| **Statistica Sinica Preprint No:** SS-2017-0153 |
|-----------------------------------------------|
| **Title** | Tensor Generalized Estimating Equations for Longitudinal Imaging Analysis |
| **Manuscript ID** | SS-2017-0153 |
| **URL** | http://www.stat.sinica.edu.tw/statistica/ |
| **DOI** | 10.5705/ss.202017.0153 |
| **Complete List of Authors** | Xiang Zhang  
Lexin Li  
Hua Zhou  
Yeqing Zhou  
Dinggang Shen and ADNI |
| **Corresponding Author** | Hua Zhou |
| **E-mail** | huazhou@ucla.edu |

Notice: Accepted version subject to English editing.
TENSOR GENERALIZED ESTIMATING EQUATIONS
FOR LONGITUDINAL IMAGING ANALYSIS

Xiang Zhang¹, Lexin Li², Hua Zhou³, Yeqing Zhou², Dinggang Shen⁴, and ADNI⁵

¹North Carolina State University, ²University of California, Berkeley
³University of California, Los Angeles, ⁴University of North Carolina, Chapel Hill
and ⁵the Alzheimer’s Disease Neuroimaging Initiative

Abstract: Longitudinal neuroimaging studies are rapidly emerging, where brain images are collected on multiple subjects at multiple time points. Analysis of such data is scientifically important, but also challenging. Brain image is in the form of multidimensional array, or tensor, which possesses both ultrahigh dimensionality and complex structure. Longitudinally repeated images and the induced temporal correlations add another layer of complexity. Despite some recent efforts, there exist very few solutions for longitudinal imaging analysis. In response to the increasing demand of analyzing longitudinal imaging data, we propose tensor generalized estimating equations (GEE) in this article. The GEE approach accounts for intra-subject correlation, whereas an imposed low rank structure on the coefficient tensor effectively reduces the dimensionality. We propose a scalable estimation algorithm, establish the asymptotic properties, and investigate sparsity regularization for the purpose of region selection. We demonstrate the proposed method through simulations and analysis of a real dataset from the Alzheimer’s Disease Neuroimaging Initiative.

Key words and phrases: Generalized estimating equations; longitudinal imaging; magnetic resonance imaging; low rank tensor decomposition; multidimensional array; tensor regression.

1. Introduction

In recent years, an increasing number of longitudinal neuroimaging studies are rapidly emerging, where brain images are collected for multiple subjects and each at multiple time points (Zhang et al., 2012). Analysis of such longitudinal images is particularly important for understanding disease progression, predicting onset of disorders, and identifying disease relevant brain regions. Our motivating example is a study from the
Alzheimer’s Disease Neuroimaging Initiative (ADNI). Alzheimer’s disease (AD) is a progressive and irreversible neurodegenerative disorder and the leading form of dementia in elderly subjects. The dataset consists of 88 subjects with mild cognitive impairment (MCI), a prodromal stage of AD. Each subject had magnetic resonance imaging (MRI) scans at 5 different time points: baseline, 6-month, 12-month, 18-month and 24-month. After preprocessing, each MRI image is a $32 \times 32 \times 32$ three dimensional array. Also measured for each subject at each visit was a cognitive score, the Mini-Mental State Examination (MMSE), which measures the disease progression. In this study, it is of scientific interest to understand association between MCI/AD and the structural brain atrophy as reflected by MRI. It is equally important to use MRI images to accurately predict AD/MCI, as an accurate diagnosis is critical for timely therapy and possible delay of the disease [Zhang et al., 2011].

Longitudinal imaging analysis, however, is particularly challenging. Each image is in the form of a multidimensional array, a.k.a. tensor, which possesses both ultrahigh dimensionality and complex structure. For instance, a $32 \times 32 \times 32$ MRI image involves $32^3 = 32,768$ parameters, whereas the number of subjects is rarely beyond hundreds. A single image also carries complex spatial correlations among its voxels, and naively turning an array into a vector would both result in extremely high dimensionality, and destroy all the inherent spatial information. Moreover, repeated images of the same subjects are also temporally correlated. Facing increasing availability of longitudinal imaging data, there is a relative paucity of effective solutions, and thus a substantial demand for a systematic development of new longitudinal imaging analysis methods.

In this article, we propose tensor generalized estimating equations for longitudinal imaging analysis. Our proposal consists of two key components: a low rank tensor decomposition and generalized estimating equations (GEE). We impose a low rank structure on the coefficient array in GEE, which implicitly utilizes the spatial structure of the image predictor. Meanwhile it substantially reduces the number of free parameters, and thus makes subsequent estimation and inference feasible. We incorporate this structure in estimating equations, so to accommodate longitudinal correlation of the data.
Within the framework of tensor GEE, we develop a scalable computational algorithm for solving the complicated tensor estimating equations. We also study the $L_1$ and SCAD type penalized tensor GEE for the purpose of finding brain subregions that are highly relevant to the clinical outcome. This region selection is itself of vital scientific interest, and corresponds to the intensively studied variable selection problem in classical regression with vector-valued predictors. Furthermore, we establish the asymptotic properties of the solution of tensor GEE. In particular, we show that the tensor GEE estimator inherits the robustness feature of the classical GEE estimator, in that the estimate is consistent even if the working correlation structure is misspecified.

Our proposal is related to but also clearly distinctive of existing works on longitudinal data and tensor data analysis. We briefly review the literature here, and point out the differences and our contributions. First, there is a long list of longitudinal data analysis ([Liang and Zeger, 1986; Prentice and Zhao, 1991; Li, 1997; Qu et al., 2000; Xie and Yang, 2003; Balan and Schiopu-Kratina, 2005; Song et al., 2009; Wang, 2011]), and variable selection for longitudinal models ([Pan, 2001; Fu, 2003; Fan and Li, 2004; Ni et al., 2010; Xue et al., 2010; Wang et al., 2012]). However, all those methods take a vector of covariates, whereas in our problem, covariates take the form of a multi-dimensional array. Second, most existing neuroimaging studies utilize only the baseline imaging data, ignoring all the information at the follow-up time points. There have recently emerged some works using longitudinal images for individual-based classification ([Misra et al., 2009; Davatzikos et al., 2009; McEvoy et al., 2011; Hinrichs et al., 2011]), or for cognitive score prediction ([Zhang et al., 2012]). All those solutions extract a vector of summary features from the longitudinal images. By contrast, we jointly model all the voxels of the image and take in a tensor predictor. There are also studies regressing longitudinal images on a vector of predictors ([Skup et al., 2012; Li et al., 2013b]), which differ from our work in that they treat image as response instead of predictor. Third, there have been a number of applications of tensor decompositions in statistical models ([Zhou et al., 2013; Zhou and Li, 2014; Aston et al., 2017; Sun et al., 2017; Raskutti and Yuan, 2016]). Our proposal shares a similar spirit, by imposing a low rank structure on the tensor
GEE coefficient for effective dimension reduction. In that sense, our work generalizes the classical GEE from a vector predictor to a tensor, and the tensor predictor regression [Zhou et al., 2013; Raskutti and Yuan, 2016] from independent imaging data to longitudinal imaging. Such a generalization may seem straightforward conceptually, but is far from trivial technically. To our knowledge, our work is the first that systematically tackles a longitudinal imaging predictor in a regression context. As such it offers both a timely response to the increasing demand of longitudinal neuroimaging, as well as a useful addition to the methodology of longitudinal data analysis.

The rest of the article is organized as follows. Section 2 proposes the tensor GEE, along with its estimation and regularization. Section 3 studies the asymptotic properties. Sections 4 and 5 present the simulations and real data analysis. Section 6 concludes with a discussion. The Supplementary Materials contain all the technical proofs.

2. Methodology

2.1. Tensor generalized estimating equations

Suppose there are \( n \) training subjects, and for the \( i \)-th subject, there are observations over \( m_i \) time points. For simplicity, we assume \( m_i = m \) and the time points are the same for all subjects. The observed data consist of \( \{(Y_{ij}, X_{ij}), i = 1, \ldots, n, j = 1, \ldots, m\} \), where \( Y_{ij} \) denotes the target response and \( X_{ij} \in \mathbb{R}^{p_1 \times \cdots \times p_D} \) is a \( D \)-dimensional array representing the image. We remark that our model can naturally incorporate an additional vector of covariates, \( Z \). However, we choose to drop this term to simplify the presentation. Write \( Y_i = (Y_{i1}, \ldots, Y_{im})^T \). A key attribute of longitudinal data is that the observations from different subjects are commonly viewed as independent, but the observations from the same subject are correlated. That is, the intra-subject covariance matrix, \( \text{Var}(Y_i) \in \mathbb{R}^{m \times m} \), is not diagonal but with some structure.

