Longitudinal LASSO: Jointly Learning Features and Temporal Contingency for Outcome Prediction

Tingyang Xu
Department of Computer Science and Engineering
University of Connecticut
Storrs, CT, USA
tix11001@engr.uconn.edu

Jiangwen Sun
Department of Computer Science and Engineering
University of Connecticut
Storrs, CT, USA
javon@engr.uconn.edu

Jinbo Bi [*]
Department of Computer Science and Engineering
University of Connecticut
Storrs, CT, USA
jinbo@engr.uconn.edu

ABSTRACT

Longitudinal analysis is important in many disciplines, such as the study of behavioral transitions in social science. Only very recently, feature selection has drawn adequate attention in the context of longitudinal modeling. Standard techniques, such as generalized estimating equations, have been modified to select features by imposing sparsity-inducing regularizers. However, they do not explicitly model how a dependent variable relies on features measured at proximal time points. Recent graphical Granger modeling can select features in lagged time points but ignores the temporal correlations within an individual's repeated measurements. We propose an approach to automatically and simultaneously determine both the relevant features and the relevant temporal points that impact the current outcome of the dependent variable. Meanwhile, the proposed model takes into account the non-*i.i.d* nature of the data by estimating the within-individual correlations. This approach decomposes model parameters into a summation of two components and imposes separate block-wise LASSO penalties to each component when building a linear model in terms of the past τ measurements of features. One component is used to select features whereas the other is used to select temporal contingent points. An accelerated gradient descent algorithm is developed to efficiently solve the related optimization problem with detailed convergence analysis and asymptotic analysis. Computational results on both synthetic and real world problems demonstrate the superior performance of the proposed approach over existing techniques.

Categories and Subject Descriptors

G.1.6 [**Numerical Analysis**]: Optimization—*Gradient methods*; H.2.8 [**Database management**]: Database Application—*Data mining*

General Terms

Algorithms, Performance, Experimentation

Keywords

Longitudinal modeling; regularization methods; sparse predictive modeling; regression

1. INTRODUCTION

A longitudinal study collects and analyzes repeated measurements of a set of features for a group of subjects through time. Longitudinal analyses are important in many areas, such as in social and behavioral science [20, 7, 4], in economics [18, 2], in climate[13, 2], and in genetics [21]. For example, to predict binge drinking of college students, a longitudinal study may be designed to monitor them weekly or even daily in terms of multiple covariates, such as, the level of stress, status of negative affects and social behaviors [4, 1]. The fluctuation of these covariates is used to analyze and predict binge drinking (the dependent or outcome variable) of a student at the current observation time point. Changes of the covariates in the proximal time points are anticipated to alter the likelihood that a student binge drinks at the current observation point. To precisely understand how covariates affect the outcome, the analysis has to model not only the current values of the covariates but also their proximal values as well as take into account the correlation structure in the repeated measurements.

Typically, longitudinal data are analyzed by extending generalized linear models (GLM) with different assumptions, such as marginal models, random effects models, and transition models [6]. For example, a marginal model regresses the outcome on the current observation of features but factors in a within-subject correlation matrix that is estimated for a few proximal time points. In contrast, a random effects model reflects the variability among individuals rather than the population average comparing with marginal models. For marginal modeling, generalized estimating equations (GEE) are the most widely used methods which estimate a predictive model to predict the current outcome together with correlations among different outcomes observed temporally. The resultant predictive models are generally more accurate than those of classic regression analysis that assumes independently and identically distributed (*i.i.d.*) observations [12]. Research on feature selection in longitudinal data leads to a new family of methods based on the

[*] Correpsondence should be adressed to Jinbo Bi.

KDD '15, August 11 - 14, 2015, Sydney, NSW, Australia
© 2015 ACM. ISBN 978-1-4503-3664-2/15/08 ...$15.00.
DOI: http://dx.doi.org/10.1145/2783258.2783403.

penalized GEE (PGEE)[8]. For random effects models, generalized linear mixture model(GLMM)[11, 15] is the major method. It explores natural heterogeneity across individuals in the regression coefficients and represents this heterogeneity by a probability distribution.

None of those extensions of GLM aim to detect causal relationships from temporal changes of covariates to the outcomes of the current effect. In many studies, it is however necessary and insightful to model simultaneously the correlation among outcome records and the lagged causal effects of covariates [1]. For example, psychologists have identified that there is lagged effect in the alcohol use behavior. An individual's drinking today may be a response to an elevated level of stress two days back rather than the current day. It is actually an important question for psychologists to find out both which temporal points and which covariates influence the current outcome the most. This lagged effect is not used by temporal marginal modeling to make predictions.

On the other hand, researchers have developed machine learning approaches for longitudinal analysis that predict an outcome using feature values at multiple time points [2, 13]. For example, graphical Granger modeling [2], and grouped graphical Granger modeling[13] are insightful to explore the influences from past temporal information present in time series data in the modeling and understanding of the causal relationships. These methods assume that past values of certain time series features causally affect an outcome variable, and hence construct a model based on these values to predict future outcomes. Often, they estimate causality relationship (causal graph) among all features. However, these methods assume *i.i.d.* samples which are clearly violated in longitudinal data, and moreover they are incapable of selecting the most influential time points.

All existing methods either assume *i.i.d.* samples in Granger causality modeling or assume correlated samples but do not model *temporal* causal effects. Therefore, we propose a new learning formulation that constructs predictive models as functions of covariants not only from the current observation but also from multiple previous consecutive observations, and simultaneously determine the temporal contingency and the most influential features. The proposed method has the following advantages:

1. The proposed method makes predictions based on lagged data from current and previous time points. It decomposes the model coefficients into a summation of two components and impose different block-wise *least absolute shrinkage and selection operators* (LASSO) to the two components. One regularizer is used to detect the contingency of specific time points whereas the other is used to select covariates.

2. The proposed method also learns simultaneously a structured correlation matrix from the data. The correlations among the outcomes themselves imply the changing trend of the outcomes in the proximal time points within each subject.

3. We develop a family of methods where the outcome variable is assumed to follow a distribution from the exponential family, including Bernoulli, Gaussian and Poisson distributions. The formulations for these distributions are discussed in Section 3.3.

4. We provide the convergence analysis in Section 3.1 and asymptotic analysis in Section 3.2 to show that the proposed algorithm can find the optimal solution for the predictive models.

We have empirically compared the proposed method against the state of the art on both synthetic and real world datasets. The computational results demonstrate the effectiveness and the capability of our approach.

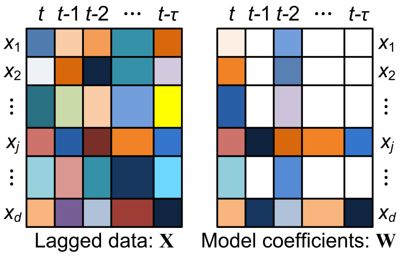

Figure 1: The outcome y_t at time t can be relevant to multiple covariates $x_1, x_2, \cdots, x_d$ observed at current and several previous time points $t-1, t-2, \cdots, t-\tau$, which forms a data matrix $\mathbf{X}$ (left). If we associate with each entry of this matrix a weight in our additive prediction model, then our model coefficients form a matrix $\mathbf{W}$ (right). If the coefficient matrix is sparse, then the resultant model will be selective in terms of covariates and time points.

2. METHOD

In our approach, the predictive model takes the form of the *trace* of the product of the lagged data $\mathbf{X}$ and the model coefficient matrix $\mathbf{W}$ as shown in Figure 1. The model coefficients are organized into a matrix rather than a vector used in traditional analysis because this way reflects the structure in the lagged data. Note that the lagged observations of y can also be included in the data matrix $\mathbf{X}$ to be used in the predictive model. For notational convenience, we just use $\mathbf{X}$ to represent the data that are used to form the model.

We first briefly review two most relevant sets of longitudinal analytics in Section 2.1 which will help elucidate the advantages of our proposed formulation.

2.1 Preliminaries

We introduce the notation that is used through out the paper. A bold lower case letter denotes a vector, such as $\mathbf{v}$. The $\|\mathbf{v}\|_p$ refers to the ℓ_p norm of a vector $\mathbf{v}$, which is formed as $\|\mathbf{v}\|_p = (\sum_{i=1}^d |v_i|^p)^{1/p}$, where v_i is the i-th component of $\mathbf{v}$ and d is the length of $\mathbf{v}$. A bold upper case letter denotes a matrix such as $\mathbf{M}$. Similarly, $\mathbf{m}_{(i,)}$, $\mathbf{m}_{(,j)}$ and m_{ij} represent the i-th row, j-th column and (i,j)-th component of $\mathbf{M}$, respectively. The Frobenius norm and $\ell_{p,q}$ norm of a matrix $\mathbf{M}$ refer, respectively, to $\|\mathbf{M}\|_F$, which is equal to $(tr(\mathbf{M}^\top \mathbf{M}))^{1/2}$, and $\|\mathbf{M}\|_{p,q}$, defined by $\left(\sum_{i=1}^n \left(\|\mathbf{m}_{(i,)}\|_q\right)^p\right)^{1/p}$, where n is the number of rows in $\mathbf{M}$, and $tr(\mathbf{M})$ indicates the trace of $\mathbf{M}$. We assume that vect$(\mathbf{M})$ is the column-major vectorization of $\mathbf{M}$, which is defined as vect$(\mathbf{M}) = (\mathbf{m}_{(,1)}^\top, \cdots, \mathbf{m}_{(,k)}^\top)^\top$ assuming k columns are in $\mathbf{M}$. Then, $\langle \mathbf{M}_1, \mathbf{M}_2 \rangle$ is the inner product of two matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ that is computed as the inner product of vect$(\mathbf{M}_1)$ and vect$(\mathbf{M}_2)$. The operator reshape$(\mathbf{v})$ re-shapes $\mathbf{v}$ into a matrix of a proper size determined by the specific context.

