\sum_{k=1}^{K} |||\mathcal{W}_k^{(r+1)} - \mathcal{W}_k^{(r)}|||_F^2 \tag{61}$$

Summing (61) over r we obtain

$$\frac{2}{KL}\left[-nF(\Phi^{(n)}) + \sum_{r=0}^{n-1} F(\Phi^{(r+1)})\right] \geq \sum_{r=0}^{n-1} r\sum_{k=1}^{K} |||\mathcal{W}_k^{(r+1)} - \mathcal{W}_k^{(r)}|||_F^2. \tag{62}$$

Combining (60) and (62) yields (27). □

Remark 3. *Lemma 3 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 4, we deduce that the minimizer of* (23) *generated by $\widetilde{\Phi}^{(r)}$ will satisfy* (56) *if $\widetilde{\Phi}^{(r)}$ does.*

7.3. 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 (28):

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

Then $(\hat{\mathcal{W}}_k)_{(k)}^{(j)} = 0$ for $j \in S_k^{*c}$. From Lemma 1 and (29), $\hat{\mathcal{W}}_k$ is a solution of (28) 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 $|||(\mathcal{A})_{(k)}^{(j)}|||_F \leq 1$ 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

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

where

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

Using (29) we deduce

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

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

References

[1] Andrew Arnold, Yan Liu, and Naoki Abe. Temporal causal modeling with graphical granger methods. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '07, pages 66–75, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-609-7. doi: 10.1145/1281192.1281203. URL `http://doi.acm.org/10.1145/1281192.1281203`.

[2] Yang Bai, Wing K. Fung, and Zhong Yi Zhu. Penalized quadratic inference functions for single-index models with longitudinal data. *Journal of Multivariate Analysis*, 100(1):152 – 161, 2009. ISSN 0047-259X. doi: http://dx.doi.org/10.1016/j.jmva.2008.04.004. URL `http://www.sciencedirect.com/science/article/pii/S0047259X08001115`.

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

[4] Jinbo Bi, Jiangwen Sun, Yu Wu, Howard Tennen, and Stephen Armeli. A machine learning approach to college drinking prediction and risk factor identification. *ACM Trans. Intell. Syst. Technol.*, 4(4):72:1–72:24, October 2013. ISSN 2157-6904. doi: 10.1145/2508037.2508053. URL `http://doi.acm.org/10.1145/2508037.2508053`.

[5] Martin Crowder. On the use of a working correlation matrix in using generalised linear models for repeated measures. *Biometrika*, 82(2):407–410, 1995. doi: 10.1093/biomet/82.2.407. URL `http://dx.doi.org/10.1093/biomet/82.2.407`.

[6] Peter Diggle, Patrick Heagerty, Kung-Yee Liang, and Scott Zeger. *Analysis of Longitudinal Data*. Oxford University Press, 2002.

[7] James H Fowler and Nicholas A Christakis. Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the framingham heart study. *BMJ*, 337, 2008. ISSN 0959-8138. doi: 10.1136/bmj.a2338. URL http://www.bmj.com/content/337/bmj.a2338.

[8] Wenjiang J. Fu. Penalized estimating equations. *Biometrics*, 59(1):126–132, 2003. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/3695820.

[9] Clive Granger. Testing for causality: A personal viewpoint. *Journal of Economic Dynamics and Control*, 2(1):329–352, 1980. URL http://EconPapers.repec.org/RePEc:eee:dyncon:v:2:y:1980:i:1:p:329-352.

[10] Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054, 1982. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1912775.

[11] Peter D. Hoff. Multilinear tensor regression for longitudinal relational data. *Ann. Appl. Stat.*, 9(3):1169–1193, 09 2015. doi: 10.1214/15-AOAS839. URL http://dx.doi.org/10.1214/15-AOAS839.

[12] Jason K. Johannesen, Jinbo Bi, Ruhua Jiang, Joshua G. Kenney, and Chi-Ming A. Chen. Machine learning identification of eeg features predicting working memory performance in schizophrenia and healthy adults. *Neuropsychiatric Electrophysiology*, 2(1):3, Feb 2016. ISSN 2055-4788. doi: 10.1186/s40810-016-0017-0. URL https://doi.org/10.1186/s40810-016-0017-0.

[13] Kung-Yee Liang and Scott L. Zeger. Longitudinal data analysis using generalised estimating equations. *Biometrika*, 73(1):13–22, 1986.

[14] Catherine Loader and Ramani S Pilla. Iteratively reweighted generalized least squares for estimation and testing with correlated data: An inference function framework. *Jour-*

nal of Computational and Graphical Statistics, 16(4):925–945, 2007. doi: 10.1198/106186007X238828. URL http://dx.doi.org/10.1198/106186007X238828.