The GEE method has been widely employed for analyzing correlated longitudinal data since the pioneer work of [Liang and Zeger, 1986]. It requires specification of the first two moments of the conditional distribution of the response given the covariates, \( \mu_{ij} = E(Y_{ij}|X_{ij}) \) and \( \sigma^2_{ij} = \text{Var}(Y_{ij}|X_{ij}) \). Following [Liang and Zeger, 1986], we assume \( Y_{ij} \) is from an exponential family with a canonical link. Then \( \mu_{ij}(B) = \mu(\theta_{ij}) \) and
\( \sigma^2_{ij}(B) = \phi \mu^{(1)}(\theta_{ij}), \; i = 1, \ldots, n, \; j = 1, \ldots, m, \) where \( \mu(\cdot) \) is a differentiable canonical link function, \( \mu^{(1)}(\cdot) \) is its first derivative, \( \theta_{ij} \) is the linear systematic part, and \( \phi \) is an over-dispersion parameter. In this article, we simply set \( \phi = 1 \), while the extension to a general \( \phi \) is straightforward. The systematic part \( \theta_{ij} \) is associated with the covariates via the equation,

\[
\theta_{ij} = \langle B, X_{ij} \rangle, \tag{2.1}
\]

where \( B \) is the coefficient tensor of the same size as \( X \) that captures effects of every array element of \( X \) on \( Y \). The inner product \( \langle B, X_{ij} \rangle = \langle \text{vec}(B), \text{vec}(X_{ij}) \rangle \), and the \( \text{vec}(B) \) operator stacks the entries of a tensor \( B \) into a column vector. The GEE estimator of \( B \) is then defined as the solution of

\[
\sum_{i=1}^{n} \frac{\partial \mu_i(B)}{\partial \text{vec}(B)} V_i^{-1} \{Y_i - \mu_i(B)\} = 0, \tag{2.2}
\]

where \( Y_i = (Y_{i1}, \ldots, Y_{im})^\top \), \( \mu_i(B) = [\mu_{i1}(B), \ldots, \mu_{im}(B)]^\top \), and \( V_i = \text{cov}(Y_i) \) is the response covariance matrix of the \( i \)-th subject. The first component in (2.2) is the derivative of \( \mu_i(B) \) with respect to the vector \( \text{vec}(B) \in \mathbb{R}^{\prod_d p_d} \). As such, there are totally \( \prod_d p_d \) estimating equations to solve in (2.2). For regression with image covariates, this dimension is prohibitively high and usually far exceeds the sample size. For instance, for a \( 32 \times 32 \times 32 \) MRI image predictor, the number of equations to solve is in the scale of \( 32^3 = 327,680 \), resulting no unique solution when the sample size is only in tens or hundreds. It thus becomes crucial to reduce the number of estimating equations.

Toward that end, we impose a low rank structure on the coefficient array \( B \). More specifically, we assume \( B \) in model (2.1) follows a canonical polyadic (CP) decomposition structure [Kolda and Bader 2009],

\[
B = \sum_{r=1}^{R} \beta_1^{(r)} \circ \cdots \circ \beta_D^{(r)}, \tag{2.3}
\]

where \( \beta_d^{(r)} \in \mathbb{R}^{p_d}, d = 1, \ldots, D, r = 1, \ldots, R, \) are all column vectors, \( \circ \) denotes the outer product, and \( B \) cannot be written as a sum of less than \( R \) outer products. The decomposition (2.3) is often represented by a shorthand, \( B = [B_1, \ldots, B_D] \), where
$\mathbf{B}_d = [\beta_d^{(1)}, \ldots, \beta_d^{(R)}] \in \mathbb{R}^{pd \times R}$. Under this structure, the systematic part in (2.1) becomes

$$\theta_{ij} = \left\langle \sum_{r=1}^{R} \beta^{(r)}_i \circ \cdots \circ \beta^{(r)}_D, \mathbf{X}_{ij} \right\rangle = \left\langle (\mathbf{B}_D \odot \cdots \odot \mathbf{B}_1) \mathbf{1}_R, \text{vec} \mathbf{X}_{ij} \right\rangle.$$  

We then propose the tensor generalized estimating equations estimator of $\mathbf{B}$, which is defined as the solution of

$$\sum_{i=1}^{n} \frac{\partial \mu_i(\mathbf{B})}{\partial \beta_B} V_i^{-1} \{ \mathbf{Y}_i - \mu_i(\mathbf{B}) \} = 0,$$  

where $\beta_B = \text{vec}(\mathbf{B}_1, \ldots, \mathbf{B}_D)$, and the subscript $B$ is to remind that $\beta$ is constructed based on the CP decomposition of a given coefficient tensor $\mathbf{B} = [\mathbf{B}_1, \ldots, \mathbf{B}_D]$. Introducing the CP structure into GEE has two important implications. First, comparing to the classical GEE (2.2), the derivative in (2.4) is now with respect to $\beta_B \in \mathbb{R}^{R \sum_d pd}$. Consequently, the number of estimating equations is reduced from the exponential order $\prod_d pd$ to the linear order $R \sum_d pd$. This substantial reduction in dimensionality is the key to enable effective estimation and inference under a limited sample size. Second, under this structure, any two elements $\beta_{i_1 \ldots i_d}$ and $\beta_{j_1 \ldots j_d}$ in $\mathbf{B}$ share common parameters if $i_d = j_d$ for any $d = 1, \ldots, D$. In consequence, the coefficients are correlated if they share the same spatial locations along any one of the tensor modes. This incorporates, implicitly, the spatial structure of the tensor coefficient.

Examining (2.4), the true intra-subject covariance structure $V_i$ is usually unknown in practice. The classical GEE adopts a working covariance matrix, specified through a working correlation matrix $\mathbf{R}$. That is, $V_i = \mathbf{A}_i^{1/2}(\mathbf{B}) \mathbf{R} \mathbf{A}_i^{1/2}(\mathbf{B})$, where $\mathbf{A}_i(\mathbf{B})$ is an $m \times m$ diagonal matrix with $\sigma_{ij}^2(\mathbf{B})$ on the diagonal and $\mathbf{R}$ is the $m$-by-$m$ working intra-subject correlation matrix. Some commonly used correlation structures include independence, autocorrelation (AR), compound symmetry, and unstructured correlation, among others. The correlation matrix $\mathbf{R}$ may involve additional parameters, which can be estimated using a residual-based moment method.

By adopting this working correlation idea, and explicitly evaluating the derivative in (2.4), we finally arrive at the formal definition of the tensor GEE estimator, which is
the solution \(\hat{B}\) of the following estimating equations
\[
\sum_{i=1}^{n} [J_1, J_2, \ldots, J_D]^\top \text{vec}(X_i) A_i^{1/2}(B) \hat{R}^{-1} A_i^{-1/2}(B)\{Y_i - \mu_i(B)\} = 0, \tag{2.5}
\]
where \(\hat{R}\) is an estimated correlation matrix, \(\text{vec}(X_i) = (\text{vec}(X_{i1}), \ldots, \text{vec}(X_{im}))\) is a \(\prod_{d=1}^{D} p_d \times m\) matrix, \(J_d\) is the \(\prod_{d=1}^{D} p_d \times Rp_d\) Jacobian matrix of the form \(\Pi_d \times [B_D \circ \cdots \circ B_{d+1} \circ B_{d-1} \circ \cdots \circ B_1] \otimes I_{p_d}\), where \(\Pi_d\) is the \((\prod_{d=1}^{D} p_d)-\text{by-}(\prod_{d=1}^{D} p_d)\) permutation matrix that reorders \(\text{vec}B_{(d)}\) to obtain \(\text{vec}B\), i.e., \(\text{vec}B = \Pi_d \times \text{vec}B_{(d)}\). Note that \(\mu^{(1)}(\theta_{ij})\) has been canceled by the diagonals on the matrix \(A_i^{-1}\) due to the property of canonical link. For ease of presentation, we denote the left hand side of equation (2.5) as \(s(B)\), and write the tensor GEE (2.5) as \(s(B) = 0\).