Assume that we are given data of m number of individuals on d number of features (independent variables) that are repeatedly measured at n_i time points for each individual i. The data of each individual i is represented by a matrix $\mathbf{X}^{(i)}$ of size $d \times n_i$, and $\mathbf{x}_t^{(i)}$ refers to the d-entry data vector of individual i at time point t. Without loss of generality, we assume that all individuals have data at the same consecutive time points ($n_i = n$) to simplify the notation and the subsequent analysis. Data on the dependent variable (outcome) is also given in $\mathbf{y}^{(i)}$ of length n that contains the observations at the n time points for individual i. Typically, a longitudinal study aims to estimate the effect of covariates on the dependent variable.

2.1.1 *Granger Causality*

The notion of *Granger Causality* was introduced by the Nobel prize winning economist, Clive Granger, and has proven useful in time series analysis [10]. It is based on the intuition that if a time series variable causally affects another, the past observations of the former should be useful in predicting the future outcome of the latter.

Specifically, a time series observation x is said to *Granger cause* another time series outcome, y, if the regressing for y in terms of past y and x is significantly better than the regressing just with past values of y. The so-called Granger test first performs two regressions:

$$y_t^{(i)} = \sum_{j=1}^{\tau} \left(a_j y_{t-j}^{(i)} + w_j^\top x_{t-j}^{(i)} \right), \tag{1}$$

and $y_t^{(i)} = \sum_{j=1}^{\tau} a_j y_{t-j}^{(i)}$, where τ is the maximum "lag" in the past observations, and then uses a hypothesis test such as an F-test to determine if the outcome y_t can be predicted significantly better from the past covariate x. Recent graphical Granger models [2, 13] extend it from a single time series covariate $\mathbf{x}$ to multiple covariates $\mathbf{X}$. They learn the coefficients $\mathbf{a}$ and $\mathbf{w}$'s with LASSO type of regularizers and evaluate if coefficients are non-zero for Granger causality.

2.1.2 *Generalized Estimating Equations (GEE)*

GEE estimates the parameters of a GLM while taking into account the correlations in the training examples. Similar to GLM, it assumes that the dependent variable comes from a class of distributions known as the exponential family. For each member in this family, there exists a link function that can be used to translate the nonlinear model into a linear model. The expectation of the outcome $y_t^{(i)}$ for subject i at time t is computed as:

$$E(y_t^{(i)}) = \mu_t^{(i)} = g^{-1}(\eta_t^{(i)}), \tag{2}$$

where $\mu_t^{(i)}$ represents the mean model, g^{-1} is the inverse of a link function g in a GLM [14], and $\eta_t^{(i)} = \left(\mathbf{x}_t^{(i)}\right)^\top \mathbf{w}$. The variance of $y_t^{(i)}$ is computed as $\text{var}(y_t^{(i)}) = \text{var}(\mu_t^{(i)})/\phi$ where ϕ is a scaling parameter that may be known or estimated.

GEE presumes a so-called working correlation structure, typically denoted by $\mathbf{R}(\boldsymbol{\alpha})$, where $\boldsymbol{\alpha}$ is a parameter to be determined from data. The common choices of $\mathbf{R}(\boldsymbol{\alpha})$ include exchangeable, tri-diagonal and the first-order autoregressive (AR(1)) formula [12]. The exchangeable correlation structure, also called *equi-correlation*, assumes that $corr(y_{it}, y_{it'}) = \alpha$ for all $t \neq t'$. The tri-diagonal structure uses a tridiagonal matrix as $\mathbf{R}(\boldsymbol{\alpha})$ where $corr(y_{it}, y_{it'}) = \alpha$ if $t' = t \pm 1$ or 0 otherwise. The AR(1) formula assumes a correlation structure along continuous time, and uses $corr(y_{it}, y_{it'}) = \alpha^{|t-t'|}$.

To estimate the regression coefficients $\mathbf{w}$, GEE uses the the estimating equations that are formulated, in general, by setting the derivative of an appropriate loss function to 0. Although a loss function may not be explicitly written out, the estimating equations always can be computed by

$$EE(\mathbf{w}, \boldsymbol{\alpha}) = \sum_{i=1}^{m} \left(\mathbf{D}^{(i)}\right)^\top \left(\mathbf{\Sigma}^{(i)}\right)^{-1} \mathbf{s}^{(i)} = 0. \tag{3}$$

where the $n \times d$ matrix $\mathbf{D}^{(i)} = \partial \boldsymbol{\mu}^{(i)} / \partial \mathbf{w}$ where $\boldsymbol{\mu}^{(i)}$ combines all $\mu_t^{(i)}, \forall t = 1, \cdots, n$ into a vector, $\mathbf{s}^{(i)} = \mathbf{y}^{(i)} - \boldsymbol{\mu}^{(i)}(\mathbf{w})$. The $n \times n$ matrix $\mathbf{\Sigma}^{(i)}$ is the estimated covariance structure as:

$$\mathbf{\Sigma}^{(i)}(\boldsymbol{\alpha}) = \left(\mathbf{A}^{(i)}\right)^{1/2} \mathbf{R}(\boldsymbol{\alpha}) \left(\mathbf{A}^{(i)}\right)^{1/2} / \phi \tag{4}$$

where $\mathbf{A}^{(i)}$ is an $n \times n$ diagonal matrix with $\text{var}(\mu_t^{(i)})$ as the t-th diagonal element. Algorithms are given in [12] to compute $\mathbf{w}$ and $\boldsymbol{\alpha}$ for the different choices of $\mathbf{R}(\boldsymbol{\alpha})$.

2.2 The Proposed Formulation

In our approach, each training example consists of the current and τ previous records of the repeated measurements. Let

$$\mathbf{X}_{(i;t)} = [\mathbf{x}_t^{(i)}, \mathbf{x}_{t-1}^{(i)}, \cdots, \mathbf{x}_{t-\tau}^{(i)}]$$

be a $d \times (\tau + 1)$ data matrix for subject i. Given T total measurements for each subject, the index t of $\mathbf{X}_{(i;t)}$ starts from $\tau + 1$ in order to have enough previous observations in the first training example. Hence, there are totally $n = T - \tau$ training examples for each subject. If $\mathbf{X}_{(i;t)}$ includes previous $\tau + 1$ values of $y^{(i)}$ as a feature, then the model $y_t^{(i)} = tr\left(\mathbf{X}_{(i;t)}^\top \mathbf{W}\right)$ where $\mathbf{W} = [\mathbf{w}_0, \mathbf{w}_1, \cdots, \mathbf{w}_\tau]$ essentially gives the same model like Eq.(1) in the graphical Granger models.

The Granger models would assume that the training examples are *i.i.d.*. However, the consecutive examples are not mutually independent because they contain overlapping records (e.g., $\mathbf{X}_{(i;t)}$ and $\mathbf{X}_{(i;t+1)}$ share $\tau - 1$ records $\mathbf{x}_t^{(i)}, \cdots, \mathbf{x}_{t-\tau+1}^{(i)}$). GEE provides a mechanism to estimate the sample correlation simultaneously while constructing predictive models, and to extend the linear models to generalized linear models. To apply GEE to our model, we replace $\eta_t^{(i)}$ used in GEE by the following formula

$$\eta_t^{(i)} = tr\left(\mathbf{X}_{(i;t)}^\top \mathbf{W}\right). \tag{5}$$

Substituting Eq.(5) for η in Eq.(2) yields a formulation similar to GEE. The regression coefficients $\mathbf{W}$ can be estimated through the well-developed GEE estimators. In particular, the quasi-likelihood methods of GEE estimate $\mathbf{W}$ by minimizing a loss function that is defined via the model deviance. The model deviance measures the difference between the log-likelihood of the estimated mean model $\boldsymbol{\mu}^{(i)}$ and that of the observed values $\mathbf{y}^{(i)}$. For instance, the model deviance for a linearly regressive response is written by $Dev^{(i)}(\mathbf{W}, \boldsymbol{\alpha}) = (\mathbf{y}^{(i)} - \boldsymbol{\mu}^{(i)})^\top \mathbf{R}(\boldsymbol{\alpha})(\mathbf{y}^{(i)} - \boldsymbol{\mu}^{(i)})$ where $\mathbf{y}^{(i)}$ contains the observed responses for subject i, and $\boldsymbol{\mu}^{(i)}$ is the estimated expectations of y for subject i. If the response follows an arbitrary distribution, the model deviance may not correspond to an explicit function. For the exponential family, it takes a special form as discussed in Theorem 1 below, which is still complicated. We denote by $Dev^{(i)}(\mathbf{W}, \boldsymbol{\alpha})$ the deviance occurred on subject i. GEE minimizes a loss function of $\sum_{i=1}^{m} Dev^{(i)}(\mathbf{W}, \boldsymbol{\alpha})$ for the optimal $\mathbf{W}$ by solving the *estimating equations*, i.e., taking the derivatives of the loss function and setting them to 0.

