The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Latent Sparse Modeling of Longitudinal Multi-Dimensional Data

Ko-Shin Chen,[1] **Tingyang Xu,**[2*] **Jinbo Bi**[1]
[1]Department of Computer Science and Engineering, University of Connecticut, Storrs, CT, USA
ko-shin.chen@uconn.edu, jinbo.bi@uconn.edu
[2] Tencent AI Lab, Shenzhen, China, tingyangxu@tencent.com

Abstract

We propose a tensor-based approach to analyze multi-dimensional data describing sample subjects. It simultaneously discovers patterns in features and reveals past temporal points that have 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 $L_{F,1}$ norm' as a grouped penalty in our formulation. Furthermore, the proposed model takes into account within-subject correlations by developing 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 for a fixed sample size. Computational results on synthetic datasets and real-file fMRI and EEG problems demonstrate the superior performance of the proposed approach over existing techniques.

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 (Arnold, Liu, and Abe 2007; Sela and Simonoff 2012) to scientific research (Arnold, Liu, and Abe 2007; Lozano et al. 2009; Wang, Zhou, and Qu 2012), health-care and medicine (Bi et al. 2013; Fowler and Christakis 2008; Stappenbeck and Fromme 2010).

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 (Liang and Zeger 1986) and quadratic inference function - QIF (Qu and Li 2006) 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 (Crowder 1995). In addition, the inverse of the matrix R is essential that may cause poor estimation when R has high dimension (Qu and Lindsay 2003). To overcome these disadvantages, Qu, Lindsay, and Li suggest 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 (Fu 2003) and penalized QIF (Bai, Fung, and Zhu 2009) 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 (Granger 1980). For example, Shen et al. point 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 (Arnold, Liu, and Abe 2007; Lozano et al. 2009) ignore the temporal correlations. Xu, Sun, and Bi have 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 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 (Hoff 2015; Zhou, Li, and Zhu 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 tensor-based predictive model as a function of covariates, not only from the current observation but also from multiple previous consecutive observations. Simultaneously the

*This work was done while T. Xu was with the Department of Computer Science and Engineering at University of Connecticut.

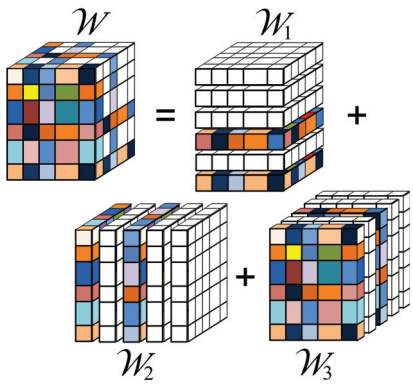

Figure 1: Case for $K = 3$: a 3-way tensor is decomposed into a summation of three 3-way tensors so that each component tensor is sparse along a particular direction.

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 (Beck and Teboulle 2009; Xu and Yin 2017) 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, followed 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.

Method

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 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.

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 $\mathcal{X}_i \in \mathbb{R}^{d_1 \times d_2 \times \ldots \times d_K}$ 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 (Wimalawarne, Tomioka, and Sugiyama 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** $L_{F,1}$ **norm** defined by

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

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.

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 $\mathcal{W} \in \mathbb{R}^{d_1 \times d_2 \times \ldots \times d_{K-1} \times d_K}$, 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 $\mathbf{y}^{(i)}$ (Liang and Zeger 1986). The model coefficients are then obtained by solving the score equation from the quasi-likelihood analysis. In our setting, it turns out to be

