SPARSE GENERALIZED FUNCTIONAL LINEAR MODEL FOR PREDICTING REMISSION STATUS OF DEPRESSION PATIENTS

YASHU LIU†, ZHI NIE†, JIAYU ZHOU†, MICHAEL FARNUM‡, VAIBHAV A NARAYAN‡, GAYLE WITTENBERG‡, JIEPING YE†

†*Department of Computer Science and Engineering, Center for Evolutionary Medicine and Informatics, The Biodesign Institute, Arizona State University, Tempe, AZ 85287, USA*

‡*Johnson & Johnson Pharmaceutical Research & Development, LLC, Titusville, NJ, USA*

E-mail: †{*Yashu.Liu, Zhi.Nie, Jiayu.Zhou, Jieping.Ye*}*@asu.edu,*

‡{*MFARNUM, VNaray16, GWittenb*}*@its.jnj.com*

Complex diseases such as major depression affect people over time in complicated patterns. Longitudinal data analysis is thus crucial for understanding and prognosis of such diseases and has received considerable attention in the biomedical research community. Traditional classification and regression methods have been commonly applied in a simple (controlled) clinical setting with a small number of time points. However, these methods cannot be easily extended to the more general setting for longitudinal analysis, as they are not inherently built for time-dependent data. Functional regression, in contrast, is capable of identifying the relationship between features and outcomes along with time information by assuming features and/or outcomes as random functions over time rather than independent random variables. In this paper, we propose a novel sparse generalized functional linear model for the prediction of treatment remission status of the depression participants with longitudinal features. Compared to traditional functional regression models, our model enables high-dimensional learning, smoothness of functional coefficients, longitudinal feature selection and interpretable estimation of functional coefficients. Extensive experiments have been conducted on the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) data set and the results show that the proposed sparse functional regression method achieves significantly higher prediction power than existing approaches.

Keywords: Depression, generalized functional linear model, STAR*D, longitudinal analysis, fused Lasso, group Lasso

1. Introduction

The increasing life expectancy of the worldwide population has led to a growing number of patients with serious mental disease such as depression. Research on the diagnosis and prognosis of these diseases has received increasing attention in the biomedical domain. Depression, or major depression (MD) is a common mental disorder affecting estimated 350 million people worldwide, featured by symptoms such as depressed mood, loss of interest or pleasure, feelings of guilt or low self-worth.[1] It is expected to be the second leading cause of disability worldwide.[2] Though the efficacy of several antidepressant medications and therapies has been proven, a universal and long-term treatment of MD has not been well explored due to its high risk of relapses and recurrences.[3]

Like many other mental conditions, major depression affects people over time and it is notorious for the chronicity. Thus, the analysis of longitudinal data is one crucial step towards the understanding and prognosis of major depression. One valuable resource for such research

is the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) trial initiated by National Institute of Mental Health (NIMH), which was originally designed for seeking the optimal combination and sequence of treatment strategies for non-psychiatric depressed patients.[3] Based on the evaluation of the therapeutic responses, participants in STAR*D may receive up to 4 levels of treatments and their information such as symptomatic status, daily functioning, treatment side effects is collected during every clinical visit.

In STAR*D, a range of clinical scales have been applied to evaluate or describe the severity of diseases. For instance, the 17-item Hamilton Rating Scale for Depression ($HRSD_{17}$) is collected via telephone interview for research purposes.[3] The 16-item Quick Inventory of Depressive Symptomatology - Clinician Rated (QIDS-C_{16}) provides the evidences for clinicians to decide whether the patients proceed to the next treatment level.[3] Exploring the longitudinal relationship between clinical measurements (input features) and therapeutic responses (outcomes) and detecting features with significant statistical power are two fundamental and important research questions. Several tools based on machine learning techniques have been developed for longitudinal study.[4–7]

In our paper, we adopt sparse functional regression for the longitudinal data analysis. Functional data (FD) refers to the data samples whose features are viewed as random functions or surfaces over one or more continuum such as time, spatial location.[8,9] For instance, the average daily temperatures observed in a weather station can be viewed as a functional data sample over time; the intensity or color composition of a brain image can be taken as a functional sample over spatial location.[9] Functional data analysis (FDA), an important branch of statistics, is referred to the statistical analysis built on functional data, where the random functions are assumed to be independent and smooth.[8–10] As the extension of classic regression methods to functional data, functional regression is used to estimate the relationship among functional features. Variant forms of functional regression are applicable in different problem setups. For example, it can be applied for regressing functional outcomes on scalar features;[9,11–13] it can also be applied on estimating relationships between functional features and scalar outcomes.[14–16] Under the assumption that the functional coefficient is sparse over time, the FliRTI model was proposed by James *et al.*[15] and it showed better predictive power than regular functional regression models. However, the FliRTI model is only limited to the settings with one functional feature. For higher flexibility, multivariate functional regression models were developed. To enhance interpretability of the multivariate functional regression model, Zhu *et al.*[17] and Gertheiss *et al.*[18] applied the group Lasso type constraint for curve (functional feature) selection. Zhu *et al.*[17] combined both functional features and scalar features together in their model, however it imposed smoothness of coefficient functions only by controlling the number of basis functions. Gertheiss *et al.*[18] introduced the sparsity-smoothness penalty for simultaneously selecting functional feature and controlling the smoothness of the coefficient functions, however it does not incorporate extra scalar features or achieve sparse feature effects over time. Fan *et al.*[19] proposed a functional additive regression (FAR) model which managed functional feature selection via concave penalties in both linear and non-linear settings, while the resulting solutions are not interpretable in term of functional feature effects over time. Therefore, there is a need to develop a general and interpretable formulation

of functional regression that simultaneously achieves functional feature selection, smoothness of functional coefficients and interpretable estimation of functional coefficients.