Now, to select among features and discover the most influential time points in predicting y over time, (and also to control the model capacity,) we apply regularizers to the model parameters. We first decompose $\mathbf{W}$ into a summation of two components as $\mathbf{W} = \mathbf{U} + \mathbf{V}$ and apply different regularizers to $\mathbf{U}$ and $\mathbf{V}$. The block-wise LASSO, such as the $\ell_{1,2}$ matrix norm, is widely-used in multi-task learning or feature selection with group structures, but has not been explored within the GEE setting. To the best of our knowledge, it has not been studied in longitudinal analytics how to produce shrinkage effects simultaneously on both features and contingent temporal records through proper regularization. The general $\ell_{1,p}$ matrix norm [23] calculates the sum of the ℓ_p norms of the rows in a matrix. Regularizers based on the $\ell_{1,p}$ norms encourage row sparsity by shrinking the entire rows to have zero entries.

In our parameter matrix $\mathbf{W}$, rows correspond to features and columns correspond to the observation time points. If we apply the $\ell_{1,2}$ norm to $\mathbf{U}$ (row-wisely), the optimal solution of $\mathbf{U}$ will contain rows with all zero entries. Thus, a selected subset of features in the $\tau + 1$ observations will be used in the predictive model to predict the current outcome. The $\ell_{1,2}$ norm of $\mathbf{V}^\top$ (column-wisely) encourages to select among columns of $\mathbf{V}$. If the k-th column of $\mathbf{V}$ contains the largest values in the selected columns, the current outcome is most contingent on the previous $(k-1)$-th record, thus having the $(k-1)$ "lagged" effect. Overall, we solve the following optimization problem for the best model parameters $\mathbf{W}$ which is computed as $\mathbf{U} + \mathbf{V}$:

$$\min_{\mathbf{U},\mathbf{V}} \quad \sum_{i=1}^{m} Dev^{(i)}(\mathbf{U} + \mathbf{V}, \boldsymbol{\alpha}) + \lambda_1 \|\mathbf{U}\|_{1,2} + \lambda_2 \|\mathbf{V}^\top\|_{1,2} \quad (6)$$

where $\mathbf{W}$ in the deviance is simply replaced by $\mathbf{U} + \mathbf{V}$.

The optimization of Eq.(6) is challenging. In general, even solving the GEE formulation is not easy as it estimates not only the model expectation but also the variance term $\mathbf{\Sigma}^{(i)}$. The algorithm that solves the GEE (i.e., the estimating equations) applies the Newton-Raphson method in the iterative reweighted least squares (IRLS) procedure [8] to estimate $\mathbf{w}$ and $\mathbf{\Sigma}^{(i)}$. However, this method does not solve any formula that uses regularizers. By modifying the Newton-Raphson method or shooting algorithm [8], it can be extended only to the regularizers that are decomposable into individual parameters w_j. For instance, the ℓ_1 vector norm of $\mathbf{w}$ can be decomposed into the summation of individual $|w_j|$, $j = 1, \cdots, d$. The $\ell_{1,2}$ matrix norm, unfortunately, can not be decomposed in such a way. Therefore, we have developed an accelerated gradient descent method based on the fast iterative shrinkage-thresholding algorithm (FISTA) [3]. Further, the following theorem shows that Eq.(6) is a convex optimization problem in terms of $\mathbf{W}$. Our algorithm can be proved to find the global optimal solution $\mathbf{W}$ of Eq.(6) when $\boldsymbol{\alpha}$ is fixed (to a consistent estimate given by GEE).

THEOREM 1. *The first term of Eq.(6) is convex and continuously differentiable with respect to* $\mathbf{U}$ *and* $\mathbf{V}$ *if the distribution of* $\mathbf{y}^{(i)}$ *is in a natural exponential family and the link function is continuous.*

PROOF. First, let us recall that the probability density function of a distribution in the exponential family takes the following form:

$$f(y_t^{(i)}) = \exp\left\{ \frac{y_t^{(i)} \eta_t^{(i)} - b(\eta_t^{(i)})}{a_t^{(i)}(\phi)} + c(y_t^{(i)}, \phi) \right\},$$

where $a_t^{(i)}(\phi)$, $b(\eta_t^{(i)})$, and $c(y_t^{(i)}, \phi)$ are known functions and specified for each member of the exponential family, and $\eta_t^{(i)}$ is a parameter in the mean as defined in Eq.(2). Typically, $a_t^{(i)}(\phi) = \phi$. Then, the deviance of the exponential family can be computed as

$$Dev = 2 \frac{\sum_{i=1}^{m} \left(y_t^{(i)} (\tilde{\eta}_t^{(i)} - \hat{\eta}_t^{(i)}) - b(\tilde{\eta}_t^{(i)}) + b(\hat{\eta}_t^{(i)}) \right)}{\phi},$$

where $\tilde{\eta}_t^{(i)}$ denotes the true value under a saturated model, $\hat{\eta}_t^{(i)}$ denotes the fitted values of the model. Thus, $\tilde{\eta}_t^{(i)}$ and $b(\tilde{\eta}_t^{(i)})$ are constant in model fitting. The derivative of b always satisfies $b'(\eta_t^{(i)}) = \mu_t^{(i)}$. Moreover, it has been proved that $b(\hat{\eta}_t^{(i)})$ is a convex function on the natural parameter space $\mathbf{H} = \{\hat{\boldsymbol{\eta}} | b(\hat{\boldsymbol{\eta}}) < \infty\}$ [19]. Thus, the deviance contains either linear terms or a convex term with respect to $\hat{\eta}$. In our model (5), $\hat{\eta}$ is linear with respect to $\mathbf{W}$. Hence, the deviance term in Eq.(6) is convex with respect to $\mathbf{U}$ and $\mathbf{V}$.

Moreover, it is true that $b'(\hat{\eta}_t^{(i)}) = \hat{\mu}_t^{(i)} = g^{-1}(\hat{\eta}_t^{(i)})$ which is the inverse of a continuous link function [19]. The first term of Eq.(6) is continuously differentiable with respect to $\mathbf{U}$ and $\mathbf{V}$. Thus, theorem 1 holds. □

2.3 Optimization Algorithm

To solve Eq.(6), we design an alternating optimization algorithm that alternates between optimizing two working sets of variables: one set consisting of $\mathbf{U}$ and $\mathbf{V}$ and the other consisting of $\boldsymbol{\alpha}$.

(a) Find $\mathbf{U}$ ***and*** $\mathbf{V}$ ***when*** $\boldsymbol{\alpha}$ ***is fixed***

When $\boldsymbol{\alpha}$ is fixed, the objective function of Eq.(6), denoted by $f(\mathbf{U}, \mathbf{V})$, is convex with a continuously differentiable part $\ell(\mathbf{U}, \mathbf{V})$ that is the deviance and a nonsmooth part $R(\mathbf{U}, \mathbf{V})$ that constitutes the two regularizers. We hence have

$$f(\mathbf{U}, \mathbf{V}) = \ell(\mathbf{U}, \mathbf{V}) + R(\mathbf{U}, \mathbf{V}).$$

We develop a FISTA algorithm in the following iterative procedure to find optimal $\mathbf{U}$ and $\mathbf{V}$.

Denote the iterates at the k-th iteration by $\mathbf{U}_k$ and $\mathbf{V}_k$. Let $\nabla_{\mathbf{U}} \ell(\mathbf{U}, \mathbf{V})$, $\nabla_{\mathbf{V}} \ell(\mathbf{U}, \mathbf{V})$ be the partial derivative of $\ell(\mathbf{U}, \mathbf{V})$ with respect to $\mathbf{U}$ and $\mathbf{V}$, respectively, For any given point $(\tilde{\mathbf{U}}, \tilde{\mathbf{V}})$, the following $Q_{L,\tilde{\mathbf{U}},\tilde{\mathbf{V}}}(\mathbf{U}, \mathbf{V})$ is a *well-defined* proximal map for the non-smooth R

$$\begin{aligned} Q_{L,\tilde{\mathbf{U}},\tilde{\mathbf{V}}}(\mathbf{U}, \mathbf{V}) = {} & \ell(\tilde{\mathbf{U}}, \tilde{\mathbf{V}}) + R(\mathbf{U}, \mathbf{V}) \\ & + \langle \nabla_{\mathbf{U}} \ell(\tilde{\mathbf{U}}, \tilde{\mathbf{V}}), \mathbf{U} - \tilde{\mathbf{U}} \rangle + \frac{L}{2} \|\mathbf{U} - \tilde{\mathbf{U}}\|_F^2 \\ & + \langle \nabla_{\mathbf{V}} \ell(\tilde{\mathbf{U}}, \tilde{\mathbf{V}}), \mathbf{V} - \tilde{\mathbf{V}} \rangle + \frac{L}{2} \|\mathbf{V} - \tilde{\mathbf{V}}\|_F^2. \end{aligned}$$

If $\ell(\mathbf{U}, \mathbf{V})$ has Lipschitz continuous gradient with Lipschitz modulis L. Then, according to the Lemma 2.1 in [3], the inequality

$$f(\mathbf{U}, \mathbf{V}) \leq Q_{L,\tilde{\mathbf{U}},\tilde{\mathbf{V}}}(\mathbf{U}, \mathbf{V}).$$

holds indicating that $Q_{L,\tilde{\mathbf{U}},\tilde{\mathbf{V}}}(\mathbf{U},\mathbf{V})$ is the upper bound of $f(\mathbf{U},\mathbf{V})$.