[15] Aurelie C. Lozano, Naoki Abe, Yan Liu, and Saharon Rosset. 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, pages 577–586, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-495-9. doi: 10.1145/1557019.1557085. URL http://doi.acm.org/10.1145/1557019.1557085.

[16] Whitney K. Newey and Daniel McFadden. Chapter 36 large sample estimation and hypothesis testing. *Handbook of Econometrics*, 4:2111 – 2245, 1994. ISSN 1573-4412. doi: http://dx.doi.org/10.1016/S1573-4412(05)80005-4. URL http://www.sciencedirect.com/science/article/pii/S1573441205800054.

[17] Annie Qu and Runze Li. Quadratic inference functions for varying-coefficient models with longitudinal data. *Biometrics*, 62(2):379–391, 2006. ISSN 1541-0420. doi: 10.1111/j.1541-0420.2005.00490.x. URL http://dx.doi.org/10.1111/j.1541-0420.2005.00490.x.

[18] Annie Qu and Bruce G. Lindsay. 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, 2003. ISSN 1467-9868. doi: 10.1111/1467-9868.00376. URL http://dx.doi.org/10.1111/1467-9868.00376.

[19] Annie Qu, Bruce G. Lindsay, and Bing Li. Improving generalised estimating equations using quadratic inference functions. *Biometrika*, 87(4):823–836, 2000. doi: 10.1093/biomet/87.4.823. URL +http://dx.doi.org/10.1093/biomet/87.4.823.

[20] Rebecca J. Sela and Jeffrey S. Simonoff. Re-em trees: a data mining approach for longitudinal and clustered data. *Machine Learning*, 86(2):169–207, Feb 2012. ISSN 1573-0565. doi: 10.1007/s10994-011-5258-3. URL https://doi.org/10.1007/s10994-011-5258-3.

[21] Li Shen, Paul M. Thompson, Steven G. Potkin, Lars Bertram, Lindsay A. Farrer, Tatiana M. Foroud, Robert C. Green, Xiaolan Hu, Matthew J. Huentelman, Sungeun Kim, John S. K.

Kauwe, Qingqin Li, Enchi Liu, Fabio Macciardi, Jason H. Moore, Leanne Munsie, Kwangsik Nho, Vijay K. Ramanan, Shannon L. Risacher, David J. Stone, Shanker Swaminathan, Arthur W. Toga, Michael W. Weiner, and Andrew J. Saykin. Genetic analysis of quantitative phenotypes in ad and mci: imaging, cognition and biomarkers. *Brain Imaging and Behavior*, 8(2):183–207, Jun 2014. ISSN 1931-7565. doi: 10.1007/s11682-013-9262-z. URL https://doi.org/10.1007/s11682-013-9262-z.

[22] Cynthia A. Stappenbeck and Kim Fromme. A longitudinal investigation of heavy drinking and physical dating violence in men and women. *Addictive Behaviors*, 35(5):479 – 485, 2010. ISSN 0306-4603. doi: http://dx.doi.org/10.1016/j.addbeh.2009.12.027. URL http://www.sciencedirect.com/science/article/pii/S0306460309003542.

[23] Ryota Tomioka and Taiji Suzuki. Convex tensor decomposition via structured schatten norm regularization. In C.j.c. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 1331–1339. 2013. URL http://media.nips.cc/nipsbooks/nipspapers/paper_files/nips26/686.pdf.

[24] Lan Wang, Jianhui Zhou, and Annie Qu. Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics*, 68(2):353–360, 2012. ISSN 1541-0420. doi: 10.1111/j.1541-0420.2011.01678.x. URL http://dx.doi.org/10.1111/j.1541-0420.2011.01678.x.

[25] Kishan Wimalawarne, Ryota Tomioka, and Masashi Sugiyama. Theoretical and experimental analyses of tensor-based regression and classification. *Neural Comput.*, 28(4):686–715, April 2016. ISSN 0899-7667. doi: 10.1162/NECO_a_00815. URL http://dx.doi.org/10.1162/NECO_a_00815.

[26] Tingyang Xu, Jiangwen Sun, and Jinbo Bi. 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, pages 1345–1354, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3664-2. doi: 10.1145/2783258.2783403. URL http://doi.acm.org/10.1145/2783258.2783403.

[27] Yangyang Xu and Wotao Yin. A globally convergent algorithm for nonconvex optimization based on block coordinate update. *Journal of Scientific Computing*, 72(2):700–734, Aug 2017. ISSN 1573-7691. doi: 10.1007/s10915-017-0376-0. URL https://doi.org/10.1007/s10915-017-0376-0.

[28] Hua Zhou, Lexin Li, and Hongtu Zhu. Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association*, 108(502):540–552, 2013. doi: 10.1080/01621459.2013.776499. URL http://dx.doi.org/10.1080/01621459.2013.776499.