In this paper, we propose a novel sparse generalized functional linear model for longitudinal biomedical data analysis, which can be applied to predict the disease status based on longitudinal features. Specifically, we empower basic functional regression models to simultaneously identify features with significant predictive power across time points with the group Lasso penalty,[20] enforce smoothness of functional coefficients with the fused Lasso penalty[21] and achieve interpretable estimations of functional coefficients with the Lasso penalty.[22] Since the unknown coefficient matrix is a multiplication factor of the penalized term, the proposed formulation is challenging to solve. Our proposed algorithm integrates the Alternating Direction Method of Multipliers (ADMM)[23] and the accelerated gradient method (AGM)[24,25] to estimate the unknown coefficient matrix. We demonstrate the effectiveness and flexibility of the proposed formulations for longitudinal data analysis using STAR*D data. Experimental results show that the proposed method achieves better prediction performance with longitudinal features than existing approaches.

The rest of the paper is organized as follows. We briefly introduce FDA and the basic functional regression model in section 2. We propose a novel sparse generalized functional linear model and present the algorithm to solve the proposed formulations in section 3. In section 4, we evaluate the proposed sparse generalized functional regression model on STAR*D data and report the experimental results. We conclude our paper in section 5.

2. Basics of Functional Regression

2.1. *Functional Data Analysis*

Functional data is usually assumed to be generated by an underlying smooth function. In practice, a functional data sample consists of sequences of numerical values (or vectors) varying over a certain continuum. For instance, the series of QIDS-C_{16} scores of a depression patient over his/her visiting time can be considered as a functional data sample. Fig. 1 gives an illustration of functional data. Sequences of QIDS-C_{16} scores of 6 depression patients are recorded over 14 weeks. In the functional context, we assume each sequence of QIDS-C_{16} scores is generated by an underlying function varying over time t (weeks).

One important issue in FDA is to recover the underlying function based on the sequences of observed numerical values. A common approach is to express the underlying function by a linear combination of basis functions using smoothing techniques. Specifically, given the evaluations of basis functions over time as features and the observed numerical values as outcomes, the coefficients of basis functions can be fitted using least square methods with roughness penalty.[9] However, the basis smoothing technique is only effective when the functional data is observed continuously or densely. When it comes to longitudinal data, observations are always sparse and irregular. Rice *et al.*[26] contrasted and compared FDA with longitudinal data analysis (LDA). James *et al.*[27] and Yao *et al.*[28] connected FDA and LDA by proposing approaches that estimate the underlying function of sparsely and irregularly observed functional data by exploiting both population and individual information. The former extended the basis smoothing technique with mixed effect models while the latter proposed the "PACE" method

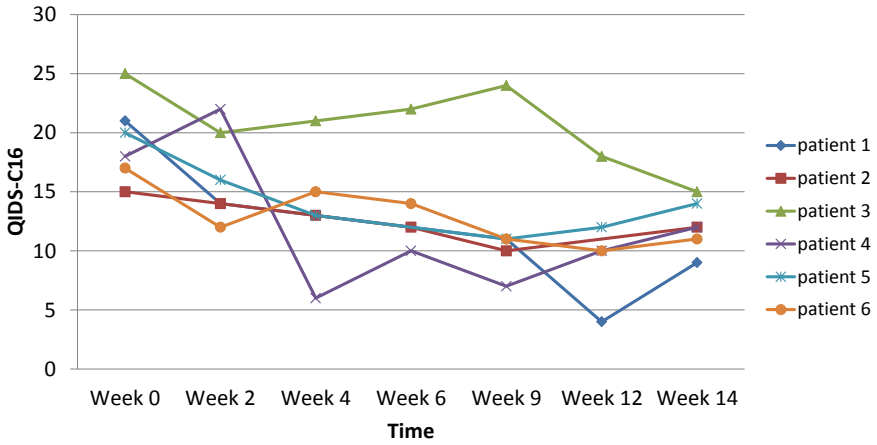

Fig. 1. Illustration of functional data. The QIDS-C_{16} scores of 6 patients are recorded over 14 weeks. In the functional context, we assume each sequence of QIDS-C_{16} scores is generated by an underlying function.

which involves the kernel smoothing technique and computing the conditional expectation. In our paper, we adopt PACE to estimate the underlying function of functional data.

2.2. *Functional Regression Model with Functional Features and Scalar Outcomes*

In classic statistical analysis, regression methods play an important role in analyzing the relationship between features (independent variables) and outcomes (dependent variables). FDA extends the philosophy of classic regression to functional data and develops functional regression which involves various models for different purposes.

When the features are functional and the outcomes are scalar, we have

$$Y_i = \alpha + \int_{\Omega_t} X_i(t)\beta(t)dt + \epsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where n is the total number of samples, Y_i is a scalar outcome of the ith sample, $X_i(t)$ is a 1-dimensional functional feature of the ith sample, $\beta(t)$ is a functional coefficient, Ω_t is the domain of continuum t, scalar α is a bias term, and ϵ_i corresponds to the scalar residual. Note that the model above only allows one functional feature, which greatly limits its application. A simple but useful extension to multiple functional features is

$$Y_i = \alpha + \sum_{k=1}^{p} \int_{\Omega_t} X_{ik}(t)\beta_k(t)dt + \epsilon_i, \quad i = 1, \ldots, n, \tag{2}$$

where $X_{ik}(t)$ is the kth functional feature of the ith sample, and $\beta_k(t)$ is the functional coefficient corresponding to the kth functional feature.

