Published in final edited form as:

Stat Appl Genet Mol Biol. 2015 February ; 14(1): 93–111. doi:10.1515/sagmb-2014-0004.

Regularization Method for Predicting an Ordinal Response Using Longitudinal High-dimensional Genomic Data

Jiayi Hou and

Department of Biostatistics, Virginia Commonwealth University. houj2@vcu.edu

Kellie J. Archer

Department of Biostatistics, Virginia Commonwealth University, 830 East Main St., Room 718, Richmond, VA 23298-0032, United States. kjarcher@vcu.edu

Abstract

An ordinal scale is commonly used to measure health status and disease related outcomes in hospital settings as well as in translational medical research. In addition, repeated measurements are common in clinical practice for tracking and monitoring the progression of complex diseases. Classical methodology based on statistical inference, in particular, ordinal modeling has contributed to the analysis of data in which the response categories are ordered and the number of covariates (p) remains smaller than the sample size (n). With the emergence of genomic technologies being increasingly applied for more accurate diagnosis and prognosis, high-dimensional data where the number of covariates (p) is much larger than the number of samples (n), are generated. To meet the emerging needs, we introduce our proposed model which is a two-stage algorithm: Extend the Generalized Monotone Incremental Forward Stagewise (GMIFS) method to the cumulative logit ordinal model; and combine the GMIFS procedure with the classical mixed-effects model for classifying disease status in disease progression along with time. We demonstrate the efficiency and accuracy of the proposed models in classification using a time-course microarray dataset collected from the Inflammation and the Host Response to Injury study.

Keywords

classification; high-dimensional data; longitudinal data; ordinal response; prediction; regularization methods

1 Introduction

Longitudinal data analysis has been playing a profound and irreplaceable role in analyzing clustered and correlated data in a variety of fields. Substantial effort has been devoted to developing statistical models and inferential procedures for explaining the relationship between independent and dependent variables. However, little work has been done in developing variable selection methods for longitudinal and clustered data, especially for longitudinal high-dimensional data ($n \ll p$) with a repeatedly measured outcome. In the

past, the formidable cost of generating and collecting longitudinal high-throughput genomic data created a lag in the application of genomic technologies for longitudinal monitoring in clinical practice and translational research. Since the 1990s, with the emergence of the ‘omics’ revolution, the cost of genomic technologies has tremendously dropped. There is now a need to develop cutting-edge data mining algorithms that are able to identify a few prognostic biomarkers from tens of thousands of genes to track the progression of complex diseases at the molecular level. However, due to confounding and multivariate dependencies in longitudinal high-dimensional data, it is very challenging to select a set of good features for building a parsimonious model and thus enhance predictability. In the past, several authors have developed methodologies for analyzing time-course gene expression data. For example, Yuan and Kendzierski (2006) proposed Hidden Markov Models to detect differential expression patterns in genes over time under multiple biological conditions. Tai and Speed (2006) developed a variation of Hotelling T^2 -statistics and the multivariate empirical Bayes statistics to detect temporal changes of certain genes in one- and two-sample cases where the expression level changes often motivate the experiments. Zhou et al. (2010) developed a factorial design by pooling information across the time course while accounting for multiple testing and nonnormality of the microarray data to effectively extract dynamic features. Zhang et al. (2013) proposed a novel ‘optimal direction’ approach to extract useful features to perform binary classification problems in a time-course microarray setting. However, most of the approaches for extracting temporal patterns of differential expression requires fixed and equal number of time points between heterogeneous samples which tremendously limits usage. Most recently, tremendous efforts have been devoted to develop a more general variable selection procedure under the generalized linear mixed model (GLMM) framework by implementing the penalization technique. Notably, Yang (2007) proposed a two-stage penalized quasi-likelihood algorithm (TPQL) to correct the biased parameter estimates for high-dimensional sparse binary longitudinal data. Schellhdorfer et al. (2013) developed a Lasso-type regularization with a cyclic coordinate descent optimization. Groll and Tutz (2012) developed a gradient ascent algorithm to maximize the approximation form of the penalized quasi-likelihood with inclusion of the L_1 penalty. The corresponding `glmmLasso` (Groll, 2014) is available on CRAN. A similar R package `glmmixedlasso` (Schellhdorfer, 2014) is also available. While most of the Lasso-type approaches for GLMM are focused on longitudinal or clustered binary and count data, to the best of our knowledge, there has not been any literature focusing on the solution for variable selection, classification and prediction for longitudinal high-dimensional data with a repeated ordinal outcome. Thus, we propose a two-stage algorithm: first, we modified the Generalized Monotone Incremental Forward Stagewise (GMIFS) algorithm (Hastie et al., 2007) to identify a set of good features whose category-specific mean values are monotonically associated with the ordinal trend in the high-dimensional setting; second, we build a parsimonious random coefficient ordinal response model using the features selected by GMIFS along with other time-dependent covariates for classification and prediction purposes.