### 2.2. Estimation and rank selection

Directly solving the tensor GEE (2.5) with respect to \(B\) can be computationally intensive, as the mean of the response given the covariates is nonlinear in the parameters and the Jacobian matrices \(J_1, \ldots, J_D\) also depend on the unknown parameters. We propose a block relaxation algorithm to iteratively solve the sub-GEE for \(B_1, \ldots, B_D\) one at a time, while keeping all other components fixed. Specifically, when updating \(B_d \in \mathbb{R}^{p_d \times R}\), the systematic part \(\theta_{ij}(B)\) can be rewritten as
\[
\theta_{ij}(B) = \langle B, X_{ij} \rangle = \langle B_d, X_{ij(d)}(B_D \circ \cdots \circ B_{d+1} \circ B_{d-1} \circ \cdots \circ B_1)\rangle,
\]
where \(X_{ij(d)}\) is the mode-\(d\) matricization of the tensor \(X_{ij}\), which flattens \(X_{ij}\) into a \(p_d \times \prod_{d' \neq d} p_{d'}\) matrix, such that the \((k_1, \ldots, k_D)\) element of \(X_{ij}\) maps to the \((k_d, l)\) element of the matrix \(X_{ij(d)}\), where \(l = 1 + \sum_{d'' \neq d} (k_{d''} - 1) \prod_{d''' < d', d''' \neq d} p_{d'''}\), and \(\circ\) denotes the Khatri-Rao product [Rao and Mitra 1971]. Consequently, the systematic part \(\theta_{ij}(B)\) becomes linear in \(B_d\). The Jacobian matrix \(J_d\) is free of \(B_d\) and depends on the covariates and fixed parameters only. Then each step reduces to a standard GEE problem with \(Rp_d\) parameters, which can be solved using standard statistical softwares. Same as the classical GEE, our tensor GEE potentially has multiple roots too. Our numerical experience has found that different starting values often lead to the same solution.

A problem of practical importance is to choose the rank \(R\) for the coefficient array \(B\) in its CP decomposition. This can be viewed as a model selection problem. [Pan]
proposed a quasi-likelihood independence model criterion for the classical GEE model selection, where the core idea is to evaluate the likelihood under the independence working correlation assumption. In our tensor GEE setup, we adopt a similar criterion,

$$\text{BIC}(R) = -2\ell(\hat{B}(R); I_m) + \log(n)p_e,$$

where $\ell(\hat{B}(R); I_m)$ is the log-likelihood evaluated at the tensor GEE estimator $\hat{B}(R)$, with a working rank $R$ and the independence working correlation structure $I_m$. For simplicity, we call this criterion BIC, since the term $\log(n)$ is used. Because the CP decomposition itself is not unique, but can be made so under some minor conditions (Zhou et al., 2013), the actual number of estimating equations, or the effective number of parameters, is of the form: $p_e = R(p_1 + p_2) - R^2$ for $D = 2$, and $p_e = R(\sum_d p_d - D + 1)$ for $D > 2$. We choose $R$ that minimizes this criterion among a series of working ranks.

2.3. Regularization for region selection

Selecting brain subregions that are highly relevant to the disease outcome is of vital scientific interest. It allows researchers to concentrate on brain subregions for improved understanding of the disease pathology, and for hypothesis generation and validation. In our setup, region selection translates to sparse estimation of the elements of the coefficient tensor $B$, and is analogous to the intensively studied variable selection in classical vector-valued regression. We adopt the $L_1$ type regularization to achieve this goal. Specifically, we consider the following regularized tensor GEE

$$n^{-1}s(B) - \begin{bmatrix}
\frac{\partial \beta_{11}(1)}{\beta_{11}} P_\lambda(|\beta_{11}(1)|, \rho_n) \\
\vdots \\
\frac{\partial \beta_{di}(r)}{\beta_{di}} P_\lambda(|\beta_{di}(r)|, \rho_n) \\
\vdots \\
\frac{\partial \beta_{DPD}(R)}{\beta_{DPD}} P_\lambda(|\beta_{DPD}(R)|, \rho_n)
\end{bmatrix} = 0,$$

(2.7)

where $P_\lambda(|\beta|, \rho_n)$ is a scalar penalty function, $\rho_n$ is the penalty tuning parameter, $\lambda$ is an index for the penalty family, and $\partial_\beta P_\lambda(|\beta|, \rho_n)$ is the subgradient with respect to the argument $\beta$. We consider two specific penalty functions: the Lasso (Tibshirani, 1996)
in which $P_\lambda(|\beta|, \rho_n) = \rho_n|\beta|$ with $\lambda = 1$, and the SCAD (Fan and Li 2001), in which
\[\frac{\partial}{\partial |\beta|} P_\lambda(|\beta|, \rho_n) = \rho_n \left \{ 1_{|\beta| \leq \rho_n} + (\lambda \rho_n - |\beta|)/((\lambda - 1)1_{|\beta| > \rho_n}) \right \}, \lambda > 2.\]

Thanks to the separability of parameters in the regularization term, the alternating updating strategy still applies. When updating $B_d$, we solve the penalized sub-GEE
\[
n^{-1} s_d(B_d) = \left( \begin{array}{c} \frac{\partial \beta^{(1)}_{d1}}{\partial \beta} P_\lambda(|\beta^{(1)}_{d1}|, \rho_n) \\ \vdots \\ \frac{\partial \beta^{(r)}_{d1}}{\partial \beta} P_\lambda(|\beta^{(r)}_{d1}|, \rho_n) \\ \vdots \\ \frac{\partial \beta^{(R)}_{dpd}}{\partial \beta} P_\lambda(|\beta^{(R)}_{dpd}|, \rho_n) \end{array} \right) = 0, \quad (2.8)
\]
where $s_d$ is the sub-estimation equation for block $B_d$, and there are $R_{pd}$ equations to solve at this step. The anti-derivative of $s_d$ is recognized as the loss of an Aitken linear model with block diagonal covariance matrix. Thus after a linear transformation of $Y_i$ and the working design matrix, the solution to (2.8) is the same as the minimizer of a regular penalized weighted least squares problem, for which many software packages exist. The fitting procedure boils down to alternating penalized weighted least squares.

We also briefly comment that, in addition to region selection, regularization is also useful to stabilize the estimates, to handle small-$n$-large-$p$, and to incorporate prior subject knowledge. The above regularization paradigm can be extended to incorporate other forms of regularization, e.g., the $L_2$ type ridge regularization, or different penalties along different modes of the tensor coefficient.

3. Theory

Next we study the asymptotic properties of the tensor GEE estimator. We first note that, in tensor GEE, there are two specifications, or potentially misspecifications. The first is the working correlation structure. We show that, the tensor GEE estimator remains consistent even if the working correlation structure is misspecified. This is an analogous result as the classical GEE, and our work builds upon and extends Xie and Yang (2003); Balan and Schiopu-Kratina (2005); Wang (2011). We achieve this by assuming the rank is fixed and known. This is similar in spirit to the classical GEE
setup, where the linear model is imposed and the rank is in effect set to one. The second specification is the working rank of the CP decomposition in tensor GEE. We show that, for the normal linear model, the rank selected by BIC under an independent correlation structure is consistent, even if this structure might have been misspecified. This justifies the BIC criterion \(^2.6\), and, to some extent, the asymptotic investigation under a known rank. We remark that, assuming a known rank is common in theoretical analysis of estimators based on low rank structures (Zhou et al., 2013; Sun and Li, 2016). We also remark that, our asymptotic study is carried out in the classical sense that the number of parameters (dimension) is fixed and the sample size goes to infinity. We believe such a fixed dimension asymptotic study is useful, as it reveals the basic properties and offers a statistical guarantee for our tensor GEE estimator. More importantly, it establishes that both our tensor estimator and the rank estimator remain consistent under a potentially incorrect working correlation structure. In principle, one can also consider the scenario where the dimension diverges to infinity along with the sample size. We have obtained some preliminary asymptotic results, but leave a comprehensive treatment of the tensor GEE under a diverging dimension for future research.

3.1. Regularity conditions

We begin with a list of regularity conditions for the asymptotics of tensor GEE with a fixed number of parameters. Let \(\|x\|\) denote the Euclidean norm of a vector \(x\) and \(\|X\|_F\) the Frobenius norm of a matrix \(X\). Denote \(N_n\) the neighborhood of the true tensor coefficient \(\{B : \|\beta_B - \beta_{B_0}\| \leq \Delta n^{-1/2}\}\) for some constant \(\Delta > 0\).