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

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 $\mathbf{R}^{-1}(\alpha) = a_1 \mathbf{M}_1 + a_2 \mathbf{M_2}$, where $\mathbf{M}_1$ is an identity, and $\mathbf{M}_2$ satisfies $(\mathbf{M}_2)_{i,j} = 1$ if $|i-j| = 1$, $(\mathbf{M}_2)_{i,j} = 0$ if $|i-j| \neq 1$. 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):

$$\mathbf{g}_m(\mathcal{W}) := \frac{1}{m} \sum_{i=1}^{m} \mathbf{g}^{(i)}(\mathcal{W}) \tag{9}$$

$$:= \frac{1}{m} \sum_{i=1}^{m} \begin{pmatrix} \left(\mathbf{D}^{(i)}\right)^\top \left(\mathbf{A}^{(i)}\right)^{-1/2} \mathbf{M}_1 \left(\mathbf{A}^{(i)}\right)^{-1/2} \mathbf{s}^{(i)} \\ \vdots \\ \left(\mathbf{D}^{(i)}\right)^\top \left(\mathbf{A}^{(i)}\right)^{-1/2} \mathbf{M}_d \left(\mathbf{A}^{(i)}\right)^{-1/2} \mathbf{s}^{(i)} \end{pmatrix}$$

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 (Qu and Li 2006) and (Qu, Lindsay, and Li 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 model (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 (2). More precisely, we solve the following optimization problem:

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

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

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

Asymptotic Analysis

In this section we establish the asymptotic normality for our *TensorQIF* model as m approaches to infinity. Below the convergence notations $A \xrightarrow{p} B$ represents 'A converges to B in probability,' i.e. $P(|A-B|>\varepsilon)\to 0$ for any $\varepsilon>0$; $F \xrightarrow{d} G$ denotes 'F converges to G in distribution, i.e. the distribution function of F converges to the distribution function of G.' We first rescale the objective function in (12):

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

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 $\mathbf{g}$:

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 $\mathcal{W}\in\Omega$, $\|\mathbf{g}(\mathcal{W})\mathbf{g}(\mathcal{W})^\top\|_F \le d_1(\mathcal{X},\mathbf{s})$, $\|\nabla_{\mathbf{w}}\mathbf{g}(\mathcal{W})\|_F \le d_2(\mathcal{X},\mathbf{s})$ 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*

$$\hat{\mathcal{W}}_m \xrightarrow{p} \mathcal{W}^*, \tag{15}$$

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

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

The proof of the theorem is based on a uniform convergence result for stochastic functions (Newey and McFadden 1994). A complete proof is given in the supplemental material.

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.

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

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

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

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:

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

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

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

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

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

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

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

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:

$$\begin{aligned}\min_{\mathcal{W}_k}\Bigg\{&\langle\nabla_{\mathcal{W}}Q^{(r)},\mathcal{W}_k-\mathcal{W}_k^{(r)}\rangle+\frac{KL}{2}\|\!|\mathcal{W}_k-\mathcal{W}_k^{(r)}|\!\|_F^2 \\ &+\lambda_k\sum_{j=1}^{d_k}\|\!|(\mathcal{W}_k)^{(j)}_{(k)}|\!\|_F\Bigg\}, \quad k\in\{1,\dots,K\}.\end{aligned} \tag{22}$$

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

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

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

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

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

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

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 $1 \le k \le K$.

2. Obtain $\Phi^{(r+1)} = (\mathcal{W}_1^{(r+1)}, \cdots, \mathcal{W}_K^{(r+1)})$ by solving (23) for each fixed $1 \le k \le K$.

3. $r = r + 1$.

Repeat 2 and 3 until convergence.

for all $r \ge 1$ and $1 \le k \le K$. 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, Sun, and Bi 2015) to find a proper approximation $\tilde{L} \ge L_Q$ and use $\tilde{L}$ as L in all of our computations. Algorithm 1 summarizes the steps for finding the optimal $\hat{\mathcal{W}}_k$.

Convergence Analysis

In this section, we prove that the sequence $\{\Phi^{(r)}\}_{r\ge 0}$ generated by Algorithm 1 will converge to a global optimal solution $\hat{\Phi}$ with a rate of $O(1/r)$ if the initial point $\Phi^{(0)}$ is located in a convex neighborhood of $\hat{\Phi}$. Loader and Pilla indicate that the function $Q(\mathcal{W})$ is not globally convex in general. Hence the standard convergence arguments such as (Beck and Teboulle 2009) cannot be applied directly. Furthermore, with the latent approach (13), 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 a convex function in $\Pi(\Phi)$. Assume $\Phi^{(0)}$ satisfies

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

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 $n \ge 1$,*

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

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 $\{\mathcal{W}(\Phi^{(n)})\}_{n\ge 0}$ generated by Algorithm 1 lies in $\Pi(\Omega)$ where Q is convex. Thus the convex inequality is always valid and Theorem 2 is established. Details are provided in the supplemental material.