3. Proposed Sparse Functional Regression Models

In this section, we propose a novel sparse generalized functional linear model which simultaneously selects useful features, enforces smoothness of functional coefficients and achieves interpretable estimations of functional coefficients. Suppose there are N samples and each sample has d scalar features $s_1, s_2, \ldots, s_d$ and p functional features $x_1(t), x_2(t), \ldots, x_p(t)$. Then, for

the ith sample, we denote $\mathbf{s}_i = (s_{i1}, s_{i2}, \ldots, s_{id})^T$ and $\mathbf{x}_i(t) = (x_{i1}(t), x_{i2}(t), \ldots, x_{ip}(t))^T$. In matrix form, we let $S \in \mathbb{R}^{N\times d}$ be the scalar data matrix with each row as a sample of d scalar features, *i.e.*, $S_{i,\cdot} = \mathbf{s}_i^T = (s_{i1}, s_{i2}, \ldots, s_{id})$, and let $X(t)$ be an $N \times p$ matrix of functions where the ith row denotes the ith sample of p functional features *i.e.*, $X_{i,\cdot}(t) = \mathbf{x}_i(t)^T = (x_{i1}(t), x_{i2}(t), \ldots, x_{ip}(t))$. Let $\mathbf{y} = (y_1, y_2, \ldots, y_N)^T \in \mathbb{R}^{N\times 1}$ be the vector of scalar outcomes, and $\mathbf{w} = (w_1, w_2, \ldots, w_d)^T$ be the coefficients of d scalar features and $\mathbf{b}(t) = (\beta_1(t), \beta_2(t), \ldots, \beta_p(t))^T$ be the vector of p functional coefficients. Moreover, we assume the functional coefficients can be represented by a set of k_b basis functions $\mathbf{\Theta}(t) = (\theta_1(t), \theta_2(t), \ldots, \theta_{k_b}(t))^T$, *i.e.,* $\mathbf{b}(t) = B\mathbf{\Theta}(t)$, where $B \in \mathbb{R}^{p\times k_b}$. Then, for a known link function $g(\cdot)$, we have the generalized functional linear model

$$g(y_i) = \alpha + \sum_{g=1}^{d} s_{ig} w_g + \sum_{h=1}^{p} \int_{\Omega_t} x_{ih}(t)\beta_h(t)dt = \alpha + \mathbf{s}_i\mathbf{w} + \int_{\Omega_t} \mathbf{x}_i(t)\mathbf{b}(t)dt. \tag{3}$$

In matrix form, we have

$$g(\mathbf{y}) = \alpha\mathbf{1} + S\mathbf{w} + \int_{\Omega_t} X(t)\beta(t)dt = \alpha\mathbf{1} + S\mathbf{w} + \int_{\Omega_t} X(t)B\mathbf{\Theta}(t)dt, \tag{4}$$

where $\mathbf{1} \in \mathbb{R}^{N\times 1}$ is a column vectors of ones. When $g(\cdot)$ is the identity function, *i.e.*, $g(u) = u$, the proposed model is functional linear regression. Then the optimization procedure involves minimizing the quadratic loss. When $g(\cdot)$ is the sigmoid function, the proposed model turns out to be a functional logistic regression, *i.e.*,

$$\text{Prob}(\mathbf{y}|S, X) = \frac{1}{1 + \exp\left(-\mathbf{y} \odot \left(\alpha\mathbf{1} + S\mathbf{w} + \int_{\Omega_t} X(t)B\mathbf{\Theta}(t)dt\right)\right)}, \tag{5}$$

where "$\odot$" denotes the componentwise multiplication. Let $\Theta \in \mathbb{R}^{k_b\times T}$ be the evaluation matrix of $\mathbf{\Theta}(t)$ at T time points, where $\Theta_{\cdot,t} \in \mathbb{R}^{k_b\times 1}$ corresponds to the evaluation at time t. Then the unknown coefficients $\alpha, \mathbf{w}$ and matrix B can be obtained by minimizing the average logistic loss (negative log-likelihood function),

$$\mathcal{L}(\alpha, \mathbf{w}, B) = \frac{1}{N}\sum_{i=1}^{N} \log\left(1 + \exp\left(-y_i(\alpha + \mathbf{s}_i\mathbf{w} + \sum_t X(t)B\Theta_{\cdot,t}\right)\right)). \tag{6}$$

The logistic loss is convex and smooth and can be solved via standard optimization methods.

When $\beta_j(t) = 0$, the changes of the jth functional feature has no effect on the outcome at time t. We apply the Lasso penalty[22] on $B\Theta$, *i.e.*, $\|B\Theta\|_1 = \sum_{j,t} |B_{j,\cdot}\Theta_{\cdot,t}|$, resulting in interpretable estimations, *i.e.*, many entries of $B\Theta$ are zero. That is, the changes of many features have no effects on prediction at some time points. For feature selection purpose, we also introduce the group Lasso penalty[20] with $\|B\Theta\|_{2,1} = \sum_{j=1,\ldots,p} \|B_{j,\cdot}\Theta\|_2$ which enforces many rows of $B\Theta$ to be zero. If the jth row of matrix $B\Theta$ is zero, then the jth feature has no predictive power along with time. In addition, we employ the fused Lasso penalty[21] on matrix $B\Theta$ to enforce the smoothness of the functional coefficients. The resulting sparse functional logistic regression model with scalar features and functional outcomes can be obtained by

$$\min_{\alpha,\mathbf{w},B} \mathcal{L}(\alpha, \mathbf{w}, B) + \lambda_1\|\mathbf{w}\|_1 + \lambda_2\|B\Theta\|_1 + \lambda_3\|B\Theta R\|_1 + \lambda_4\|B\Theta\|_{2,1}, \tag{7}$$

where R is a T by $T-1$ sparse matrix with $R_{j,j} = 1, R_{j+1,j} = -1$, and $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are the tuning parameters. We solve problem (7) by alternately minimizing over α and $\mathbf{w}$ with B

fixed, and minimizing over B with α and $\mathbf{w}$ fixed. Note that the penalty of $\mathbf{w}$ only involves the Lasso penalty, and we have already known that the optimization in terms of $\mathbf{w}$ (*i.e.,* sparse logistic regression problem) can be solved efficiently by the Accelerated Gradient Methods (AGM).[24,25,29]