(A1) For some constant \(c_1 > 0\), \(\|X_{ij}\|_F \leq c_1, i = 1, \ldots, n, j = 1, \ldots, m\).

(A2) The true value \(B_0\) of the unknown parameter lies in the interior of a compact parameter space \(B\) and follows a rank-\(R\) CP structure defined in \(^2.3\).

(A3) Let \(I(B) = n^{-1} \sum_{i=1}^{n}[J_1, J_2, \ldots, J_D] \text{vec}X_i \text{vec}^T \text{vec}X_i[J_1, J_2, \ldots, J_D]\). There exist two constants \(0 < c_2 < c_3\) such that, \(c_2 \leq \lambda_{\text{min}}(I(B)) \leq \lambda_{\text{max}}(I(B)) \leq c_3\) over the set \(N_n\), where \(\lambda_{\text{min}}\) and \(\lambda_{\text{max}}\) are smallest and largest eigenvalue, respectively. Additionally, on the same set \(I(B)\) has a constant rank.
(A4) The true intra-subject correlation matrix $R_0$ has bounded eigenvalues from zero and infinity. There exists a positive definite matrix $\tilde{R}$ with eigenvalues bounded away from zero and infinity, such that $\|\hat{R}^{-1} - \tilde{R}^{-1}\|_F = O_p(n^{-1/2})$, where $\hat{R}$ is an estimator of the correlation matrix.

(A5) For $\delta > 0$ and $c_4 > 0$, $E(\|A_i^{-1/2}(B_0)(Y_i - \mu_i(B_0))\|^{2+\delta} \leq c_4$ for all $1 \leq i \leq n$.

(A6) For some constant $c_5 > 0$, $\|\partial \theta_{ij}(B)/\partial \beta_B\| \leq c_5$, $i = 1, \ldots, n$, $j = 1, \ldots, m$.

(A7) Denote by $\mu^{(k)}(\theta_{ij})$, $i = 1, \ldots, n$, $j = 1, \ldots, m$, and $k = 2, 3$, the $k$-th derivative of $\mu(\theta_{ij})$. For some positive constants $c_6 < c_7$ and $c_8$, we have $c_6 < |\mu^{(1)}(\theta_{ij})| < c_7$, and $|\mu^{(k)}(\theta_{ij})| < c_8$, over the set $N_n$.

(A8) Denote by $H_{ij}(B) = \partial^2 \theta_{ij}(B)/\partial \beta_B \partial \beta_B^T$. That is, $H_{ij}(B)$ is the Hessian matrix of the linear systematic part $\theta_{ij}$. There exist two positive constants $c_9 < c_{10}$ such that, for $i = 1, \ldots, n$ and $j = 1, \ldots, m$, $c_9 \leq \lambda_{\min}(H_{ij}(B)) \leq \lambda_{\max}(H_{ij}(B)) \leq c_{10}$ over the set $N_n$.

A few remarks are in order. Conditions (A2) and (A3) are required for model identifiability of tensor GEE [Zhou et al., 2013]. We observe that, the matrix $I(B)$ in (A3) is an $R \sum_{d=1}^D p_d \times R \sum_{d=1}^D p_d$ matrix, and thus (A3) is much weaker than the nonsingularity condition on the design matrix if one were to directly vectorize the tensor covariate. Condition (A4) is commonly imposed in the GEE literature. It only requires that $\hat{R}$ be a consistent estimator of some $\tilde{R}$, in the sense $\|\hat{R}^{-1} - \tilde{R}^{-1}\|_F = O_p(n^{-1/2})$. $\tilde{R}$ needs to be well behaved in that it is positive definite with bounded eigenvalues from zero and infinity, but $\hat{R}$ does not have to be the true intra-subject correlation $R_0$. This condition essentially leads to the robust feature in Theorem 1 that the tensor GEE estimate is consistent even if the working correlation structure is misspecified. Condition (A5) regulates the tail behavior of the residuals so that the noise cannot accumulate too fast, and we can employ the Lindeberg-Feller central limit theorem to control the asymptotic behavior of the residuals. Condition (A6) states that the gradients of the systematic part are well-defined. Condition (A7) concerns the canonical link and generally holds for
common exponential families, for example, the binomial and the Poisson distributions. Condition (A8) ensures that the Hessian matrix $H(B)$ of the linear systematic part, which is highly sparse, is well-behaved in a neighborhood of the true value.

### 3.2. Consistency and asymptotic normality

Before we turn to the asymptotics of the tensor GEE estimator, we address two components involved in the estimating equations: the initial estimator and the correlation estimator. Recall the tensor GEE estimator $\hat{B}$ is obtained by solving the equations

$$\sum_{i=1}^{n}[J_1, \ldots, J_D]^\top \text{vec} X_i A_i^{1/2}(B) \hat{R}^{-1} A_i^{-1/2}(B) \{Y_i - \mu_i(B)\} = 0,$$

where $\hat{R}$ is any estimator of the intra-subject correlation matrix satisfying condition (A4). Note that $\hat{R}$ is often obtained via residual-based moment method, which in turn requires an initial estimator of $B_0$. Next, we examine some frequently used estimators of $\hat{B}$ and $\hat{R}$.

A customary initial estimator $\hat{B}$ in the GEE literature is the one that assumes an independent working correlation. That is, one completely ignores potential intra-subject correlation, and the corresponding tensor GEE becomes

$$\sum_{i=1}^{n}[J_1, \ldots, J_D]^\top \text{vec} X_i \{Y_i - \mu_i(B)\} = 0.$$

Denoting the equations as $s_{\text{init}}(B) = 0$, and the solution as $\hat{B}_{\text{init}}$, the next Lemma shows that it is a consistent estimator of the true $B_0$.

**Lemma 1.** Under conditions (A1)-(A3) and (A5)-(A8), there exists a root $\hat{B}_{\text{init}}$ of the equations $s_{\text{init}}(B) = 0$ satisfying

$$\|\beta_{\hat{B}_{\text{init}}} - \beta_{B_0}\| = O_p(n^{-1/2}).$$

Here $\beta_B = \text{vec}(B_1, \ldots, B_D)$, and is constructed based on the CP decomposition of a given tensor $B = [B_1, \ldots, B_D]$, as defined before. Given a consistent initial estimator of $B_0$, there exist multiple choices for the working correlation structure, e.g., autocorrelation, compound symmetry, and the nonparametric structure (Balan and Schiopu-Kratina, 2005). We investigate those choices in Sections 4 and 5.
Next we establish the consistency and asymptotic normality of the tensor GEE estimator from (2.5).

**Theorem 1.** Under conditions (A1)-(A8), there exists a root $\hat{B}$ of the equations $s(B) = 0$ satisfying that

$$\|\beta_{\hat{B}} - \beta_{B_0}\| = O_p(n^{-1/2}).$$

The key message of Theorem 1, as implied by condition (A4), is that the consistency of the tensor coefficient estimator $\hat{B}$ does *not* require the estimated working correlation $\hat{R}$ to be a consistent estimator of the true correlation $R_0$. This protects us from potential misspecification of the intra-subject correlation structure. Such a robustness feature is well known for GEE estimator with vector-valued covariates. Theorem 1 confirms and extends this result to the tensor GEE case with image covariates. We also remark that, although the asymptotics of the classical GEE can in principle be generalized to the tensor data by directly vectorizing the coefficient array, the ultrahigh dimensionality of the parameters would have made the regularity conditions such as (A3) unrealistic. By contrast, Theorem 1 ensures that one could still enjoy the consistency and robustness properties, by taking into account the structural information of the tensor coefficient under the GEE framework. Under condition (A4), we define

$$\tilde{M}_n(B) = \sum_{i=1}^{n} [J_1, \ldots, J_D]^\top \operatorname{vec} X_i A_i^{1/2}(B) \hat{R}^{-1} R_0 \hat{R}^{-1} A_i^{1/2}(B) \operatorname{vec}^\top X_i [J_1, \ldots, J_D],$$

$$\tilde{D}_{n1}(B) = \sum_{i=1}^{n} [J_1, \ldots, J_D]^\top \operatorname{vec} X_i A_i^{1/2}(B) \hat{R}^{-1} A_i^{1/2}(B) \operatorname{vec}^\top X_i [J_1, \ldots, J_D].$$

As we show in the appendix, $\tilde{M}_n(B)$ approximates the covariance matrix of $s(B)$ in (2.5), while $\tilde{D}_{n1}(B)$ approximates the leading term of the negative gradient of $s(B)$ with respect to $\beta_{B}$. The next theorem establishes the asymptotic normality of the tensor GEE estimator.