Starting from an initial point $(\mathbf{U}_0,\mathbf{V}_0)$, we iteratively search for the optimal solution. At each iteration k, we first use the iterates $(\mathbf{U}_{k-1},\mathbf{V}_{k-1})$ and $(\mathbf{U}_{k-2},\mathbf{V}_{k-2})$ to compute (at the first iteration, $(\tilde{\mathbf{U}}_1,\tilde{\mathbf{V}}_1)=(\mathbf{U}_0,\mathbf{V}_0)$)

$$\begin{aligned}\tilde{\mathbf{U}}_k &= \mathbf{U}_{k-1}+\left(\frac{t_{k-1}-1}{t_k}\right)(\mathbf{U}_{k-1}-\mathbf{U}_{k-2}),\\ \tilde{\mathbf{V}}_k &= \mathbf{V}_{k-1}+\left(\frac{t_{k-1}-1}{t_k}\right)(\mathbf{V}_{k-1}-\mathbf{V}_{k-2}),\end{aligned} \tag{7}$$

where t_k is a scalar and updated at each iteration as:

$$t_{k+1}=\frac{1+\sqrt{1+4t_k^2}}{2}. \tag{8}$$

Then, we solve the following problem

$$\begin{aligned}\min_{\mathbf{U},\mathbf{V}} \quad & \langle\nabla_{\mathbf{U}}\ell_k,\mathbf{U}-\tilde{\mathbf{U}}_k\rangle+\frac{L}{2}\|\mathbf{U}-\tilde{\mathbf{U}}_k\|_F^2\\ & +\langle\nabla_{\mathbf{V}}\ell_k,\mathbf{V}-\tilde{\mathbf{V}}_k\rangle+\frac{L}{2}\|\mathbf{V}-\tilde{\mathbf{V}}_k\|_F^2\\ & +R(\mathbf{U},\mathbf{V})\end{aligned} \tag{9}$$

for a solution $(\mathbf{U}_k,\mathbf{V}_k)$, where $\nabla_{\mathbf{U}}\ell_k$ and $\nabla_{\mathbf{V}}\ell_k$ are respectively the partial derivatives of ℓ computed at $(\tilde{\mathbf{U}}_k,\tilde{\mathbf{V}}_k)$, and L acts as a learning step size.

Since there is no interacting term between $\mathbf{U}$ and $\mathbf{V}$ in Eq.(9), the problem can be decomposed into two separate subproblems as follows:

$$\min_{\mathbf{U}}\langle\nabla_{\mathbf{U}}\ell_k,\mathbf{U}-\tilde{\mathbf{U}}_k\rangle+\frac{L}{2}\|\mathbf{U}-\tilde{\mathbf{U}}_k\|_F^2+\lambda_1\|\mathbf{U}\|_{1,2}, \tag{10}$$

$$\min_{\mathbf{V}}\langle\nabla_{\mathbf{V}}\ell_k,\mathbf{V}-\tilde{\mathbf{V}}_k\rangle+\frac{L}{2}\|\mathbf{V}-\tilde{\mathbf{V}}_k\|_F^2+\lambda_2\|\mathbf{V}^\top\|_{1,2}. \tag{11}$$

The two subproblems share the same structure and thus can be solved following the same procedure. Hence, we only show how to solve (10) for the best $\mathbf{U}$.

Eq.(10) is equivalent to the following problem

$$\min_{\mathbf{U}}\frac{1}{2}\left\|\mathbf{U}-\left(\tilde{\mathbf{U}}_k-\frac{1}{L}\nabla_{\mathbf{U}}\ell_k\right)\right\|_F^2+\frac{\lambda_1}{L}\|\mathbf{U}\|_{1,2}$$

after omitting constants, and this problem has a closed-form solution where each row of $\mathbf{U}_k$, $\mathbf{U}^k_{(i,)}$ is:

$$\mathbf{U}^k_{(i,)}=\max\left(0,1-\frac{\lambda_1}{L\|\mathbf{P}^{(k)}_{(i,)}\|_2}\right)\mathbf{P}^{(k)}_{(i,)},$$

and $\mathbf{P}^{(k)}=\tilde{\mathbf{U}}_k-\frac{1}{L}\nabla_{\mathbf{U}}\ell_k$. The gradient vector $\nabla_{\mathbf{U}}\ell_k$ (i.e., the gradient of the deviance) can be computed by Eq.(3) with the fixed $\boldsymbol{\alpha}$, i.e.

$$\nabla_{\mathbf{U}}\ell_k=\text{reshape}\left(\sum_{i=1}^{m}\left(\mathbf{D}^{(i)}\right)^\top\left(\mathbf{\Sigma}^{(i)}\right)^{-1}\mathbf{s}_k^{(i)}\right) \tag{12}$$

where $\mathbf{s}_k^{(i)}=\mathbf{y}^{(i)}-\boldsymbol{\mu}^{(i)}$, and $\mu_t^{(i)}=g^{-1}(tr(\mathbf{X}_{(i;t)}^\top(\tilde{\mathbf{U}}_k+\tilde{\mathbf{V}}_k)))$.

In the above discussion, the Lipschitz modulus L is computed and given. However, the calculation of L can be computational expensive. We therefore follow the similar argument in [9] to find a proper approximation L_k at each iteration k starting from $L_0>0$. Recall that the Lipschits constant L is defined:

$$L=\max_{\mathbf{W}}\lambda_{\max}(\nabla\nabla\ell_{\mathbf{W}})$$

where $\lambda_{\max}(\cdot)$ indicates the maximum singular value of the Hessian of ℓ. Decompose the Hessian matrix $\nabla\nabla\ell_{\mathbf{W}}|_{\mathbf{W}\to 0}$ into $\mathbf{M}^\top\mathbf{M}$ where $\mathbf{M}\in\mathbb{R}^{d(\tau+1)\times q}$ and q is the rank of the Hessian matrix. We have an upper bound of L as follows:

$$L\le||\mathbf{M}||_{\infty,1}||\mathbf{M}^\top||_{\infty,1}. \tag{13}$$

We use the upper bound $\tilde{L}$ in Eq.(13) as L in our iterations. Using this upper bound may increase the number of iterative steps for convergence. Algorithm 1 summarizes the steps for finding optimal $\mathbf{U}$ and $\mathbf{V}$ with fixed $\boldsymbol{\alpha}$.

Algorithm 1 Search for optimal $\mathbf{U}$ and $\mathbf{V}$ with fixed $\boldsymbol{\alpha}$

```
Input: $\mathbf{X}$, $\mathbf{y}$, $\mathbf{\Sigma}$, $\lambda_1$, $\lambda_2$
Output: $\mathbf{U}$, $\mathbf{V}$
1. $k=1$, compute $\tilde{L}$ and initialize $t_1=1$, $\mathbf{U}_0=\tilde{\mathbf{U}}_1=\mathbf{0}$ and $\mathbf{V}_0=\tilde{\mathbf{V}}_1=\mathbf{0}$;
2. Solve Eq.(9) to obtain $\mathbf{U}_k$ and $\mathbf{V}_k$.
3. Compute $t_{k+1}$ by Eq.(8).
4. Compute $\tilde{\mathbf{U}}_{k+1}$ and $\tilde{\mathbf{V}}_{k+1}$ by Eq.(7).
5. $k=k+1$.
Repeat 2 $\sim$ 5 until convergence.
```

(b) Find $\boldsymbol{\alpha}$ when $\mathbf{U}$ and $\mathbf{V}$ **are fixed**

When $\mathbf{U}$ and $\mathbf{V}$ are fixed, the regularizers no longer appear in the objective of Eq.(6). Eq.(6) is degenerated into just the GEE formula with $\boldsymbol{\alpha}$ as the variables. Hence, $\boldsymbol{\alpha}$ can be estimated via the standard GEE procedure, i.e., from the current Pearson residuals defined by:

$$\gamma_t^{(i)}=\frac{y_t^{(i)}-tr\left(\left(\mathbf{X}_{(i;t)}\right)^\top(\mathbf{U}+\mathbf{V})\right)}{(\sigma_{t,t}^{(i)})^{(1/2)}}.$$

where $\sigma_{t,t}^{(i)}$ is the t-th diagonal entry in the matrix $\mathbf{\Sigma}^{(i)}$ [12]. The specific estimator of $\boldsymbol{\alpha}$ depends on the choices of $\mathbf{R}(\boldsymbol{\alpha})$. This GEE-based procedure has been shown to find a *consistent* estimate of $\boldsymbol{\alpha}$ [12].

Let $N=mn$ be the total number of training examples, and $p=d(\tau+1)$ be the practical number of parameters in $\mathbf{W}$. A general approach to estimating $\mathbf{R}$ is given by:

$$r_{j,k}=\sum_{i=1}^{m}\frac{\gamma_j^{(i)}\gamma_k^{(i)}}{N-p}, \tag{14}$$

for $j=1,\cdots,n$, and $k=1,\cdots,n$. In addition, the scaler parameter ϕ in Eq.(4) can be estimated as follows:

$$\phi=(N-p)/\sum_{i=1}^{m}\sum_{t=1}^{n}\left(\gamma_t^{(i)}\right)^2. \tag{15}$$

Algorithm 2 depicts the overall procedure for solving Eq.(6).