The proposed sparse functional logistic regression model is much more challenging to solve than usual multi-task learning algorithms since the unknown coefficient matrix B is a multiplication factor of the penalized term $B\Theta$. In this paper, we integrate the Alternating Direction Method of Multipliers (ADMM)[23] and AGM[24,25] to solve B. When α and $\mathbf{w}$ are fixed, we write the objective (7) in the following form:

$$\begin{aligned} &\min_{B} \mathcal{L}(\alpha, \mathbf{w}, B) + \lambda_2\|Z\|_1 + \lambda_3\|ZR\|_1 + \lambda_4\|Z\|_{2,1} \\ &s.t.\ B\Theta = Z. \end{aligned} \tag{8}$$

Then, the augmented Lagrangian function is given by

$$L_\rho(B, Z, \xi) = \mathcal{L}(\alpha, \mathbf{w}, B) + \lambda_2\|Z\|_1 + \lambda_3\|ZR\|_1 + \lambda_4\|Z\|_{2,1} + \langle \xi, B\Theta - Z\rangle + \frac{\rho}{2}\|B\Theta - Z\|_F^2, \tag{9}$$

where $\xi \in \mathbb{R}^{p\times T}$ is the lagrangian dual variable and ρ is a penalty parameter. The ADMM-based procedures for solving the unknown matrix B in the proposed sparse functional logistic regression at the kth iteration can be described as follows:[23]

$$B^{(k)} := \min_B \mathcal{E}(B) = \min_B \mathcal{L}(\alpha^{(k)}, \mathbf{w}^{(k)}, B) + \langle \xi^{(k-1)}, B\Theta - Z^{(k-1)}\rangle + \frac{\rho}{2}\|B\Theta - Z^{(k-1)}\|_F^2, \tag{10}$$

$$Z^{(k)} := \min_Z \lambda_2\|Z\|_1 + \lambda_3\|ZR\|_1 + \lambda_4\|Z\|_{2,1} + \langle \xi^{(k-1)}, B^{(k)}\Theta - Z\rangle + \frac{\rho}{2}\|B^{(k)}\Theta - Z\|_F^2, \tag{11}$$

$$\xi^{(k)} := \xi^{(k-1)} + \rho(B^{(k)}\Theta - Z^{(k)}). \tag{12}$$

For the B-update step (10), the unknown matrix B can be solved by the accelerated gradient descent method[24,25,29] with gradient

$$\nabla_B \mathcal{E}(B) = -\frac{1}{N}U^T(\mathbf{1} - \mathbf{p}) + \rho(B\Theta + \frac{\xi^{(k-1)}}{\rho} - Z^{(k-1)})\Theta^T,$$

where

$$U = [y_1 \sum_t X_{1,\cdot}(t)^T\Theta_{\cdot,t}^T, y_2 \sum_t X_{2,\cdot}(t)^T\Theta_{\cdot,t}^T, \ldots, y_N \sum_t X_{N,\cdot}(t)^T\Theta_{\cdot,t}^T]^T,$$

$$\mathbf{p} = \mathbf{1}./\left(\mathbf{1} + \exp\left(-\mathbf{y} \odot \left(\alpha^{(k)} + S\mathbf{w}^{(k)} + \sum_t X(t)B\Theta_{\cdot,t}\right)\right)\right),$$

and "./" denotes the componentwise division. For the Z-update step (11), it has been shown that the proximal operator can be solved efficiently in two stages as stated in the following lemma,[30]

Lemma 3.1. *Given vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{1\times T}$, penalty parameters γ_1, γ_2, γ_3, and the sparse matrix R defined as above, let*

$$\begin{aligned} \mathcal{F}(\mathbf{v}) &= \arg\min_{\mathbf{u}} \frac{1}{2}\|\mathbf{u} - \mathbf{v}\|_2^2 + \gamma_1\|\mathbf{u}\|_1 + \gamma_2\|\mathbf{u}R\|_1, \\ \mathcal{G}(\mathbf{v}) &= \arg\min_{\mathbf{u}} \frac{1}{2}\|\mathbf{u} - \mathbf{v}\|_2^2 + \gamma_3\|\mathbf{u}\|_2, \\ \mathcal{FG}(\mathbf{v}) &= \arg\min_{\mathbf{u}} \frac{1}{2}\|\mathbf{u} - \mathbf{v}\|_2^2 + \gamma_1\|\mathbf{u}\|_1 + \gamma_2\|\mathbf{u}R\|_1 + \gamma_3\|\mathbf{u}\|_2. \end{aligned} \tag{13}$$

Then, the following relationship holds

$$\mathcal{F}\mathcal{G}(\mathbf{v}) = \mathcal{G}(\mathcal{F}(\mathbf{v})). \tag{14}$$

The details of the proposed algorithm for solving (7) are summarized in Algorithm 1. Steps 3 and 4 are solved by AGM, and readers may refer to Liu *et al.*[29] for more details. In terms of solving the fused Lasso problem (step 6), readers may refer to Liu *et al.*[31] The ADMM parameter ρ is fixed as a constant in our experiment.