**Theorem 2.** Under conditions (A1)-(A8), for any vector $b \in \mathbb{R}^{\sum_{d=1}^{D} p_d}$ such that $\|b\| = 1$, we have

$$b^\top \tilde{M}_n^{-1/2}(B_0) \tilde{D}_{n1}(B_0) (\beta_{\hat{B}} - \beta_{B_0}) \rightarrow \operatorname{Normal}(0, 1) \text{ in distribution.}$$
3.3. Rank selection consistency

Next we establish that the rank selected by BIC in (2.6) under the independent working correlation is a consistent estimator of the true rank. This result is useful in two ways. First, it justifies, to some extent, the asymptotic study in the previous section under a known rank. Second, it improves our understanding of the interaction between the working correlation and the rank specification. That is, the rank selected under a potentially misspecified correlation structure remains consistent. We also remark that, this rank selection consistency result is not available in Zhou et al. (2013), and our result is the first of its kind. For simplicity, we only consider the Gaussian linear model case, and leave the GLM case for future research.

We employ the same regularity conditions (A1)-(A8) in Section 3.2, except that we replace (A3) by the following condition:

(A3*) There exist two positive constants $c_1^* < c_2^*$ such that $c_1^* \leq \lambda_{\min}(I(B)) \leq \lambda_{\max}(I(B)) \leq c_2^*$, for all parameter points $B$ in the interior of the parameter space. In addition, the rank is constant over the set $\{B : \|\beta_B - \beta_{B_0}\| \leq \triangle n^{-1/2}\}$ for some $\triangle > 0$.

The reason for requiring (A3*) is that we need to characterize the behavior of some underfitted estimators with rank smaller than the true rank. These underfitted estimators may not reside in the neighborhood of the true parameters. We note that, however, (A3*) is a fairly mild condition, and the difference between (A3*) and (A3) is small. This is because, when the dimension is fixed, $I(B)$ has fixed dimensions, then the condition on the bounded eigenvalues is essentially requiring the matrix to be non-singular.

The next theorem establishes the rank selection consistency.

Theorem 3. Let $\hat{R} = \arg \min BIC(R)$, and $R_0 = \text{rank}(B_0)$. For the Gaussian linear model, under conditions (A1)-(A8) and the modified condition (A3*), we have

$$\Pr(\hat{R} = R_0) \to 1, \text{ as } n \to \infty.$$
overfitted model with a higher rank nor the underfitted model with an insufficient rank is favored by BIC.

3.4. Region selection consistency

Recall that, under the CP structure, the element in the coefficient tensor $B$ can be written as $\beta_{i_1 \ldots i_D} = \sum_{r=1}^{R} \beta_{1i_1}^{(r)} \times \cdots \times \beta_{Di_D}^{(r)}$. In the imaging application where $D = 3$, for example, the region at $(i_1, i_2, i_3)$ is non-active if $\beta_{i_1,i_2,i_3} = 0$. This can be induced if one of $\{\beta_{1i_1}^{(r)}, \beta_{2i_2}^{(r)}, \beta_{3i_3}^{(r)}\}$ is zero for each $r = 1, \ldots, R$. Therefore, correctly recovering the sparsity pattern of $\beta_B$ results in a selection of active regions of $B$. We next establish that, for the SCAD regularized tensor GEE in (2.7), this selection is consistent.

**Theorem 4.** Under conditions (A1)-(A8), $\rho_n = o(1)$ and $n^{-1/2} \log n = o(\rho_n)$, there exists one solution, $\hat{\beta}_B$, to the SCAD regularized tensor GEE such that

$$\Pr(\text{supp}(\beta_B) = \text{supp}(\beta_{B_0})) \to 1, \text{ as } n \to \infty,$$

where $\text{supp}(\beta)$ denotes the support of the vector $\beta$.

This theorem states that the support of true tensor coefficient, $\text{supp}(\beta_{B_0})$, can be recovered with high probability using the SCAD regularized tensor GEE, and as such establishes the region selection consistency in the context of tensor GEE.

4. Simulations

We have carried out intensive simulations to investigate the finite sample performance of our proposed tensor GEE approach. We adopt the following simulation setup. We generate the responses according to the normal linear model

$$Y_i \sim \text{MVN}(\mu_i, \sigma^2 R_0), \quad i = 1, \ldots, n,$$

where $Y_i = (Y_{i1}, \ldots, Y_{im})^T$, $\mu_i = (\mu_{i1}, \ldots, \mu_{im})^T$, $\sigma^2$ is a scale parameter, and $R_0$ is the true $m \times m$ intra-subject correlation matrix. We choose $R_0$ to be of an exchangeable (compound symmetric) structure with the off-diagonal coefficient $\rho_n = 0.8$. The mean function is of the form $\mu_{ij} = \gamma^T Z_{ij} + \langle B, X_{ij} \rangle$, $i = 1, \ldots, n, j = 1, \ldots, m$, where $Z_{ij} \in \mathbb{R}^5$ denotes the additional covariates, with all elements generated from a standard normal
distribution, and $\gamma \in \mathbb{R}^5$ is the corresponding coefficient vector, with all elements equal to one; $X_{ij} \in \mathbb{R}^{64 \times 64}$ denotes the 2D matrix covariate, again with all elements from standard normal, and $B \in \mathbb{R}^{64 \times 64}$ is the matrix coefficient. The entries of $B$ take the value of 0 or 1, and contains a series of shapes as shown in Figure 1 including “square”, “T-shape”, “disk”, “triangle”, and “butterfly”. Our goal is to recover those shapes in $B$ by inferring the association between $Y_{ij}$ and $X_{ij}$.

4.1. Signal recovery

As the true signal in reality is rarely of an exact low rank structure, the tensor GEE model essentially provides a low rank approximation to the true signal. Thus our first task is to verify if such an approximation is adequate, in that it can recover the true signal area and shape to a reasonable level. We set $n = 500$ and $m = 4$ and show the tensor GEE estimates and the corresponding BIC values under three working ranks $R = 1, 2, \text{ and } 3$ in Figure 1. We first assume that the correlation structure is correctly specified, and study potential misspecification in the next section. In this setup, “square” has the true rank equal to 1, “T-shape” has the rank 2, and the remaining shapes have a rank much larger than 3. It is clearly seen from Figure 1 that, the tensor GEE produces a reasonable recovery of the true signal, even for the signals with a high rank, e.g., “disk” and “butterfly”. All shapes can be clearly recognized, even though the surrounding area is gray and noisy. Moreover, the BIC criterion (2.6) successfully identifies the correct or best approximate the rank for all the signals.

4.2. Effect of correlation specification

We have shown that the tensor GEE estimator remains asymptotically consistent even when the working correlation structure is misspecified. However, this describes only the large sample behavior. In this section, we investigate potential effect of correlation misspecification when the sample size is small or moderate.

We choose the “butterfly” signal and fitted the tensor GEE model with three different working correlation structures: exchangeable, autoregressive of order one (AR-1), and independent. Table 1 reports the averages and standard errors (in parenthesis) out
Figure 1: True and recovered image signals by the tensor GEE with varying ranks. $n = 500, m = 4$. The correlation structure is correctly specified. TR($R$) means estimate from the rank-$R$ tensor model.
Table 1: Bias, variance, and MSE of the tensor GEE estimates under various working correlation structures. The result is based on 100 simulation replicates. The true intra-subject correlation is exchangeable with $\rho_n = 0.8$.

| $n$ | $m$ | Working Correlation | Bias$^2$ | Variance | MSE         |
|-----|-----|---------------------|----------|----------|-------------|
| 50  | 10  | Exchangeable        | 122.0    | 383.6    | **505.6(7.9)** |
|     |     | AR-1                | 139.1    | 530.0    | 669.1(15.8)  |
|     |     | Independence        | 119.1    | 393.9    | 513.0(11.0)  |
| 100 | 10  | Exchangeable        | 85.8     | 128.9    | **214.7(2.2)** |
|     |     | AR-1                | 88.0     | 159.1    | 247.1(3.0)   |
|     |     | Independence        | 93.0     | 141.2    | 234.2(2.8)   |
| 150 | 10  | Exchangeable        | 86.1     | 51.3     | **137.2(0.6)** |
|     |     | AR-1                | 85.6     | 56.0     | 141.6(0.6)   |
|     |     | Independence        | 84.9     | 62.3     | 147.2(0.9)   |

Of 100 simulation replicates of the squared bias, the variance, and the mean squared error (MSE) of the tensor GEE estimate. We observe that the estimator based on the correct working correlation structure, i.e., the exchangeable structure, performs better than those based on misspecified correlation structures. When the sample size is moderate ($n = 100$), all the estimators have comparable bias, while the difference in MSE mostly comes from the variance part of the estimator. This agrees with the theory that the choice of the working correlation structure affects the asymptotic variance of the estimator. When the sample size becomes relatively large ($n = 150$), all the estimators perform similarly by the scaling term of $n^{-1/2}$ on the variance. When the sample size is small ($n = 50$), all the estimators have relatively large bias, while the independence working structure yields similar results as the exchangeable structure. This suggests that, when the sample size is limited, using a simple independence working structure is probably preferable compared to a more complex correlation structure.