Algorithm 2 Main algorithm - Jointly select features and temporal points

```
Input: $\mathbf{X}$, $\mathbf{y}$, $\lambda_1$, $\lambda_2$
Output: $\mathbf{U}$, $\mathbf{V}$
1. Set $\mathbf{R}(\alpha)=\mathbf{I}$;
2. Solve for $\mathbf{U}$ and $\mathbf{V}$ using Algorithm 1.
3. Estimate $\alpha$ using a proper estimator in [12] and compute $\mathbf{R}(\alpha)$ by Eq.(14) and $\phi$ by Eq.(15).
Repeat 2 $\sim$ 3 until convergence.
```

3. THEORETICAL ANALYSIS

We provide a convergence analysis for Algorithm 1 and an asymptotic analysis for the proposed formulation.

3.1 Convergence Analysis

We show that Algorithm 1 converges to the optimal solution with a convergence rate of $O(1/k^2)$. The proof follows largely the arguments in [3]. We only provide a sketch here.

THEOREM 2. *Let $\mathbf{U}_k$ and $\mathbf{V}_k$ be the pair of the matrix generated by Algorithm 1. Then for any $k \geq 1$*

$$f(\mathbf{U}_k, \mathbf{V}_k) - f(\hat{\mathbf{U}}, \hat{\mathbf{V}}) \leq \frac{2\tilde{L}\left(||\mathbf{U}_0 - \hat{\mathbf{U}}||_F^2 + ||\mathbf{V}_0 - \hat{\mathbf{V}}||_F^2\right)}{(k+1)^2}$$

where $(\hat{\mathbf{U}}, \hat{\mathbf{V}})$ is a globally optimal solution of Eq.(6).

PROOF. We start with defining the following quantities

$$\begin{aligned}
v_k =& f(\mathbf{U}_k, \mathbf{V}_k) - f(\hat{\mathbf{U}}, \hat{\mathbf{V}}),\\
a_k =& \frac{2}{L_k} t_k^2 v_k,\\
b_k =& ||t_k\mathbf{U}_k - (t_k - 1)\mathbf{U}_{k-1} - \hat{\mathbf{U}}||_F^2\\
&+ ||t_k\mathbf{V}_k - (t_k - 1)\mathbf{V}_{k-1} - \hat{\mathbf{V}}||_F^2,\\
c =& ||\tilde{\mathbf{U}}_1 - \hat{\mathbf{U}}||_F^2 + ||\tilde{\mathbf{V}}_1 - \hat{\mathbf{V}}||_F^2\\
=& ||\mathbf{U}_0 - \hat{\mathbf{U}}||_F^2 + ||\mathbf{V}_0 - \hat{\mathbf{V}}||_F^2,
\end{aligned}$$

where $\tilde{\mathbf{U}}_1 = \mathbf{U}_0$, $\tilde{\mathbf{V}}_1 = \mathbf{V}_0$, and subsequent $\tilde{\mathbf{U}}_k$ and $\tilde{\mathbf{V}}_k$ are defined by Eq.(7). Following the proof of Theorem 4.4 in [3], in the first iteration, given $t_1 = 1$, we have $a_1 = \frac{2}{L_1}v_1$, and $b_1 = ||\mathbf{U}_1 - \hat{\mathbf{U}}||_F^2 - ||\mathbf{V}_1 - \hat{\mathbf{V}}||_F^2$. We show that $a_1 + b_1 \leq c$ by applying Lemma 2.3 in [3], which yields

$$\begin{aligned}
&f(\hat{\mathbf{U}}, \hat{\mathbf{V}}) - f(\mathbf{U}_1, \mathbf{V}_1) = -v_1\\
\geq& \frac{L_1}{2}||\mathbf{U}_1 - \tilde{\mathbf{U}}_1||_F^2 + L_1\langle \tilde{\mathbf{U}}_1 - \hat{\mathbf{U}}, \mathbf{U}_1 - \tilde{\mathbf{U}}_1\rangle\\
&+ \frac{L_1}{2}||\mathbf{V}_1 - \tilde{\mathbf{V}}_1||_F^2 + L_1\langle \tilde{\mathbf{V}}_1 - \hat{\mathbf{V}}, \mathbf{V}_1 - \tilde{\mathbf{V}}_1\rangle\\
=& \frac{L_1}{2}(||\mathbf{U}_1 - \hat{\mathbf{U}}||_F^2 - ||\tilde{\mathbf{U}}_1 - \hat{\mathbf{U}}||_F^2)\\
&+ \frac{L_1}{2}(||\mathbf{V}_1 - \hat{\mathbf{V}}||_F^2 - ||\tilde{\mathbf{V}}_1 - \hat{\mathbf{V}}||_F^2).
\end{aligned}$$

Reorganizing the above inequality yields

$$\begin{aligned}
\frac{2}{L_1}t_1^2 v_1 + ||\mathbf{U}_1 - \hat{\mathbf{U}}||_F^2 + ||\mathbf{V}_1 - \hat{\mathbf{V}}||_F^2 \leq&\\
||\tilde{\mathbf{U}}_1 - \hat{\mathbf{U}}||_F^2 + ||\tilde{\mathbf{V}}_1 - \hat{\mathbf{V}}||_F^2&
\end{aligned}$$

Thus, $a_1 + b_1 \leq c$ holds.

Then, according to Lemma 4.1 in [3], we have for every $k \geq 1$, $a_k - a_{k+1} \geq b_{k+1} - b_k$, together with $a_1 + b_1 \leq c$, which derives into the following inequality,

$$c \geq a_1 + b_1 \geq a_2 + b_2 \geq \cdots \geq a_k + b_k \geq a_k.$$

Therefore, we obtain that

$$\frac{2}{L_k}t_k^2 v_k \leq ||\mathbf{U}_0 - \hat{\mathbf{U}}||_F^2 + ||\mathbf{V}_0 - \hat{\mathbf{V}}||_F^2, \tag{16}$$

Given t_k is updated according to Eq.(8), it is easy to show that $t_k \geq \dfrac{(k+1)}{2}$. Substituting this inequality into Eq.(16) yields

$$v_k \leq \frac{2L_k\left(||\mathbf{U}_0 - \hat{\mathbf{U}}||_F^2 + ||\mathbf{V}_0 - \hat{\mathbf{V}}||_F^2\right)}{(k+1)^2}$$

By the Remark 3.2 in [3] and the inequality (13), we also know that an upper bound of L_k is $\tilde{L}$. Hence,

$$f(\mathbf{U}_k, \mathbf{V}_k) - f(\hat{\mathbf{U}}, \hat{\mathbf{V}}) \leq \frac{2\tilde{L}\left(||\mathbf{U}_0 - \hat{\mathbf{U}}||_F^2 + ||\mathbf{V}_0 - \hat{\mathbf{V}}||_F^2\right)}{(k+1)^2}$$

In our algorithm, we set $L_k = \tilde{L}, \forall k$. □

REMARK 1. *The loss function, $\ell(\mathbf{U}, \mathbf{V})$, of an exponential distribution has Lipschitz continuous gradient within the range $\{||\mathbf{U}||_{1,2} \leq \delta_1, ||\mathbf{V}^\top||_{1,2} \leq \delta_2\}$ where δ_1, δ_2 are constant values in terms of λ_1, λ_2, respectively to guarantee the non-trivial step size $\frac{\lambda}{L}$. Otherwise, it may lead to a suboptimal solution.*

3.2 Asymptotic Analysis

To facilitate the asymptotic analysis, we re-write the notation as follows: let

$$\boldsymbol{\beta} = [\text{vect}(\mathbf{U})^\top, \text{vect}(\mathbf{V})^\top]^\top, \quad \mathbf{H}^{(i)} = [\mathbf{h}_{\tau+1}^{(i)}, \cdots, \mathbf{h}_n^{(i)}]$$

and

$$\mathbf{h}_t^{(i)} = [\text{vect}(\mathbf{X}_{i;t})^\top, \text{vect}(\mathbf{X}_{i;t})^\top]^\top$$

where one block $\mathbf{X}_{i;t}$ corresponds to $\mathbf{U}$ and the other to $\mathbf{V}$. Then, correspondingly, we have $\eta_t^{(i)} = (\mathbf{h}_t^{(i)})^\top\boldsymbol{\beta}$, and $f(\mathbf{U}, \mathbf{V})$ can be re-written as $f(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) + R(\boldsymbol{\beta}; \lambda_1, \lambda_2)$.

Solve Eq.(6) yields a solution to the penalized estimating equations:

$$\sum_i (\mathbf{D}^{(i)})^\top(\mathbf{\Sigma}^{(i)})^{-1}\mathbf{s}^{(i)} + \lambda\frac{\partial R(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = 0 \tag{17}$$

assuming $\lambda_1 = \lambda_2 = \lambda$ for notational convenience which will not change the property. Given our model definition (5), $\mathbf{D}^{(i)} = \mathbf{A}^{(i)}(\mathbf{H}^{(i)})^\top$. The first term in (17) is the estimating functions in GEE [12] whereas the second term corresponds to the regularizers. The asymptotic property of Eq.(6) can be naturally derived from the results in [12] which have proved that the estimating equations $L(\boldsymbol{\beta}) = \sum_i(\mathbf{D}^{(i)})^\top(\mathbf{\Sigma}^{(i)})^{-1}\mathbf{s}^{(i)}$ of GEE gives a consistent estimator of $\boldsymbol{\beta}$. We extend the same argument to our formulation Eq.(6) in Theorem 3 under the following regularity conditions: $\mathbf{H}^{(i)}$ is bounded, and $\lim_{m\to\infty}(\sum_i \mathbf{H}^{(i)})/m = \mathbf{H}^{(0)}$, and $(\mathbf{H}^{(i)})^\top\mathbf{H}^{(i)}$ are not singular, and the following limit is also not singular

$$\lim_{m\to\infty}(\sum_i(\mathbf{H}^{(i)})^\top\mathbf{H}^{(i)})/m;$$

Moreover, $L(\boldsymbol{\beta})$ is twice continuously differentiable with respect to $\boldsymbol{\beta}$, and $\partial L/\partial\boldsymbol{\beta}$ is positive definite.