Algorithm 1 Sparse Functional Logistic Regression

Input: $\mathbf{y} \in \mathbb{R}^{N\times 1}$, $S \in \mathbb{R}^{N\times d}$, $X_{(j)} \in \mathbb{R}^{N\times p}, j = 1,\ldots,T$, $\Theta \in \mathbb{R}^{k_b\times T}$, $R \in \mathbb{R}^{T\times(T-1)}$, $\rho \in \mathbb{R}$
Output: $\alpha \in \mathbb{R}$, $\mathbf{w} \in \mathbb{R}^{d\times 1}$, $B \in \mathbb{R}^{p\times k_b}$

1: Initialize starting points $\alpha^{(0)}$, $\mathbf{w}^{(0)}$, $B^{(0)}$, $\xi^{(0)}$, $Z^{(0)}$
2: **for** $k = 1 : \mathcal{K}$ **do**
3: $(\alpha^{(k)}, \mathbf{w}^{(k)}) := \arg\min_{\alpha,\mathbf{w}} \mathcal{L}(\alpha, \mathbf{w}, B^{(k-1)}) + \lambda_1\|\mathbf{w}\|_1$
4: $B^{(k)} := \arg\min_B \mathcal{L}(\alpha^{(k)}, \mathbf{w}^{(k)}, B) + \langle \xi^{(k-1)}, B\Theta - Z^{(k-1)}\rangle + \frac{\rho}{2}\|B\Theta - Z^{(k-1)}\|_F^2$
5: **for** $i = 1 : p$ **do**
6: $u_i := \arg\min_z \frac{\rho}{2}\|B_{i,\cdot}^{(k)}\Theta + \xi_{i,\cdot}^{(k-1)}/\rho - z\|_2^2 + \lambda_2\|z\|_1 + \lambda_3\|zR\|_1$
7: $Z_{i,\cdot}^{(k)} := \arg\min_z \frac{\rho}{2}\|u_i - z\|_2^2 + \lambda_4\|z\|_2$
8: **end for**
9: $\xi^{(k)} := \xi^{(k-1)} + \rho(B^{(k)}\Theta - Z^{(k)})$
10: **end for**

4. Experiments

In this section, we evaluate the proposed sparse functional logistic regression model on the STAR*D data set. We use the functional data analysis code[a] for the construction and evaluation of basis functions. We also use the PACE package[b] for estimating the underlying smooth functions of functional data.

4.1. *STAR*D Data Set*

The STAR*D project consists of four treatment levels aimed to help outpatients achieve depressive symptom remission with measurement-based care treatment.[32] Throughout STAR*D study, the QIDS-C_{16} score, which measures the general symptoms of depression, provides clinicians evidences for deciding the remission status of patients.[32] All the participants enrolled to STAR*D receive the same antidepressant treatment at level 1, where the selective serotonin reuptake inhibitor citalopram is used. If the participant's therapeutic response is satisfactory, *i.e.*, QIDS-C_{16} <= 5, he/she will be recommended to the follow-up phase. If the initial therapy is not sufficiently effective on the participants, they will be recommended to the level 2 treatment. At level 2, participants will enter into a set of randomized clinical trials. In a similar

[a]http://www.psych.mcgill.ca/misc/fda/software.html
[b]http://www.stat.ucdavis.edu/PACE/

manner, participants who fail to achieve satisfactory responses will enter level 2A, level 3 and up to level 4. In this paper, we concentrate on the longitudinal analysis of the data collected from level 1 and level 2. Since all the participants receive the same treatment at level 1, the questions we aim to address in this paper are: Can we use the level 1 information to predict the participant's remission status at level 2? Which features are most important for the prediction of remission status? How do the important longitudinal features affect the prediction over time?