Nevertheless, we should bear in mind that the above observations are for the average behavior of the estimate. Figure 2 shows two snapshots of the estimated signals under the three working correlations with $n = 100$. The top panel is one replicate where the estimates are “close” to the average in the sense that the bias, variance and MSE values for this single data realization are similar to those averages reported in Table 1. Consequently, the visual qualities of the three recovered signals are similar. The bottom
Figure 2: Snapshots of tensor GEE estimation with different working correlation structures. The true correlation is an equicorrelated structure. The comparison is row-wise. The first row shows a replicate where the estimates are “close” to the average behavior, and thus the visual quality of the estimates under different correlations structures are similar. The second row shows a replicate where the estimates are “far away” from the average, then the estimate under the correct correlation structure (panel 1) is superior than those under incorrect structures.

panel, on the other hand, shows another replicate where the estimates are “far away” from the average. Then for this particular data, the quality of the estimated signal under the correct working correlation structure is superior than the ones under the incorrect specifications. Such an observation suggests that, as long as the sample size of the longitudinal imaging study is moderate to large, a longitudinal model should be favored over the one that totally ignores potential intra-subject correlation.

4.3. Regularized estimation and comparison

We next study the empirical performance of the regularized tensor GEE (denoted as “regularization”), and also compare with some alternative solutions: the tensor GEE without regularization (“no regularization”), the Lasso regularized vector GEE applied to the vectorized image predictor \cite{Fu2003}, “Fu-Lasso”, the SCAD regularized vector
GEE (Wang et al., 2012, “Wang-SCAD”), and the Sandwich Estimator (Guillaume et al., 2014, “SwE”). We adopt the same simulation setup as in Section 4.1, vary the sample size \( n \), and fix \( m = 4 \). For our regularized tensor GEE, we have implemented both the Lasso and SCAD penalty, and found their performances visually very similar. As such, we only report the results based on SCAD here. The penalty parameter is tuned based on an independent validation dataset. It is also noteworthy that, the Sandwich Estimator of Guillaume et al. (2014) treats the image as a response, which is different from our method that treats the image as a predictor. We have used the software provided by Guillaume et al. (2014) for its calculation. We have experimented with various shapes and obtained similar results. In the interest of space, we only report the results of “T-shape” and “butterfly” in Figures 3 and 4. In both cases, our regularized tensor GEE is seen to outperform the alternative solutions, especially when the sample size is limited.

4.4. Computation time

In this section, we investigate the computation time of our proposed tensor GEE. We consider the same simulation setup as in Section 4.1, but vary the sample size and the image dimension. First, we set \( m = 10 \), the matrix covariate of dimension 64 × 64, and increase \( n \) from 50 to 500 by an increment of 50. Second, we set \( n = 200 \), \( m = 4 \), and increase the matrix covariate dimension from 32 × 32 to 128 × 128 by an increment of 16. All simulations have been carried out on a laptop computer with Intel Xeon 2.60 GHz processor. Figure 3 reports the average computation time, in seconds, along with its confidence interval, based on 100 data replications for various signal shapes. Overall we have observed that the computation time of our method is reasonable.

5. Real Data Analysis

5.1. Alzheimer’s disease

Alzheimer’s Disease (AD) is a progressive and irreversible neurodegenerative disorder and the leading form of dementia in elderly subjects. It is characterized by gradual impairment of cognitive and memory functions, and it has been projected to quadruple in its prevalence by the year 2050 (Brookmeyer et al., 2007). Amnestic mild cognitive
Figure 3: Comparison of the tensor GEE with and without regularization, the Lasso regularized vector GEE (Fu, 2003, “Fu-Lasso”), the SCAD regularized vector GEE (Wang et al., 2012, “Wang-SCAD”), and the Sandwich Estimator (Guillaume et al., 2014, “SwE”). The sample size $n$ varies and $m = 4$. The matrix covariate is of size $64 \times 64$ and the true signal shape is “T-shape”.
Figure 4: Comparison of the tensor GEE with and without regularization, the Lasso regularized vector GEE (Fu 2003, “Fu-Lasso”), the SCAD regularized vector GEE (Wang et al. 2012, “Wang-SCAD”), and the Sandwich Estimator (Guillaume et al. 2014, “SwE”). The sample size $n$ varies and $m = 4$. The matrix covariate is of size $64 \times 64$ and the true signal shape is “butterfly”.
Figure 5: Computation time, in seconds, of the tensor GEE with varying sample size and image dimension for various signal shapes.

Impairment (MCI) is a prodromal stage to Alzheimer’s disease, and individuals with MCI may convert to AD at an annual rate as high as 15% (Petersen et al., 1999). There is a pressing need for accurate and early diagnosis of AD and MCI, and for monitoring the disease progression. The data we analyzed is obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI). It consists of $n = 88$ MCI subjects with longitudinal MRI images of white matter at baseline, 6-month, 12-month, 18-month and 24-month ($m = 5$). Also recorded was the Mini Mental State Examination (MMSE) score. It measures the orientation to time and place, the immediate and delayed recall of three words, the attention and calculations, language, and visuoconstructional functions (Folstein et al., 1975), and is our response variable. All MRI images have been preprocessed, and a detailed preprocessing protocol is given in Zhang et al. (2012). We recognize the importance of preprocessing in imaging analysis. In our study, we focus on the analysis after proper preprocessing, and we aim at two targets. One is to predict the future clinical scores based on the data at previous time points. Here the goal is not to use MRI to replace cognitive test, but instead, to better understand the association between brain structure and cognition as the disease progress. The second target is to identify brain subregions that are highly relevant to the disorder, so to better understand the disease pathology. We fit the proposed tensor GEE to this data. The rank is fixed at 3,
since this rank value has been shown to provide a reasonable tradeoff between dimension reduction and model flexibility (Zhou et al., 2013).

5.2. Prediction and disease prognosis

By averaging consecutive time points, we first downsize the original 256-dimensional MRI images to a smaller dimension, 32, 64, and 128-dimensional, respectively. This downsizing step is to sacrifice image resolution, but is to facilitate the computation and reduce the dimensionality. It is a tradeoff given the limited sample size and the huge number of unknown parameters. See Li et al. (2013a) for an alternative way of image downsizing. Next we consider two ways of evaluating the prediction accuracy.

We first use the data in early months to predict the “future” cognitive outcome in the last month of scan. This evaluation scheme is useful to understand disease progression, and has been often used in longitudinal imaging analysis, e.g., Zhang et al. (2012). Specifically, we fit the tensor GEE using the data of all subjects from baseline to 12-month, and use the prediction of MMSE at 18-month to select the tuning parameter. With the selected tuning parameter, we then refit the model using the data from baseline to 18-month, and finally evaluate the prediction accuracy of all subjects using the “future” MMSE score at 24-month, based on the rooted mean squared error (RMSE), \( \sqrt{\frac{1}{n} \sum_{i=1}^{n} (Y_{im} - \hat{Y}_{im})^2} \). Table 2 summarizes the results. It is seen that the MRI images of three different sizes yield similar results. The best RMSE achieved by our tensor GEE is 2.147 under an AR(1) working correlation structure, the SCAD penalty, and the downsized image dimension 32 \times 32 \times 32. This is only slightly worse than the best reported RMSE of 2.035 in Zhang et al. (2012). Note that Zhang et al. (2012) used multiple imaging modalities as well as additional biomarkers, which are supposed to improve the prediction accuracy, while our study utilized only one imaging modality.