The rest of this paper is organized as follows: In Section (2), we start by briefly reviewing the framework of the classical ordinal model (Section (2.1)) and the random coefficient ordinal response model (Section (2.2)). These two models are useful for analyzing the

traditional data in which the response categories are ordered and the number of covariates (p) is smaller than the sample size (n). We then discuss the GMIFS algorithm originally proposed by Hastie et al. (2007) to address the dimension reduction problem for data with a continuous or dichotomous response (Section (2.3)). Our extension of GMIFS for selecting features from data with an ordinal or a longitudinal ordinal response immediately follows in Section (2.4). The statistical methodology to fit a random coefficient ordinal response model and usage of the parameter estimates of the fixed effects and empirical Bayes estimates of the random effects to obtain the fitted ordinal response are discussed in Section (2.5). We apply the proposed method to three simulation datasets and compare the results with existing methods in Sections (3). A description of the motivating time-course microarray data from the Inflammation and the Host Response to Injury study, also known as the Glue Grant data is in Section (4.1). The proposed method is applied to the Glue Grant data and demonstrates superior performance in Sections (4.2) and (4.3). The conclusions and discussion are in Section (5).

2 Methods

2.1 Cumulative Logit Ordinal Model

The ordinal model, which was extended primarily from the logistic and probit regression models, have been actively used to analyze ordinal response data. In particular, the framework derived under the proportional odds assumption (McCullagh, 1980) has served as a bedrock for modern ordinal analysis. To build an ordinal model, we start by letting Y_i be a categorical response for observation i with C categories, where Y_i follows a multinomial distribution with trial size 1 and $P(Y_i = c) = \pi_{ic}$. Let γ_{ic} be a function of probabilities $(\pi_{i1}, \dots, \pi_{iC})$. Suppose a monotone, differentiable link function $g(\cdot)$ connects γ_{ic} to the linear component $\alpha_c + \mathbf{x}_i^T \boldsymbol{\beta}$ such that

$$g(\gamma_{ic}) = \alpha_c + \mathbf{x}_i^T \boldsymbol{\beta}, c = 1, \dots, C - 1 \quad (1)$$

where α_c denotes the category-specific intercept; $\boldsymbol{\beta}$ is a $p \times 1$ vector representing the coefficients associated with explanatory variables $\mathbf{x}_i$. Under the proportional odds assumption, $\boldsymbol{\beta}$ has the same effects for each $g(\gamma_c)$. In other words, the explanatory variable does not have a category-specific effect. In this paper, we will only discuss models under the proportional odds assumption unless stated otherwise. We primarily review the most commonly used ordinal model: the cumulative logit model. Let γ_{ic} measure the probability of response Y_i falling into no greater than the c^{th} category where $\gamma_{ic} = P(Y_i \leq c | \mathbf{x}_i)$. Then the cumulative logit ordinal model can be written as:

$$\log \left(\frac{\gamma_{ic}}{1 - \gamma_{ic}} \right) = \log \left(\frac{P(Y_i \leq c | \mathbf{x}_i)}{P(Y_i > c | \mathbf{x}_i)} \right) = \alpha_c + \mathbf{x}_i^T \boldsymbol{\beta}, c = 1, \dots, C - 1. \quad (2)$$