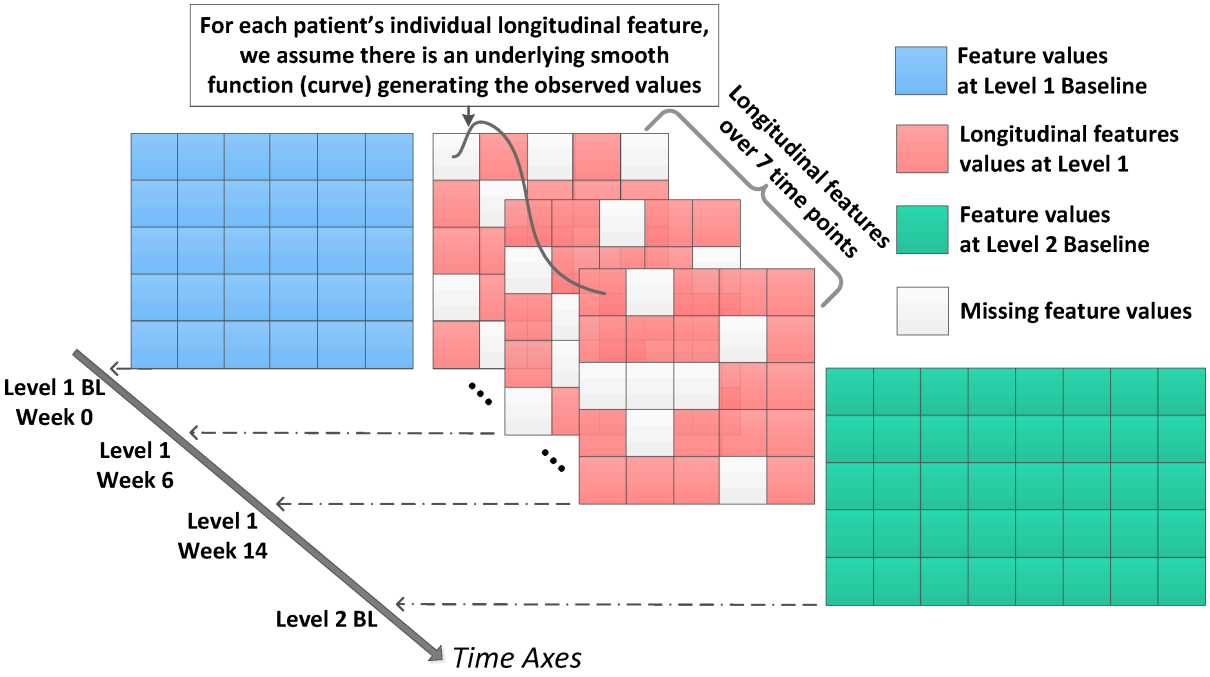

Fig. 2. This figure gives a brief description of the STAR*D data used in our experiment. In our experiment, we use the scalar features at level 1 baseline and level 2 baseline and the longitudinal features at level 1. Note that the longitudinal data is observed sparsely and irregularly. Therefore, if we align the longitudinal data in a regular time grid as shown in the figure, some feature values will appear as the "missing" values. In our experiment, we first estimate the underlying curves using PACE[28] and then evaluate them on a dense grid of time points.

At each level of treatment, clinical visit information including medication names and doses, side effect intensity, burden and frequency, is collected every 2 or 3 weeks during the acute treatment stage. In our study, we name the time point based on the duration time from baseline to the clinical visit, *e.g.*, "W2" refers to the time point 2 weeks after baseline. At level 1, there are 7 time points scheduled for the regular visits, *i.e.*, W0, W2, W4, W6, W9, W12, W14. Fig. 2 gives a brief description of the STAR*D used in our experiment. The red matrices in Fig. 2 refer to the longitudinal data stated above. Besides the longitudinal data, the enrolment information involving Cumulative Illness Rating Scale, demographics, HRSD, Medication History, Protocol Eligibility, Psychiatric History, and some level 1 W0 information without further follow-up are also available for the prediction. Those information refers to the level 1 baseline scalar features and corresponds to the blue matrix in Fig. 2. Moreover, we also

want to test the predictive power of the information collected at level 2 baseline (*i.e.*, week 0 at level 2) since the participants did not receive new treatments at that time. These level 2 baseline scalar features are diagramed as the green matrix in Fig. 2.

In STAR*D, there are 730 participants in total entering level 2 with their level 1 longitudinal information recorded until W12 or W14. After eliminating extremely sparse observations, we have a total number of 596 samples with 202 longitudinal features (LV1_L) available for analysis. There are 1438 level 1 baseline scalar features (LV1_S) and 1667 level 2 baseline scalar features (LV2_S) where the missing values for categorical features are imputed by zeros and the missing values for continuous features are imputed by the mean values. For research purposes, the 16-item Quick Inventory of Depressive Symptomatology – Self-Report (QIDS-SR_{16}) is used as outcomes in many existing studies.[32] In our experiments, we also adopt QIDS-SR_{16} as the criterion of remission, and define QIDS-SR_{16} $<= 5$ as remission and a QIDS-SR score of > 5 as non-remission. We evaluate our proposed sparse functional logistic regression model on differentiating the remission cohort from non-remission cohort with available level 1 and level 2 features. Moreover, we classify the remission samples and a subgroup of non-remission samples whose QIDS-SR_{16} $>= 11$; this subgroup is sometimes referred to as severe depression.[32] Detailed sample statistics are shown in Table 1.

Table 1. The sample statistics of the STAR*D data used in our experiments. Group (All) refers to all qualified samples for the experiment; and Group (Sub) refers to the remaining samples after removing those with QIDS-SR_{16} between 6 and 10.

Cohort	Remission (+)	Non-remission (−)	Total	Data Name	Dim
Group (All)	240	356	596	LV1_S	1438
Group (Sub)	240	161	401	LV1_L	202
				LV2_S	1667

4.2. *Predicting Remission Status at Level 2*

We compare the proposed sparse functional logistic regression with two classic multivariate classifiers, *i.e.*, Random Forest and sparse logistic regression on exactly the same training and testing sets. Our report presents the average accuracy, sensitivity and specificity and the corresponding standard deviations obtained from the 5-fold cross-validation. In all the experiments, both the parameters of sparse functional logistic regression and the sparse logistic regression are tuned via 5-fold cross-validation in the training process. We use B-spline basis functions in our proposed sparse functional logistic regression model, which are the common choice for approximating non-periodic functions.[9]

We first conduct the classification experiment on the level 1 longitudinal data. For classifiers such as Random Forest and sparse logistic regression, the input data is the the average of the longitudinal features over time. The detailed report is shown in Table 2. Compared with Random Forest and sparse logistic regression, the classification performance achieved by the proposed sparse functional logistic regression is consistently better demonstrating the effectiveness of the sparse functional logistic regression in capturing the temporal information. In addition, we observe that, when the sparse functional logistic regression is applied, the level

Table 2. Comparisons of classification performance between Random Forest, sparse logistic regression and sparse functional logistic regression on the longitudinal STAR*D data.