Group Support: Values of λ_k's and L

In this section we focus on the linear model in which

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

for some true tensor coefficient $\mathcal{W}^*$, where $\tau \le t \le T$. Let $\mathcal{D} := \nabla_{\mathcal{W}} Q(\mathcal{W}^*)$. Motivated by the algorithm, we consider the following optimization problem for a fixed k:

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

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 $(\mathcal{W}_k^*)_{(k)}^{(j)} \ne 0$ if and only if $j \in S_k^*$. The KKT conditions for solutions of (28) yield

Theorem 3. *Assume $\frac{\lambda_k}{2} \ge \max_{1\le j\le d_k} \|\mathcal{D}_{(k)}^{(j)}\|_F$. Then* (28) *has a solution $\hat{\mathcal{W}}_k$ such that*

$$\{j : (\hat{\mathcal{W}}_k)_{(k)}^{(j)} \ne 0\} := \hat{S}_k \subset S_k. \quad (29)$$

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

Theorem 3 is proved in the supplemental material.

Empirical Evaluation

In this section we present the results of both synthetic and real-life fMRI/EEG examples. We test the efficiency and effectiveness of the proposed method *TensorQIF* comparing to existing methods. The datasets containing continuous responses are examined by the following methods: *TensorQIF*, Linear Regression (LR), Least Absolute Shrinkage and Selection Operator (LASSO), QIF (Qu, Lindsay, and Li 2000), PQIF (Bai, Fung, and Zhu 2009), GEE (Liang and Zeger 1986), PGEE, and Graphical Granger Modeling (Lozano et al. 2009). For GEE and PGEE, the presumed correlation is set as the 1st-order autoregressive structure (AR(1)). Namely, $corr(y_t^{(i)}, y_{t'}^{(i)}) = \alpha^{|t-t'|}$ for some $0 < \alpha < 1$. The coefficient of determination, R^2, is employed to evaluate the performance of these predicting models. An R^2 of 1 indicates perfect fitness while an R^2 of 0 indicates that the model does not fit the data.

Synthetic Data

We constructed synthetic datasets containing 150 subjects with 20 time points per subject. The data $\mathbf{X}_t^{(i)}$ at each time t is a matrix with various sizes in $\{5\times 5, 10\times 10, 15\times 15\}$. We generate $\mathbf{X}_t^{(i)}$'s from the normal distribution $N(0, 2^2)$ and $\tau = 4$. In this setting, the coefficient $\mathcal{W}$ is then a 3-way tensor, i.e. $K = 3$. Let $\mathcal{W} = \mathcal{W}_1 + \mathcal{W}_2 + \mathcal{W}_3$ be the decomposition. We generate $\mathcal{W}_1, \mathcal{W}_2$ and $\mathcal{W}_3$ such that each $\mathcal{W}_k$ simulates the latent pattern along mode-k: $\mathcal{W}_1$ has patterns (i.e. non-zero values) in the 1st and the 3rd feature along mode-1; $\mathcal{W}_2$ has patterns in the 2nd and the 3rd feature along mode-2; $\mathcal{W}_3$ selects lagged patterns at the 1st, 3rd, and 5th lagged time points. All nonzero entries of $\mathcal{W}_k$'s are from the uniform distribution $U(0, 3^2)$. We further add the residual $\mathbf{s}^{(i)}$ and $\sin(t)$ to the mean model in order to generate the outcome variable $\mathbf{y}^{(i)}$ for each subject i. The residual $\mathbf{s}^{(i)}$'s are generated from multivariate normal distribution with an

Table 1: Comparison of R^2 values achieved by our method and the other methods on the synthetic datasets with different tensor sizes and on the real-life fMRI dataset.