THEOREM 3. *Assume that: (1) $\hat{\boldsymbol{\alpha}}$ is a consistent estimator given $\boldsymbol{\beta}$; (2) $\hat{\phi}$ is a consistent estimator given $\boldsymbol{\beta}$; and (3) the tuning parameter $\lambda_m = o(\sqrt{m})$. Under the regularity conditions listed above, optimizing Eq.(6) yields an asymptotically consistent and normally distributed estimator $\hat{\boldsymbol{\beta}}$, that is:*

$$\sqrt{m}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \to_d N(0, \mathbf{\Sigma}) \quad as \quad m \to \infty$$

where $\boldsymbol{\beta}^*$ is the true model coefficients in a model of $E(y_t^{(i)}) = g^{-1}((\mathbf{h}_t^{(i)})^\top \boldsymbol{\beta})$ and Σ is a positive definite variance-covariance matrix (see [12] for details of Σ).

Proof. Multiplying $1/m$ to both sides of Eq.(17) yields

$$\frac{1}{m}\sum_i (\mathbf{D}^{(i)})^\top (\boldsymbol{\Sigma}^{(i)})^{-1}\mathbf{s}^{(i)} + \frac{\lambda_m}{m}\frac{\partial R(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0. \tag{18}$$

It is known that solving $\frac{1}{m}\sum_i (\mathbf{D}^{(i)})^\top (\boldsymbol{\Sigma}^{(i)})^{-1}\mathbf{s}^{(i)} = 0$ yields an estimate of $\hat{\boldsymbol{\beta}}$ that is asymptotically consistent with $\boldsymbol{\beta}^*$:

$$\sqrt{m}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \to_d N(0, \boldsymbol{\Sigma}) \quad \text{as} \quad m \to \infty \quad [12].$$

Since our regularizer R (based on the $\ell_{1,2}$ matrix norm) is Lipschitz continuous, its partial derivative $\partial R(\boldsymbol{\beta})/\partial \boldsymbol{\beta}$ is bounded. The second term of Eq.(18) vanishes when $m \to \infty$, and thus the conclusion holds. □

Recall how $\hat{\boldsymbol{\alpha}}$ and $\hat{\phi}$ are estimated in the proposed method. Those estimates from the Pearson residuals are consistent. Thus, the estimate $\hat{\boldsymbol{\beta}}$ in the proposed method is asymptotically consistent and normally distributed according to Theorem 3.

3.3 Exemplar Exponential Families with Lipschitz Condition

The purposed algorithm is suitable to optimize any loss function that has Lipschitz continuous gradient. In this section, we discuss that three exemplar exponential families: Gaussian, Bernoulli, and Poisson, satisfy the Lipschitz condition. We specify how to compute the gradient of the loss function for these distributions. The gradients will instantiate (and replace) Eq.(12) used in our algorithm.

3.3.1 Gaussian Distribution

If the outcome follows a Gaussian distribution, then the outcome y is linearly regressive in terms of the covariates in the observations. The mean and the conditional covariance of y with a working correlation structure $\mathbf{R}(\boldsymbol{\alpha})$ are calculated as:

$$E(y_t^{(i)}) = \mu_t^{(i)} = tr\left(\mathbf{X}_{(i;t)}^\top \mathbf{W}\right),$$
$$cov(\mathbf{y}^{(i)}) = \boldsymbol{\Sigma}^{(i)} = \mathbf{R}(\boldsymbol{\alpha}),$$

so the gradient $\nabla_{\mathbf{U}}\ell_k$ in Eq.(12) at the k-th iteration can be computed as

$$\nabla_{\mathbf{U}}\ell_k = \text{reshape}\left(\sum_{i=1}^m \left(\mathbf{D}^{(i)}\right)^\top (\mathbf{R}(\boldsymbol{\alpha}))^{-1}\,\mathbf{s}_k^{(i)}\right),$$

where $\mathbf{D}^{(i)} = \frac{\partial \boldsymbol{\mu}^{(i)}}{\partial \text{vect}(\tilde{\mathbf{U}}_k)} = \left[\text{vect}\left(\mathbf{X}_{(i;1)}\right), \ldots, \text{vect}\left(\mathbf{X}_{(i;n)}\right)\right]^\top$, and $\mathbf{s}_k^{(i)} = \mathbf{y}^{(i)} - \left(\mathbf{D}^{(i)}\right)^\top \text{vect}(\tilde{\mathbf{U}}_k)$. The gradient $\nabla_{\mathbf{V}}\ell_k$ can be similarly computed. Hence, the gradient is linear in terms of $\boldsymbol{\beta}$, and thus Lipschitz continuous.

3.3.2 Bernoulli Distribution

If the generalized variables μ follow a Bernoulli distribution and the outcomes are binary variables. The relationship between the outcome and covariates can be learned by a logistic regression which is a special case of the GLM with the Bernoulli assumption. Hence, the mean and the conditional covariance of y with the working correlation structure $\mathbf{R}(\boldsymbol{\alpha})$ are formulated as

$$E(y_t^{(i)}) = \mu_t^{(i)} = \frac{\exp(\eta_t^{(i)})}{1 + \exp(\eta_t^{(i)})} \tag{19}$$
$$cov(\mathbf{y}^{(i)}) = \boldsymbol{\Sigma}^{(i)} = \frac{\left(\mathbf{A}^{(i)}\right)^{1/2}\mathbf{R}(\boldsymbol{\alpha})\left(\mathbf{A}^{(i)}\right)^{1/2}}{\phi}$$

where $\mathbf{A}^{(i)} = \text{diag}\left(\langle \boldsymbol{\mu}^{(i)}, 1 - \boldsymbol{\mu}^{(i)}\rangle\right)$ $= \text{diag}\left(\frac{\exp(\eta_t^{(i)})}{\left(1+\exp(\eta_t^{(i)})\right)^2}\right)$ and $\eta_t^{(i)} = tr(\mathbf{X}_{(i;t)}^\top \mathbf{W})$.

The gradient $\nabla_{\mathbf{U}}\ell_k$ in Eq.(12) can be written as:

$$\text{reshape}\left(\left(\mathbf{D}^{(i)}\right)^\top (\mathbf{A}^{(i)})^{-1/2}\mathbf{R}(\boldsymbol{\alpha})^{-1}(\mathbf{A}^{(i)})^{-1/2}\mathbf{s}_k^{(i)}\right)$$

where $\mathbf{D}^{(i)} = \frac{\partial \boldsymbol{\mu}^{(i)}}{\partial \boldsymbol{\eta}^{(i)}} \times \frac{\partial \boldsymbol{\eta}^{(i)}}{\partial \text{vect}(\tilde{\mathbf{U}}_k)}$ $= \mathbf{A}^{(i)}\left[\text{vect}\left(\mathbf{X}_{(i;1)}\right), \ldots, \text{vect}\left(\mathbf{X}_{(i;n)}\right)\right]^\top$, and $\mathbf{s}_k^{(i)} = \mathbf{y}^{(i)} - \boldsymbol{\mu}^{(i)}(\tilde{\mathbf{U}}_k)$. The gradient $\nabla_{\mathbf{V}}\ell_k$ can be similarly computed.

3.3.3 Poisson Distribution

If the generalized variables μ follow a Poisson distribution and the outcomes contain count values. The relationship of the outcome and covariates is learned by a Poisson regression. The mean and the conditional covariance of y with the working correlation structure $\mathbf{R}(\boldsymbol{\alpha})$ are formulated as

$$E(y_t^{(i)}) = \mu_t^{(i)} = \exp(\eta_t^{(i)})$$
$$cov(\mathbf{y}^{(i)}) = \boldsymbol{\Sigma}^{(i)} = \frac{\left(\mathbf{A}^{(i)}\right)^{1/2}\mathbf{R}(\boldsymbol{\alpha})\left(\mathbf{A}^{(i)}\right)^{1/2}}{\phi}$$

where $\mathbf{A}^{(i)} = \text{diag}\left((\boldsymbol{\mu}^{(i)})'\right) = \text{diag}\left(\exp(\eta_t^{(i)})\right)$. The gradient $\nabla_{\mathbf{U}}\ell_k$ can be computed using the general formula Eq.(12). The loss function of Poisson regression does not have globally Lipschitz continuous gradient. But the regularized loss function is equivalent to requiring the constraints, $||\mathbf{U}||_{1,2} \leq \delta_1$ and $||\mathbf{V}^\top||_{1,2} \leq \delta_2$ [17] for appropriate values of δ_1 and δ_2 that are determined according to λ_1 and λ_2. The loss function of Poisson regression does have Lipschitz continuous gradient within the confined region.