We next consider the leave-one-out cross-validation evaluation, which is useful to understand the generalization capability across different individuals. Specifically, we leave all the scans of a single subject as the testing set, and fit the tensor GEE on the rest of the data as the training set. We tune the regularization parameter through a five-fold cross-validation on the training set. We evaluate the prediction accuracy using
Table 2: Prediction of the MMSE score at a “future” time for all subjects.

| Working Correlation | Independence | Equicorrelated | AR(1) | Unstructured |
|---------------------|--------------|----------------|-------|--------------|
| Image dimension 32 × 32 × 32 | | | | |
| regularization (Lasso) | 2.460 | 2.349 | 2.270 | 2.570 |
| regularization (SCAD) | 2.324 | 2.202 | **2.147** | 2.674 |
| no regularization | 2.526 | 2.427 | 2.429 | 2.628 |
| Image dimension 64 × 64 × 64 | | | | |
| regularization (Lasso) | 2.364 | **2.153** | 2.245 | 2.771 |
| regularization (SCAD) | 2.627 | 2.517 | 2.659 | 2.924 |
| no regularization | 4.490 | 4.154 | 4.776 | 3.749 |
| Image dimension 128 × 128 × 128 | | | | |
| regularization (Lasso) | 2.369 | 2.315 | **2.293** | 2.702 |
| regularization (SCAD) | 2.815 | 2.874 | 3.663 | 3.037 |
| no regularization | 6.805 | 5.008 | 4.036 | 7.979 |

the RMSE of the predicted MMSE score averaged across all months for the testing subject. Table 3 reports both the mean and standard deviation (in the parenthesis) of RMSE averaged across all the subjects. The best RMSE achieved by our tensor GEE is 3.172 again under an AR(1) working correlation structure, the SCAD penalty, and the downsized image dimension 32 × 32 × 32. This is slightly worse than the best RMSE in Table 2, as is expected. Meanwhile, the two tables have exhibited a consistent pattern, in that the tensor GEE with regularization outperforms the one without regularization.

5.3. Region selection

Next we investigate brain region selection using the regularized tensor GEE. We have applied both the Lasso and SCAD penalties, and due to graphical similarity of the results, we report the SCAD estimate only. Figure 6 shows the estimate (marked in red) overlaid on an image of an arbitrarily chosen subject, with three views, top, side and bottom, respectively. The identified anatomical regions mainly correspond to cerebral cortex, part of temporal lobe, parietal lobe, and frontal lobe (Braak and Braak, 1991; Desikan et al., 2009; Yao et al., 2012). With AD, patients experience significant widespread damage over the brain, causing shrinkage of brain volume (Yao et al., 2012; Harasy et al., 1999) and thinning of cortical thickness (Desikan et al.)
Table 3: Prediction of the MMSE score of a “future” subject at all times, using leave-one-out cross-validation.

| Working Correlation | Independence | Equicorrelated | AR(1) | Unstructured |
|---------------------|--------------|----------------|-------|--------------|
|                     | Image dimesion $32 \times 32 \times 32$ | | | |
| regularization (Lasso) | 3.225(1.851) | 3.404(1.710) | 3.272(1.886) | 3.982(2.557) |
| regularization (SCAD) | 3.250(1.928) | 3.392(1.624) | **3.172(1.551)** | 3.790(2.634) |
| no regularization | 4.271(2.936) | 4.063(2.571) | 4.415(3.186) | 4.492(3.294) |
|                     | Image dimesion $64 \times 64 \times 64$ | | | |
| regularization (Lasso) | 3.381(1.949) | 3.825(1.973) | 3.333(1.877) | 3.645(1.955) |
| regularization (SCAD) | **3.282(1.761)** | 3.414(1.723) | 3.592(2.166) | 3.873(1.937) |
| no regularization | 4.670(2.179) | 5.025(2.851) | 4.681(1.870) | 4.452(2.353) |
|                     | Image dimesion $128 \times 128 \times 128$ | | | |
| regularization (Lasso) | 3.409(1.743) | 3.968(1.850) | **3.296(1.983)** | 3.301(1.574) |
| regularization (SCAD) | 4.123(2.326) | 3.929(1.895) | 3.696(2.065) | 3.780(1.862) |
| no regularization | 5.605(3.716) | 5.532(4.713) | 5.654(2.969) | 6.037(9.493) |

The affected brain regions include those involved in controlling language (Broca’s area) (Harasty et al., 1999), reasoning (superior and inferior frontal gyri) (Harasty et al., 1999), part of sensory area (primary auditory cortex, olfactory cortex, insula, and operculum) (Braak and Braak, 1991; Lee et al., 2013), somatosensory association area (Yao et al., 2012; Tales et al., 2005; Mapstone et al., 2003), memory loss (hippocampus) (den Heijer et al., 2010), and motor function (Buchman and Bennett, 2011). It is noted that these regions are affected starting at different stages of AD, indicating the capability of the proposed method to locate brain atrophies as the disease progresses. Specifically, hippocampus, which is highly correlated to memory loss, is commonly detected at the earliest stage of the disease. Regions related to language, communication, and motor functions are normally detected at the later stages of the disease. The fact that our findings are consistent with results reported in previous studies, particularly the longitudinal studies, demonstrates the efficacy of our proposed method in identifying correct biomarkers that are closely related to AD and MCI.

6. Discussions

In this article, we have proposed tensor GEE for longitudinal imaging analysis. With
the increasing availability of longitudinal imaging data and the relative paucity of effective analytical solutions, our proposal provides a timely and useful response. It combines the GEE approach for handling longitudinal correlation, and the low rank decomposition for vast dimension reduction and tensor structure preservation. The proposed algorithm scales with imaging data size and is easy to implement using existing statistical softwares. Simulation studies and real data analysis have shown promise of our method for the purpose of both signal recovery and outcome prediction.

Our method involves two specifications, the working correlation structure and the working rank of the tensor coefficient. We make a few remarks on their selection in practice, and potential consequence of their misspecification. For the working correlation structure, our study has shown that, asymptotically, the tensor GEE estimator remains consistent even if the correlation structure is misspecified. In practice, our numerical investigation has suggested that, a simple independence working correlation is probably preferable when the sample size is limited, while a data adaptive choice of a suitable working correlation is preferable when the sample size is ample. This is useful, since a good number of multi-center large scale longitudinal imaging datasets such as ADNI are currently emerging. The same correlation structure selection problem is also encountered in the classical vector-valued GEE, and more discussion can be found in Pan and Connett (2002). For the working rank of the tensor CP decomposition, we again show that, asymptotically, the BIC criterion under the independent correlation structure could
select the true rank with probability approaching one, even if this correlation structure is misspecified. In practice, the rank selection reflects a bias-variance tradeoff. When the selected rank is smaller than truth, the resulting estimator is biased, but involves a smaller number of unknown parameters, and thus is less variable. When the selection is greater than the truth, the estimator becomes unbiased, but is also more variable with a larger number of parameters. In general, our experiences suggest that the reduced rank structure provides a reasonable approximation of the coefficient tensor.

Numerous problems remain open and warrant for future research. One is rank selection, including the selection consistency for a more general family of models, its convergence rate, and its selection under a diverging dimension. We note that this problem is not yet fully solved even in the context of tensor predictor regression on a single image observation per subject, and is particularly challenging. The other is asymptotic study of our tensor GEE with a diverging dimension. This is important to improve our understanding of the properties of tensor GEE. We have obtained some preliminary results, extending those of the vector GEE (Wang, 2011) to the tensor version. However, the asymptotic properties under a diverging dimension intertwine with the diverging rank selection, and a comprehensive study warrants substantial future research.

Supplementary Materials

The proofs of the main theorems and some technical lemmas are available in the online supplementary material. A Matlab software package is available upon request.

Acknowledgements

The work is partially supported by NSF grants DMS-1310319, DMS-1613137, and NIH grants HG006139, GM105785, GM53275, and AG034570.

References

Aston, J. A. D., Pigoli, D., and Tavakoli, S. (2017). Tests for separability in nonparametric covariance operators of random surfaces. *The Annals of Statistics, 45*(4):1431–1461.

Balan, R. M. and Schiopu-Kratina, I. (2005). Asymptotic results with generalized estimating
equations for longitudinal data. *The Annals of Statistics*, 33(2):522–541.