$d_1 \times d_2 \times (\tau+1)$	LR	LASSO	QIF	PQIF	GEE	PGEE	Granger	TensorQIF
$5 \times 5 \times 5$	0.097	0.167	0.234	0.245	0.097	0.149	0.684	**0.986**
$10 \times 10 \times 5$	0.053	0.149	0.188	0.316	0.056	0.137	0.581	**0.983**
$15 \times 15 \times 5$	0.025	0.137	0.158	0.173	0.079	0.117	0.435	**0.976**
fMRI	0.007	0.027	0.032	0.033	-	-	0.075	**0.259**

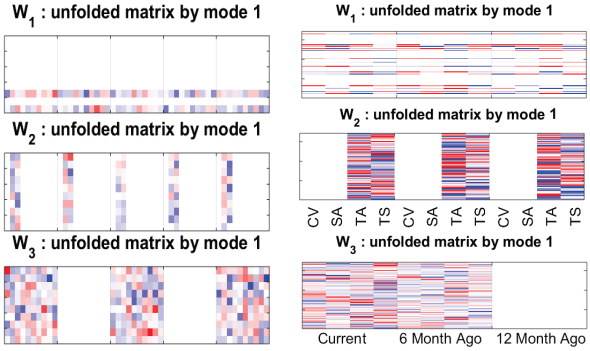

Figure 2: (Left) The model coefficients obtained by *TensorQIF* on synthetic data. (Right) Feature groups (slices) selected by *TensorQIF* for predicting MMSE scores.

AR(1) correlation structure at $\alpha = 0.6$. Then the outcome $y_t^{(i)}$ is computed as

$$y_t^{(i)} = \langle \mathcal{X}_{(i;t)}, (\mathcal{W}_1 + \mathcal{W}_2 + \mathcal{W}_3) \rangle + s_t^{(i)} + \sin(t).$$

In our experiments, λ's are tuned as $\lambda_1 = \lambda_2 = \lambda_3 = 0.3$ based on cross validation within training. We randomly select 80% of the subjects for training and the rest for testing.

Table 1 provides the R^2 comparison results between *TensorQIF* and the other seven methods for three different sizes synthetic datasets. The proposed method *TensorQIF* outperforms the traditional regression methods in all comparison scenarios in terms of predicting accuracy. LR and LASSO have poor results that might due to the number of records are relatively small comparing to total features in data tensor. For the marginal models, QIF-based methods performs better than the baseline methods because they can efficiently estimate the coefficients even though the correlation matrix is misspecified while GEE-based methods performs poorer because they misspecify the correlation matrix. The Graphical Granger Modeling shows relatively high performance because it models the effects from lagged time points. However, it's performance suffers from the high correlations within the subjects.

In Figure 2 (Left) we plot the mode-1 unfolded matrices resolved from *TensorQIF* on the dataset of the size $10 \times 10 \times 5$. Red (blue) color indicates that the corresponding features are positive (negative) predictors of the response variable. Features with white color are not selected. We see all subtensors (matrices) of $\mathcal{W}_1, \mathcal{W}_2$ and $\mathcal{W}_3$ capture the designed structure of the synthetic data. This explains the reason of achieving around 0.98 R^2 by the proposed method. Figure 3 illustrates the result from the Granger model. It's tensor coefficients are reshaped to align the matrices in Figure 2. It clearly showed the wrong selections on the first lagged time point which was due to the correlations within subjects.

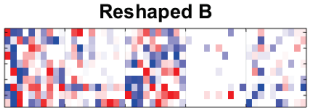

Figure 3: The model coefficients resolved from the Granger model (Lozano et al. 2009) on synthetic data.

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 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 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

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

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

parameter values that gave the best two-fold cross validation performance. As shown in Table 1, our method performs the best predictions.

Moreover, 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 2 (Right), 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.

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 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.