4. EMPIRICAL EVALUATION

We validated the proposed approach by comparing it to several most relevant and recent methods. Three GLM-based [16] methods: GEE [12], GLMM [11, 15], and RE-EM tree[1] [18] were compared. The recent graphical Granger modeling[2] [13] and a support vector machine based method called CSVM were also used. RE-EM tree and graphical Granger modeling could only be applied to regression problems (linearly regressive data from Gaussian distributions), and CSVM was only suitable to classification tasks (logistically regressive data from Bernoulli distributions). We named our approach by LGL (longitudinal group lasso). The normalized mean squared error (nMSE), which is the MSE divided by the variance of y [22, 9], was used to measure regression performance. The area under the ROC curve (AUC) [5] was used to measure classification performance.

[1] An R package is available in the Comprehensive R Archive Network (CRAN)

[2] downloaded from the author's website http://www-bcf.usc.edu/~liu32/code.html

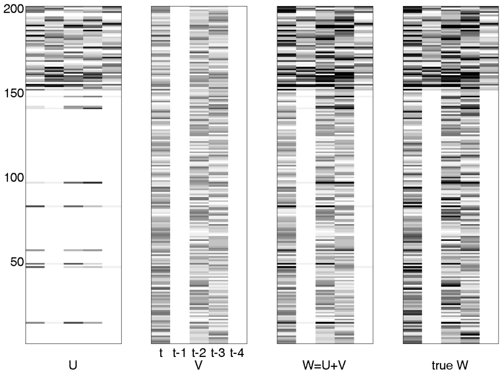

Figure 2: The model constructed by our approach LGL on a synthetic dataset.

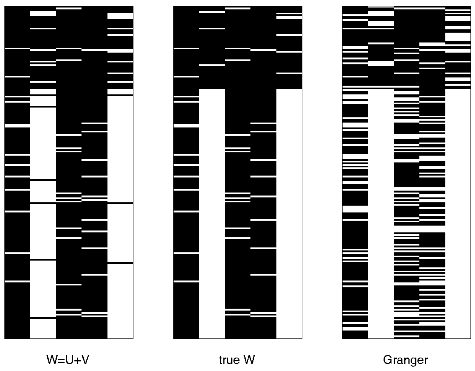

Figure 3: Comparison between the constructed models by LGL and Granger.

4.1 Synthetic Data

We generated a data matrix $\mathbf{X} \in \mathbb{R}^{d \times Tm}$ from the normal distribution $N(0, 16)$, where $d = 200$, $T = 30$, and $m = 400$. All training examples $\mathbf{X}_{(i;t)} (i = 1, \cdots, m, \forall t = \tau+1, \cdots, T)$ and $\tau = 4$ were formed from the matrix $\mathbf{X}$. Then, $\mathbf{U}$ and $\mathbf{V}$ were generated from the normal distribution $N(0, 49)$. We set the rows corresponding to features from 1 to 150 in $\mathbf{U}$ to zero and the columns 2 and 5 of $\mathbf{V}$ to zero, and computed $\mathbf{W} = \mathbf{U} + \mathbf{V}$. The residuals $\mathbf{s}^{(i)}$ of every subject were generated from a multivariate normal distribution of different variances, $N(0, 1^2), N(0, 2^2), N(0, 3^2)$. The covariance matrix of the residual followed different working correlation structures $\mathbf{R}(\alpha)$ with the parameter $\alpha = 0.64$. We generated 9 sets of regression residuals by choosing different combinations of the variances and the working correlation structures. Finally, the outcome variables $\mathbf{y}^{(i)}$ were computed as

$$\mathbf{y}^{(i)} = \left[\text{vect}\left(\mathbf{X}_{(i;\tau+1)}\right), \ldots, \text{vect}\left(\mathbf{X}_{(i;n)}\right)\right]^\top \text{vect}(\mathbf{U}+\mathbf{V}) + \mathbf{s}^{(i)}.$$

The above procedure produced regression data. Using the same data $\mathbf{X}$, the outcome $y_t^{(i)}$ of a classification problem was generated from the Bernoulli Distribution with $\text{B}(1, \mu_t^{(i)})$ where we used Eq.(19) with the regression $\mathbf{y}^{(i)}$ to obtain $\boldsymbol{\mu}^{(i)}$. We hence obtained totally 18 synthesized data with 9 datasets for each distribution. We used the 25 early records of each subject to compose the training data and the rest 5 records to form test data.

Table 1 shows the results where we can see that LGL outperformed all other methods on all the simulated datasets. The proposed method with correct correlation assumptions always performed the best. The graphical Granger modeling performed reasonably well but lacked of consideration of temporal correlation in the consecutive records. When the simulated noise increased, the performance of all methods had dropped as expected. We further demonstrate the selected features and temporal contingency. Figure 2 shows the constructed $\mathbf{U}, \mathbf{V}$, and $\mathbf{W}$ by the LGL on the regression data with the AR(1) covariance structure and $N(0, 3^2)$ residual where darker colors indicate larger values (and white means 0). Most of the features from 150 to 200 were selected in $\mathbf{U}$ and the correct columns (i.e., $1, 3, 4$) were selected in $\mathbf{V}$. We compared our approach with the Granger model that also learned $\mathbf{W}$ in Figure 3. Obviously, the Granger model excluded too many variables in the model. These results demonstrate the capability of LGL in terms of simultaneously capturing the important features and lagged effects.

4.2 Real-world Data

We tested our approach on two real-world datasets: the college alcohol use dataset; and the national longitudinal survey of youth (NLSY) dataset[3]. All comparison methods were used except GLMM due to its prohibitive computational costs. The college alcohol use dataset consisted of data from 504 college students on 52 variables in a period of continuous 30 days. The 52 variables measured each subject on daily stress, moods, emotion and substance use behavior. One of the variables measured the number of night-time drinks, which was our outcome variable, forming a regression problem. We also predicted the binge drinking behavior which is defined as having 5 or more night-time drinks, which formed a classification problem. The NLSY dataset consisted of 11 yearly data for 3,376 subjects on 27 variables. The outcome variable measured the number of days that a subject had binge drinking in past 30 days, forming a regression problem. The other 26 variables measured features, such as smoking, drug use, family support and education.

For the college alcohol use data, we experimented with using the last $t = 3, 5, 8, 10$ days of records as test data, and the rest for training. We found $\tau = 3$ was feasible. Larger τ would not change the results because the extra time points would be excluded by our model. However, it practically would cut down the sample size of each subject. The parameters λ_1 and λ_2 in our approach and any tuning parameters in other methods were tuned in a three-fold cross validation within the training data. Table 2 shows the results where our approach LGL outperformed other methods in most settings. Among the four different correlation assumptions, LGL with AR(1) obtained the best performance on three of the four settings. The results also confirmed that modeling the correlation among repeated observations improved prediction performance [12]. We also observed that for instance, 16 out of 51 variables were selected when we used the last 5 days to test binge drinking prediction. Features related to exited mood, under stress and interacting with friends during night time were the risk factors for binge drinking. The past 3 days were all included in the model, showing there was "lagged" effects in alcohol use. The effect of past days was reduced with prolonged time lag.

For the NLSY dataset, we experimented respectively with using the last one, two and three years from each subject for test and the rest in training. We also considered $\tau = 3$, which means we used 3 year lagged data to predict the current year's behavior. All tuning parameters were tuned us-

[3] http://www.bls.gov/nls/nlsy97.htm

Table 1: Comparison of different algorithms on synthetic data: (top) regression; (bottom) classification.

			LGL				GEE						
	Structures	e	AR(1)	exchangeable	Tri-diag	ind	AR(1)	exchangeable	Tri-diag	ind	GLMM	RE-EM tree	Granger
Regression	AR(1)	$N(0,1^2)$	**0.0018**	0.0020	0.0019	0.0020	0.6613	0.6615	0.6614	0.6617	0.6657	0.9873	0.0664
		$N(0,2^2)$	**0.0025**	0.0026	0.0028	0.0039	0.7223	0.7236	0.7224	0.7242	0.7323	0.9998	0.0667
		$N(0,3^2)$	**0.0032**	0.0034	0.0036	0.0038	0.7191	0.7185	0.7182	0.7192	0.7179	0.9924	0.0676
	exchangeable	$N(0,1^2)$	0.0018	0.0016	**0.0015**	0.0022	0.6872	0.6875	0.6872	0.6873	0.6914	0.9977	0.0656
		$N(0,2^2)$	0.0024	**0.0023**	0.0024	0.0025	0.6927	0.6930	0.6927	0.6930	0.6931	0.9982	0.0691
		$N(0,3^2)$	0.0027	**0.0026**	0.0028	0.0032	0.7204	0.7204	0.7204	0.7205	0.7204	0.9797	0.0635
	Tri-diag	$N(0,1^2)$	0.0021	0.0021	**0.0021**	0.0022	0.7514	0.7514	0.7514	0.7514	0.7515	0.9925	0.0665
		$N(0,2^2)$	0.0018	0.0023	**0.0013**	0.0026	0.6790	0.6792	0.6791	0.6793	0.6840	0.9991	0.0680
		$N(0,3^2)$	0.0033	0.0035	**0.0031**	0.0041	0.7226	0.7235	0.7226	0.7226	0.7222	0.9998	0.0660