Experimental Results Using All Samples			
	LV1_L	LV1_S+LV1_L	LV1_S+LV1_L+LV2_S
Random Forest			
Accuracy (%)	68.63 ± 1.16	69.47 ± 2.23	68.46 ± 3.12
Sensitivity(%)	68.83 ± 3.31	68.83 ± 4.28	67.71 ± 5.17
Specificity(%)	68.33 ± 2.72	70.42 ± 2.72	69.58 ± 2.38
Sparse Logistic Regression			
Accuracy (%)	65.94 ± 2.31	66.28 ± 3.19	63.76 ± 2.30
Sensitivity(%)	64.61 ± 5.19	66.02 ± 4.92	60.41 ± 4.43
Specificity(%)	67.92 ± 6.00	66.67 ± 2.08	68.75 ± 6.91
Sparse Functional Logistic Regression			
Accuracy (%)	**69.79 ± 2.38**	**70.30 ± 1.60**	**70.30 ± 1.95**
Sensitivity(%)	**70.83 ± 5.31**	**72.50 ± 5.78**	**73.33 ± 6.97**
Specificity(%)	**69.10 ± 3.20**	**68.82 ± 2.48**	**68.27 ± 3.12**
Experimental Results Using Samples With QIDS_C$_{16}$ ≤ 5 and QIDS_C$_{16}$ ≥ 11			
	LV1_L	LV1_S+LV1_L	LV1_S+LV1_L+LV2_S
Random Forest			
Accuracy (%)	73.84 ± 6.66	74.09 ± 6.56	74.08 ± 4.74
Sensitivity(%)	75.25 ± 11.50	75.23 ± 8.44	73.35 ± 10.46
Specificity(%)	72.92 ± 4.89	73.33 ± 6.32	74.58 ± 2.72
Sparse Logistic Regression			
Accuracy (%)	72.58 ± 3.99	73.08 ± 4.23	74.82 ± 4.53
Sensitivity(%)	68.37 ± 7.11	70.87 ± 7.41	75.19 ± 9.74
Specificity(%)	75.42 ± 3.09	74.58 ± 3.09	74.58 ± 2.28
Sparse Functional Logistic Regression			
Accuracy (%)	**77.81 ± 4.11**	**77.82 ± 4.89**	**77.07 ± 4.54**
Sensitivity(%)	**79.17 ± 2.08**	**78.75 ± 4.01**	**77.92 ± 3.78**
Specificity(%)	**75.81 ± 10.77**	**76.46 ± 9.51**	**75.83 ± 10.45**

1 and level 2 baseline scalar features are not helpful for improving classification performance. We obtain similar observations when applying Random Forest and sparse logistic regression. Since only using longitudinal data at level 1 leads to satisfactory classification performance, we may conclude that most of the information in the level 1 and level 2 baseline data is captured by the longitudinal data at level 1. The experimental results further demonstrate the importance of mining longitudinal data.

Besides the superior predictive performance, the sparse functional logistic regression is also capable of selecting significant longitudinal features and giving interpretable solutions. In our experiments, most meaningful and interesting longitudinal features including side effect frequency, side effect burden, and QIDS-C current score, are selected as a result of the group sparsity constraint. Moreover, the functional coefficients can be visualized to provide deeper insights for understanding the effects of longitudinal features on predicting remission status over time. In Fig. 3, we show 6 important functional coefficients obtained from the task of differentiating participants with QIDS-SR$_{16}$ <= 5 from those with QIDS-SR$_{16}$ >= 11. From the figure, we can see that the effect of FG-FISGQ = 0 (side effect frequency) decreases from W0 to W6 and then increases over time. For feature FG-GRSEB = 0 (side effect burden),

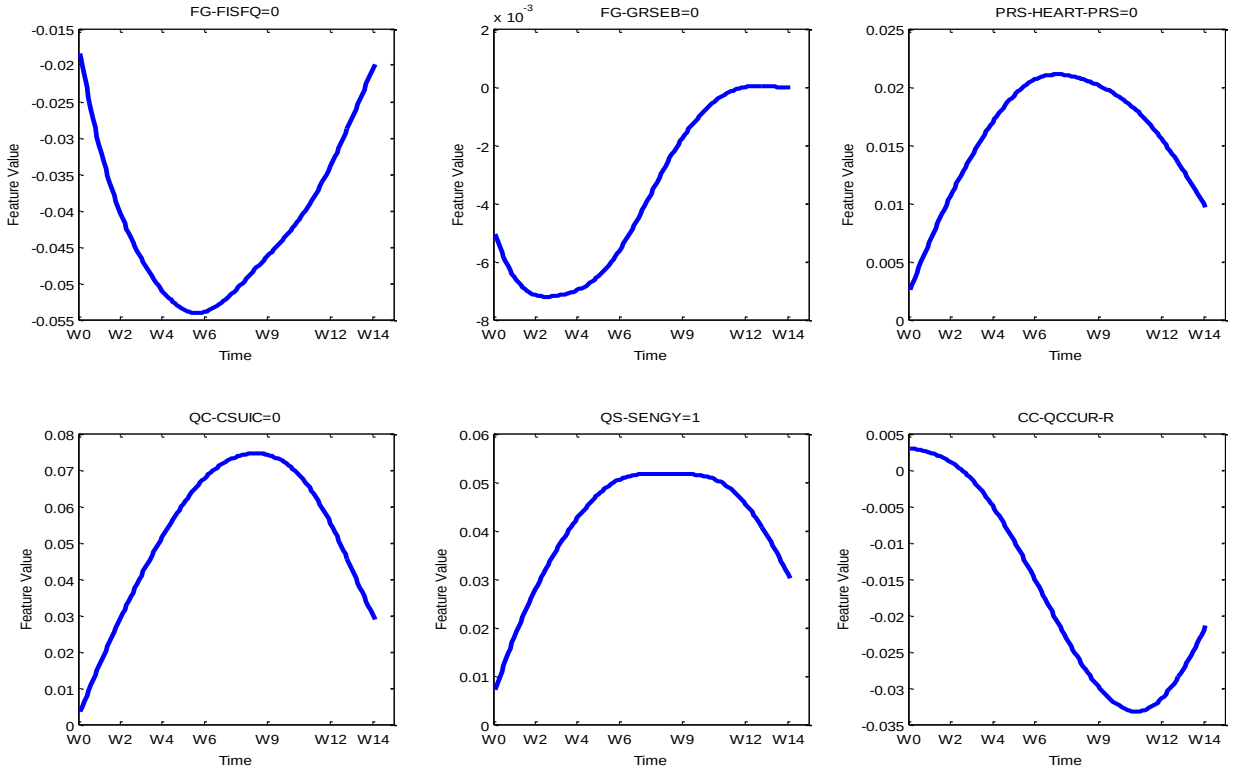

Fig. 3. The functional coefficients of some selected longitudinal features via sparse functional logistic regression. Feature FG-FISGQ= 0 relates to side effect frequency; feature FG-GRSEB= 0 relates to side effect burden; feature PRS-HERAT-PRS= 0 relates to clinical measurements of heart; feature QC-CSUIC= 0 relates to QIDS Suicidal ideation; feature QS-SENGY= 0 relates to QIDS Energy/fatigability; and feature CC-QCCUR= R relates to QIDS-C current score.