Correspondingly, we can rewrite γ_{ic} as a function of the linear component and each probability $\pi_{ic}(\mathbf{x}_i)$ can be calculated as the difference between two adjacent γ_{ic} , where

$$\pi_{ic}(\mathbf{x}_i) = \gamma_{ic} - \gamma_{i,c-1} = P(Y_i \leq c | \mathbf{x}_i) - P(Y_i \leq c-1 | \mathbf{x}_i) = \frac{\exp(\alpha_c + \mathbf{x}_i^T \boldsymbol{\beta})}{1 + \exp(\alpha_c + \mathbf{x}_i^T \boldsymbol{\beta})} - \frac{\exp(\alpha_{c-1} + \mathbf{x}_i^T \boldsymbol{\beta})}{1 + \exp(\alpha_{c-1} + \mathbf{x}_i^T \boldsymbol{\beta})}. \quad (3)$$

Since π_{ic} has to be nonnegative, a constraint on the intercepts $-\infty = \alpha_0 < \alpha_1 < \dots < \alpha_{C-1} < \alpha_C = \infty$ should be imposed on equation (2).

2.2 Random Coefficient Model with an Ordinal Response

Statistical methods suitable for modeling clustered or longitudinal data with an ordinal response have become increasingly important in a variety of fields. A large amount of work has been done to model a longitudinal ordinal response through different approaches. For example, Harville and Mee (1984) initiated a mixed model procedure for analyzing clustered data with an ordinal response where the random effects estimates were approximated through a Taylor series expansion. Ezzet and Whitehead (1991) implemented the Newton-Raphson procedure to fit a random-effects model with an ordinal response. Besides the full likelihood approaches, Yang (2001) used marginal quasi-likelihood (MQL) and predictive quasi-likelihood (PQL) to obtain the parameter estimates which are less computationally demanding but could be subject to larger bias. Heagerty and Zeger (1996) developed the estimating equations method with a mixture of parametric approach and semiparametric alternatives to provide computational ease and model robustness. Here, we primarily derive the generalized linear mixed model (GLMM) approach for modeling a longitudinal ordinal response as discussed by Hedeker and Gibbons (1994).

The random coefficient model has its best usage in scenarios when time-dependent repeated measurements are collected where the random effects in the model can capture the subject-specific variations in the data. Under the GLMM framework, the random coefficient ordinal response model is constructed by adding extra subject-specific random effect $\mathbf{u}_i$ to the traditional ordinal model which allows the mean response to vary among individuals. For each subject i , measurements are repeatedly taken at $j = 1, \dots, n_i$ timepoints where the number of repeated measurements can vary from subject to subject. We exemplify the random coefficient ordinal model for the i^{th} subject using the cumulative logit link

$\log\left(\frac{\gamma_{ic}}{1 - \gamma_{ic}}\right)$, where γ_{ic} is a vector of length n_i and $\gamma_{ic} = (\gamma_{i1c}, \dots, \gamma_{ijn_i c}, \dots, \gamma_{in_i c})$;

$\log\left(\frac{\gamma_{ic}}{1 - \gamma_{ic}}\right)$ represents a matrix of dimension $n_i \times C$,

$$\log\left(\frac{\gamma_{ic}}{1 - \gamma_{ic}}\right) = \begin{pmatrix} \log \frac{\gamma_{i11}}{1 - \gamma_{i11}} & \dots & \log \frac{\gamma_{i1c}}{1 - \gamma_{i1c}} & \dots & \log \frac{\gamma_{i1C}}{1 - \gamma_{i1C}} \\ \log \frac{\gamma_{i21}}{1 - \gamma_{i21}} & \dots & \log \frac{\gamma_{i2c}}{1 - \gamma_{i2c}} & \dots & \log \frac{\gamma_{i2C}}{1 - \gamma_{i2C}} \\ \vdots & & \vdots & & \vdots \\ \log \frac{\gamma_{i,n_i 1}}{1 - \gamma_{i,n_i 1}} & \dots & \log \frac{\gamma_{i,n_i c}}{1 - \gamma_{i,n_i c}} & \dots & \log \frac{\gamma_{i,n_i C}}{1 - \gamma_{i,n_i C}} \end{pmatrix}_{n_i \times C} \quad (4)$$