			LGL				GEE				
	Structures	e	AR(1)	exchangeable	Tri-diag	ind	AR(1)	exchangeable	Tri-diag	ind	CSVM
Classification	AR(1)	$N(0,1^2)$	**96.490%**	96.485%	96.485%	96.417%	77.691%	77.700%	77.699%	77.715%	76.644%
		$N(0,2^2)$	**96.442%**	96.431%	96.432%	96.653%	74.682%	74.727%	74.682%	74.731%	75.249%
		$N(0,3^2)$	**95.921%**	95.917%	95.917%	95.805%	77.704%	77.746%	77.708%	77.754%	77.547%
	exchangeable	$N(0,1^2)$	95.913%	**95.937%**	95.912%	95.883%	76.115%	75.812%	76.114%	75.923%	75.232%
		$N(0,2^2)$	95.139%	**95.161%**	95.147%	95.150%	70.290%	70.231%	70.275%	70.206%	71.687%
		$N(0,3^2)$	94.127%	94.091%	**94.135%**	93.470%	73.839%	73.782%	73.831%	73.776%	73.894%
	Tri-diag	$N(0,1^2)$	95.976%	95.941%	**95.978%**	95.889%	77.628%	77.634%	77.625%	77.617%	76.778%
		$N(0,2^2)$	95.231%	95.231%	**95.245%**	94.395%	72.132%	72.060%	72.126%	72.054%	71.615%
		$N(0,3^2)$	95.092%	95.087%	**95.094%**	94.231%	77.755%	77.533%	77.748%	77.637%	77.572%

Table 2: Comparison of different algorithms on the college alcohol use dataset: (top) predicting the number of night-time drinks (regression); (bottom) predicting the occurrence of binge drinking (classification).

		LGL				GEE					
	# observations	AR(1)	exchangeable	tri-diag	ind	AR	exchangeable	tri-diag	ind	RE-EM tree	Granger
Regression	3	**0.933513**	0.933863	0.935120	0.961841	1.064792	1.073358	1.063948	1.065760	1.115627	1.369948
	5	0.951999	0.954740	**0.951953**	0.976299	1.051219	1.067303	1.049305	1.072745	1.005753	1.420547
	8	**0.759935**	0.760450	0.760136	0.762205	0.787731	0.793329	0.787497	0.794089	0.759968	0.909706
	10	**0.769303**	0.769492	0.769428	0.774937	0.812622	0.818834	0.812011	0.806301	0.774797	0.940940

		LGL				GEE				
	# observations	AR(1)	exchangeable	tri-diag	ind	AR	exchangeable	tri-diag	ind	CSVM
Classification	3	79.737%	75.677%	79.772%	78.579%	78.401%	74.145%	78.650%	77.831%	**80.698%**
	5	**83.290%**	77.237%	83.070%	82.323%	80.371%	78.363%	80.646%	80.438%	83.187%
	8	**88.570%**	87.331%	87.936%	87.787%	85.999%	86.330%	85.714%	86.014%	88.017%
	10	**89.484%**	87.574%	88.853%	88.578%	85.979%	86.622%	85.721%	85.783%	89.041%

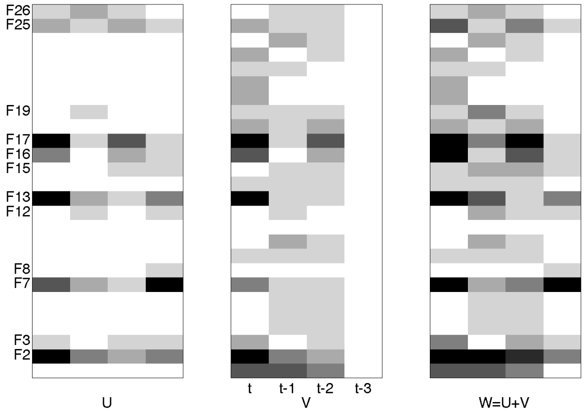

Figure 4: The model constructed by our approach on the NLSY dataset.

ing a within-training two-fold cross validation. The results are reported in Table 3. For any assumption of the working correlation structure, LGL had comparative performance with RE-EM tree and consistently outperformed GEE in all of the three experiments. LGL with tri-diagonal correlation performed the best on this dataset. The results here again show that taking care of the correlation among repeated observations improves the performance (given we see that LGL with the independent correlation assumption had the worst performance among all LGL variants).

The gray map of **U**, **V** and **W** constructed by LGL is shown in Figure 4 to illustrate an example for the tri-diagonal working correlation assumption. Out of the 26 features, 12 were selected by LGL and we list them below.

F2: # days of smoking a cigarette in the past 30 days
F3: Received a training certificate or vocational license
F7: The grade began during the academic year
F8: # months that respondent did not attend school during the academic year
F12: The college degree working toward or attained
F13: The highest grade completed as of the survey year
F15: The highest grade attended as of the survey day
F16: The highest grade completed as of the survey day
F17: # days of using marijuana in the past 30 days
F19: # times of using some drug or other substance right before school or during school or work hours
F25: As the victim of a violent crime in the survey year
F26: Divorced parents.

This list shows that a subject's smoking, drug use, education background and family support influenced his or her drinking behavior. Figure 4 demonstrates that the data in the third prior year might be obsolete to predict this year's behavior as LGL only selected the past two years for use in the model as seen in the plot of **V**.

Table 3: Comparison of different algorithms on the NLSY dataset in terms of test nMSE values.

	LGL				GEE					
# observations	AR(1)	exchangeable	tri-diag	ind	AR	exchangeable	tri-diag	ind	RE-EM tree	Granger
1	0.906552	0.908932	0.904760	0.909446	0.911543	0.918691	0.911885	0.914043	**0.904260**	1.370135
2	0.888608	0.891761	**0.887294**	0.891051	0.898132	0.904225	0.897920	0.898320	0.888822	1.363714
3	0.885448	0.885814	**0.883617**	0.887579	0.892963	0.895863	0.892633	0.890937	0.883958	1.360430

5. DISCUSSION

We have proposed a new learning formulation for longitudinal analytics. Unlike existing methods, the proposed approach can simultaneously determine the temporal contingency and the influential features in predicting an outcome over time. The model parameter matrix is computed by the summation of two component matrices: one matrix reflects the selection among covariates; and the other characterizes the dependency along the temporal line. Moreover, our approach simultaneously models the sample correlations in the longitudinal data while constructing a predictive model. The related optimization problem can be efficiently solved by a new accelerated gradient descent algorithm. Convergence analysis shows that the algorithm can find the global optimal solution for the model with a quadratic convergence rate. An asymptotic analysis shows that the solution of our formulation is a consistent estimate of the model parameters. Hence, the proposed approach solves an underdeveloped problem - jointly learning the relevant features and determining how current outcome relies on past observations. Empirical studies on both synthetic and real-world problems demonstrate the superior performance of the proposed approach over the state of the art.

Acknowledgments

This work was supported by NSF grants IIS-1320586, DBI-1356655 and NIH grant R01DA037349. Jinbo Bi was also supported by NSF grants IIS-1407205 and IIS-1447711.

6. REFERENCES

[1] S. Armeli, T. S. Conner, J. Cullum, and H. Tennen. A longitudinal analysis of drinking motives moderating the negative affect-drinking association among college students. *Psychology of Addictive Behaviors*, 24(1):38–47, 2010.

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

[3] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

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

[5] C. D. Brown and H. T. Davis. Receiver operating characteristics curves and related decision measures: A tutorial. *Chemometrics and Intelligent Laboratory Systems*, 80(1):24–38, 2006.

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

[7] J. H. Fowler and N. A. Christakis. Dynamic spread of happiness in a large social network: Longitudinal analysis over 20 years in the framingham heart study editorial comment. *Journal of Urology*, 181(5):2258–2259, 2009.

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

[9] P. Gong, J. Ye, and C. Zhang. Robust multi-task feature learning. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 895–903, New York, NY, USA, 2012. ACM.

[10] C. W. Granger. Testing for causality: a personal viewpoint. *Journal of Economic Dynamics and Control*, 2:329–352, 1980.

[11] N. M. Laird and J. H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974, 1982.

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

[13] A. Lozano, N. Abe, Y. Liu, and S. Rosset. Grouped graphical granger modeling methods for temporal causal modeling. *Proceedings of the 15th ACM International Conference on Knowledge Discovery and Data Mining*, pages 577–585, 2009.

[14] P. McCullagh and J. A. Nelder. *Generalized linear models (Second edition)*. London: Chapman & Hall, 1989.

[15] C. McCulloch and S. Searle. *Generalized, Linear, and Mixed Models.* Wiley, New York, NY, USA, 2001.

[16] U. Olsson. *Generalized linear models*, volume 18. 2002.

[17] M. R. Osborne, B. Presnell, and B. A. Turlach. On the lasso and its dual. *Journal of Computational and Graphical statistics*, 9(2):319–337, 2000.

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

[19] T. A. Severini. *Elements of Distribution Theory*, volume 17. Cambridge University Press, 2005.

[20] C. A. Stappenbeck and K. Fromme. A longitudinal investigation of heavy drinking and physical dating violence in men and women. *Addict Behav*, 35(5):479–85, 2010.

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

[22] Y. Zhang and D.-Y. Yeung. Multi-task learning using generalized t process. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, 2010.

[23] Y. Zhang, D.-Y. Yeung, and Q. Xu. Probabilistic multi-task feature selection. In *Advances in Neural Information Processing Systems*, pages 2559–2567, 2010.