Braak, H. and Braak, E. (1991). Neuropathological stageing of Alzheimer-related changes. *Acta Neuropathologica*, 82(4):239–259.

Brookmeyer, R., Johnson, E., Ziegler-Graham, K., and Arrighi, H. M. (2007). Forecasting the global burden of Alzheimers disease. *Alzheimer's & Dementia*, 3(3):186 – 191.

Buchman, A. and Bennett, D. (2011). Loss of motor function in preclinical Alzheimers disease. *Expert Review Neurotherapeutics*, 11(5):665–676.

Davatzikos, C., Xu, F., An, Y., Fan, Y., and Resnick, S. M. (2009). Longitudinal progression of Alzheimer's-like patterns of atrophy in normal older adults: the spare-ad index. *Brain*, 132(8):2026–2035.

den Heijer, T., van der Lijn, F., Koudstaal, P. J., Hofman, A., van der Lugt, A., Krestin, G. P., Niessen, W. J., and Breteler, M. M. B. (2010). A 10-year follow-up of hippocampal volume on magnetic resonance imaging in early dementia and cognitive decline. *Brain*, 133(4):1163–1172.

Desikan, R., Cabral, H., Hess, C., Dillon, W., Salat, D., Buckner, R., Fischl, B., and Initiative, A. D. N. (2009). Automated MRI measures identify individuals with mild cognitive impairment and Alzheimer's disease. *Brain*, 132(3):2048–2057.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Fan, J. and Li, R. (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *Journal of the American Statistical Association*, 99(467):710–723.

Folstein, M. F., Folstein, S. E., and McHugh, P. R. (1975). Mini-mental state: A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 12(3):189 –198.

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

Guillaume, B., Hua, X., Thompson, P. M., Waldorp, L., Nichols, T. E., Initiative, A. D. N., et al. (2014). Fast and accurate modelling of longitudinal and repeated measures neuroimaging data. *NeuroImage*, 94:287–302.

Harasty, J. A., Halliday, G. M., Kril, J. J., and Code, C. (1999). Specific temporoparietal gyral atrophy reflects the pattern of language dissolution in Alzheimer’s disease. *Brain*, 122(4):675–686.
Hinrichs, C., Singh, V., Xu, G., and Johnson, S. C. (2011). Predictive markers for AD in a multi-modality framework: An analysis of MCI progression in the ADNI population. *NeuroImage*, 55(2):574 – 589.

Kolda, T. G. and Bader, B. W. (2009). Tensor decompositions and applications. *SIAM Review*, 51(3):455–500.

Lee, T. M., Sun, D., Leung, M.-K., Chu, L.-W., and Keysers, C. (2013). Neural activities during affective processing in people with Alzheimer’s disease. *Neurobiology of Aging*, 34(3):706 – 715.

Li, B. (1997). On the consistency of generalized estimating equations. In *Selected Proceedings of the Symposium on Estimating Functions (Athens, GA, 1996)*, volume 32 of *IMS Lecture Notes Monograph Series*, pages 115–136. Institute of Mathematical Statistics, Hayward, CA.

Li, X., Zhou, H., and Li, L. (2013a). Tucker tensor regression and neuroimaging analysis. *arXiv preprint arXiv:1304.5637*.

Li, Y., Gilmore, J. H., Shen, D., Styner, M., Lin, W., and Zhu, H. (2013b). Multiscale adaptive generalized estimating equations for longitudinal neuroimaging data. *NeuroImage*, 72:91 – 105.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

Mapstone, M., Steffenella, T., and Duffy, C. (2003). A visuospatial variant of mild cognitive impairment: getting lost between aging and AD. *Neurology*, 60:802–808.

McEvoy, L. K., Holland, D., Hagler, D. J., Fennema-Notestine, C., Brewer, J. B., and Dale, A. M. (2011). Mild cognitive impairment: Baseline and longitudinal structural MR imaging measures improve predictive prognosis. *Radiology*, 259(3):834–843.

Misra, C., Fan, Y., and Davatzikos, C. (2009). Baseline and longitudinal patterns of brain atrophy in MCI patients, and their use in prediction of short-term conversion to AD: Results from ADNI. *NeuroImage*, 44(4):1415 – 1422.

Ni, X., Zhang, D., and Zhang, H. H. (2010). Variable selection for semiparametric mixed models in longitudinal studies. *Biometrics*, 66(1):79–88.

Pan, W. (2001). Akaike’s information criterion in generalized estimating equations. *Biometrics*, 57(1):120–125.
Pan, W. and Connett, J. E. (2002). Selecting the working correlation structure in generalized estimating equations with application to the lung health study. *Statistica Sinica*, 12(2):475–490.

Petersen, R., Smith, G., Waring, S., Ivnik, R., Tangalos, E., and Kokmen, E. (1999). Mild cognitive impairment: clinical characterization and outcome. *Archives of Neurology*, 56:303–308.

Prentice, R. L. and Zhao, L. P. (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics*, 47(3):825–839.

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

Rao, C. R. and Mitra, S. K. (1971). *Generalized Inverse of Matrices and its Applications*. John Wiley & Sons, Inc., New York-London-Sydney.

Raskutti, G. and Yuan, M. (2016). Convex regularization for high-dimensional tensor regression. *arXiv*.

Skup, M., Zhu, H., and Zhang, H. (2012). Multiscale adaptive marginal analysis of longitudinal neuroimaging data with time-varying covariates. *Biometrics*, 68(4):1083–1092.

Song, P. X.-K., Jiang, Z., Park, E., and Qu, A. (2009). Quadratic inference functions in marginal models for longitudinal data. *Statistics in Medicine*, 28(29):3683–3696.

Sun, W. and Li, L. (2016). Sparse tensor response regression and neuroimaging analysis. *arXiv preprint arXiv:1609.04523*.

Sun, W., Lu, J., Liu, H., and Cheng, G. (2017). Provable sparse tensor decomposition. *Journal of the Royal Statistical Society, Series B*, 79(3):899–916.

Tales, A., Haworth, J., Nelson, S., J. Snowden, R., and Wilcock, G. (2005). Abnormal visual search in mild cognitive impairment and Alzheimer’s disease. *Neurocase*, 11(1):80–84.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B.*, 58(1):267–288.

Wang, L. (2011). GEE analysis of clustered binary data with diverging number of covariates. *The Annals of Statistics*, 39(1):389–417.

Wang, L., Zhou, J., and Qu, A. (2012). Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics*, 68(2):353–360.
Xie, M. and Yang, Y. (2003). Asymptotics for generalized estimating equations with large cluster sizes. *The Annals of Statistics*, 31(1):310–347.

Xue, L., Qu, A., and Zhou, J. (2010). Consistent model selection for marginal generalized additive model for correlated data. *Journal of the American Statistical Association*, 105(492):1518–1530.

Yao, Z., Hu, B., Liang, C., Zhao, L., Jackson, M., and the Alzheimer’s Disease Neuroimaging Initiative (2012). A longitudinal study of atrophy in amnestic mild cognitive impairment and normal aging revealed by cortical thickness. *PLoS One*, 7(11):e48973.

Zhang, D., Shen, D., and Alzheimer’s Disease Neuroimaging Initiative (2012). Predicting future clinical changes of MCI patients using longitudinal and multimodal biomarkers. *PLoS One*, 7(3):e33182.

Zhang, D., Wang, Y., Zhou, L., Yuan, H., Shen, D., and the Alzheimers Disease Neuroimaging Initiative (2011). Multimodal classification of Alzheimer’s disease and mild cognitive impairment. *Neuroimage*, 55(3):856 – 867.

Zhou, H. and Li, L. (2014). Regularized matrix regression. *Journal of the Royal Statistical Society. Series B*, 76:463–483.

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.

Department of Statistics, North Carolina State University, Raleigh, NC 27697, USA.
E-mail: xzhang23@ncsu.edu

Division of Biostatistics, University of California, Berkeley, CA 94720, USA.
E-mail: lexinli@berkeley.edu

Department of Biostatistics, University of California, Los Angeles, CA 90095, USA.
E-mail: huazhou@ucla.edu

Division of Biostatistics, University of California, Berkeley, CA 94720, USA.
E-mail: yqzhou@berkeley.edu

Department of Radiology, University of North Carolina, Chapel Hill, NC 27599, USA.
E-mail: dinggang.shen@med.unc.edu