we observe that the effects are zero during W12 to W14, which indicates the corresponding feature values during that period make no contributions to the prediction. The sparse and interpretable results are due to the use of Lasso and fused Lasso penalty. From all the plots, we also observe that all the obtained functional coefficients are smooth, which demonstrates the effectiveness of the fused lasso penalty in controlling the smoothness of coefficient functions.

5. Conclusions

In this paper, we propose a novel sparse generalized functional linear model for the longitudinal analysis of STAR*D data. Compared to traditional functional regression models, our model has the advantages of simultaneously achieving high-dimensional learning, smoothness of functional coefficients, longitudinal feature selection and interpretable estimation of functional coefficients. We conduct extensive experiments on the STAR*D data set and the experimental results demonstrate that the proposed sparse functional regression model achieves significantly higher longitudinal prediction power than existing approaches.

Since the proposed models have shown great effectiveness in capturing temporal information, we intend to apply them to investigate the predictive effects of the biomarkers on other longitudinal problems such as the progression of Alzheimer's Disease (AD). Moreover, we plan

to further study the theoretical properties of the proposed models in the future.

6. Acknowledgments

This work is support by NIH (R01 LM010730) and NSF (IIS-0953662, MCB-1026710, and CCF-1025177).

References

1. World Federation for Mental Health, *Archives of Neurology* (2012).
2. A. Haden and B. Campanini, *The World Health Report 2001: Mental health: new understanding, new hope* (World health organization (WHO), 2001).
3. M. Fava, *et al.*, *Psychiatric Clinics of North America* **26**, 457 (2003).
4. H. Wang, *et al.*, *Bioinformatics* **28**, i619 (2012).
5. H. Wang, *et al.*, High-order multi-task feature learning to identify longitudinal phenotypic markers for alzheimer's disease progression prediction, in *NIPS*, 2012.
6. J. Zhou, L. Yuan, J. Liu and J. Ye, A multi-task learning formulation for predicting disease progression, in *ACM SIGKDD*, 2011.
7. J. Zhou, J. Liu, V. A. Narayan and J. Ye, *NeuroImage* **78**, 233 (2013).
8. H.-G. Müller, *StatProb: The Encyclopedia Sponsored by Statistics and Probability Societies*.
9. J. Ramsay and B. Silverman, *Functional Data Analysis*, Springer Series in Statistics (Springer, 2005).
10. F. Ferraty, P. Vieu, *Nonparametric functional data analysis: theory and practice* (Springer, 2006).
11. Q. Shen and J. Faraway, *Statistica Sinica* **14**, 1239 (2004).
12. J.-M. Chiou, H.-G. Muller, J.-L. Wang *et al.*, *Statistica Sinica* **14**, 675 (2004).
13. X. Yang, Q. Shen, H. Xu and S. Shoptaw, *Statistics in medicine* **26**, 1552 (2007).
14. T. T. Cai and P. Hall, *The Annals of Statistics* **34**, 2159 (2006).
15. G. M. James, J. Wang and J. Zhu, *The Annals of Statistics* **37**, 2083 (2009).
16. H. Cardot, *et al.*, *Computational statistics & data analysis* **51**, 4832 (2007).
17. H. Zhu and D. D. Cox, *Lecture Notes-Monograph Series* , 173 (2009).
18. J. Gertheiss, A. Maity and A.-M. Staicu, *Stat* (2013).
19. Y. Fan and G. M. James, Functional additive regression, *Under Review*, 2012.
20. J. Friedman, T. Hastie and R. Tibshirani, *arXiv preprint arXiv:1001.0736* (2010).
21. R. Tibshirani, *et al.*, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**, 91 (2004).
22. R. Tibshirani, *Journal of the Royal Statistical Society. Series B (Methodological)*, 267 (1996).
23. S. Boyd, *et al.*, *Foundations and Trends® in Machine Learning* **3**, 1 (2011).
24. Y. Nesterov, A method of solving a convex programming problem with convergence rate $O(1/k^2)$, in *Soviet Mathematics Doklady*, (2)1983.
25. Y. Nesterov, *Mathematical Programming* **103**, 127 (2005).
26. J. A. Rice, *Statistica Sinica* **14**, 631 (2004).
27. G. M. James, T. J. Hastie and C. A. Sugar, *Biometrika* **87**, 587 (2000).
28. F. Yao, *et al.*, *Journal of the American Statistical Association* **100**, 577 (2005).
29. J. Liu, J. Chen and J. Ye, Large-scale sparse logistic regression, in *ACM SIGKDD*, 2009.
30. J. Zhou, J. Liu, V. A. Narayan and J. Ye, Modeling disease progression via fused sparse group lasso, in *ACM SIGKDD*, 2012.
31. J. Liu, *et al.*, An efficient algorithm for a class of fused lasso problems, in *ACM SIGKDD*, 2010.
32. A. Rush, *et al.*, *American Journal of Psychiatry* **163**, 1905 (2006).