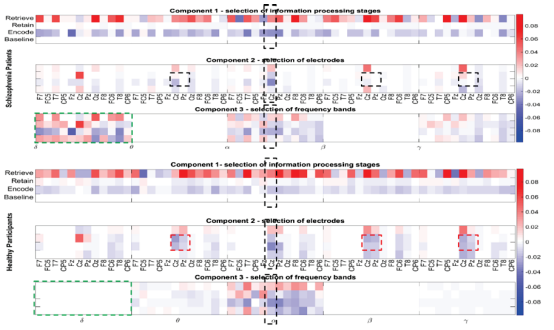

Figure 4: Rows, columns and slices selected by *TensorQIF* for SZ (the top panel) and HC (the bottom panel). The figure is also provided in the supplemental material.

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 hyperparameters λ_1, λ_2, and λ_3 in our approach and GEE/PGEE (one parameter) 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 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 (Xu, Sun, and Bi 2015). Groups differed in two main ways: (1) centroparietal theta, beta, and gamma during encoding and retention predicted higher accuracy in HC (Figure 4; red border), and (2) delta activity across stages and electrodes (Figure 4; green border) predicted lower accuracy in SZ. 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 (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.

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. Empirical studies on both synthetic and real-life problems demonstrate the superior performance of the proposed method.

Acknowledgements

This work was partially supported by NSF grants DBI-1356655, CCF-1514357, IIS-1718738, as well as NIH grants R01DA037349 and K02DA043063 to Jinbo Bi.

References

Arnold, A.; Liu, Y.; and 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. New York, NY, USA: ACM.

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

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

Bi, J.; Sun, J.; Wu, Y.; Tennen, H.; and Armeli, S. 2013. A machine learning approach to college drinking prediction and risk factor identification. *ACM Trans. Intell. Syst. Technol.* 4(4):72:1–72:24.

Crowder, M. 1995. On the use of a working correlation matrix in using generalised linear models for repeated measures. *Biometrika* 82(2):407–41–.

Diggle, P.; Heagerty, P.; Liang, K.-Y.; and Zeger, S. 2002. *Analysis of Longitudinal Data*. Oxford University Press.

Fowler, J. H., and Christakis, N. A. 2008. Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the framingham heart study. *BMJ* 337.

Fu, W. J. 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, L. P. 1982. Large sample properties of generalized method of moments estimators. *Econometrica* 50(4):1029–1054.

Hoff, P. D. 2015. Multilinear tensor regression for longitudinal relational data. *Ann. Appl. Stat.* 9(3):1169–1193.

Johannesen, J. K.; Bi, J.; Jiang, R.; Kenney, J. G.; and Chen, C.-M. A. 2016. Machine learning identification of eeg features predicting working memory performance in schizophrenia and healthy adults. *Neuropsychiatric Electrophysiology* 2(1):3.

Liang, K.-Y., and Zeger, S. L. 1986. Longitudinal data analysis using generalised estimating equations. *Biometrika* 73(1):13–22.

Loader, C., and Pilla, R. S. 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, A. C.; Abe, N.; Liu, Y.; and 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. New York, NY, USA: ACM.

Newey, W. K., and McFadden, D. 1994. Chapter 36 large sample estimation and hypothesis testing. *Handbook of Econometrics* 4:2111 – 2245.

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

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

Sela, R. J., and Simonoff, J. S. 2012. Re-em trees: a data mining approach for longitudinal and clustered data. *Machine Learning* 86(2):169–207.

Shen, L.; Thompson, P. M.; Potkin, S. G.; Bertram, L.; Farrer, L. A.; Foroud, T. M.; Green, R. C.; Hu, X.; Huentelman, M. J.; Kim, S.; Kauwe, J. S. K.; Li, Q.; Liu, E.; Macciardi, F.; Moore, J. H.; Munsie, L.; Nho, K.; Ramanan, V. K.; Risacher, S. L.; Stone, D. J.; Swaminathan, S.; Toga, A. W.; Weiner, M. W.; and Saykin, A. J. 2014. Genetic analysis of quantitative phenotypes in ad and mci: imaging, cognition and biomarkers. *Brain Imaging and Behavior* 8(2):183–207.

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

Wang, L.; Zhou, J.; and Qu, A. 2012. Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics* 68(2):353–360.

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

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

Xu, T.; Sun, J.; and 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. New York, NY, USA: ACM.

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