and n_i is the number of repeated measurements for subject i . Each entry in (4) is linked to a linear predictor consisting of the intercept α_c , fixed effects $\boldsymbol{\beta}$, and random effects $\mathbf{u}_i$ which defines a random coefficient ordinal response model for the j^{th} measurement on the i^{th} subject:

$$\log\left(\frac{\gamma_{ijc}}{1-\gamma_{ijc}}\right) = \log\left(\frac{P(Y_{ij} \leq c | \mathbf{x}_{ij}, \mathbf{u}_i)}{P(Y_{ij} > c | \mathbf{x}_{ij}, \mathbf{u}_i)}\right) = \alpha_c + \mathbf{x}_{ij}^T \boldsymbol{\beta} + \mathbf{z}_{ij} \mathbf{u}_i \quad (5)$$

where α_c denotes the category-specific intercept; $\boldsymbol{\beta}$ is a $p \times 1$ vector of coefficients associated with explanatory variables $\mathbf{x}_{ij}$ of dimension $p \times 1$; $\mathbf{z}_{ij}$ is a 1×2 design matrix that includes the intercept and the timepoints when the j^{th} measurement of the i^{th} subject was taken. Correspondingly, $\mathbf{u}_i^T = (u_{1i}, u_{2i})$ is a vector that follows a bivariate normal

distribution with mean $\mathbf{0}$ and variance $\mathbf{G}_i$ and $\mathbf{G}_i = \begin{pmatrix} \sigma_{u_1}^2 & \sigma_{u_1, u_2} \\ \sigma_{u_1, u_2} & \sigma_{u_2}^2 \end{pmatrix}$. Similar to (3), $\pi_c(\mathbf{x}_{ij}, \mathbf{u}_i)$ can be calculated using two adjacent γ_{ijc} , where

$$\begin{aligned} \pi_c(\mathbf{x}_{ij}, \mathbf{u}_i) &= \gamma_{ijc} - \gamma_{ij, c-1} = P(Y_{ij} \leq c | \mathbf{x}_{ij}, \mathbf{u}_i) - P(Y_{ij} \leq c-1 | \mathbf{x}_{ij}, \mathbf{u}_i) \\ &= \frac{\exp(\alpha_c + \mathbf{x}_{ij}^T \boldsymbol{\beta} + \mathbf{z}_{ij} \mathbf{u}_i)}{1 + \exp(\alpha_c + \mathbf{x}_{ij}^T \boldsymbol{\beta} + \mathbf{z}_{ij} \mathbf{u}_i)} - \frac{\exp(\alpha_{c-1} + \mathbf{x}_{ij}^T \boldsymbol{\beta} + \mathbf{z}_{ij} \mathbf{u}_i)}{1 + \exp(\alpha_{c-1} + \mathbf{x}_{ij}^T \boldsymbol{\beta} + \mathbf{z}_{ij} \mathbf{u}_i)}. \end{aligned} \quad (6)$$

2.3 Generalized Monotone Incremental Forward Stagewise Regression

Generalized Monotone Incremental Forward Stagewise (GMIFS) regression is a slow learning procedure that provides a greedy approximation to the proposed function. GMIFS works by updating one coefficient β_j at a time using a small incremental amount to yield a penalized solution. As a typical GMIFS consists of hundreds of thousands of tiny little steps before converging, it was once considered as an inefficient algorithm and neglected for a long time. This method started gaining enormous attention when Hastie et al. (2007) discovered the learning method Boosting (Schapire et al., 1998) is essentially the combination of a sequence of adaptively constructed basis functions and the GMIFS procedure. The other interesting property of GMIFS discovered by Efron et al. (2004) is GMIFS can approximate the solution of an L_1 -penalized regression (LASSO) where the regularization path looks the same under a certain scenario. For data with a continuous response, GMIFS works by iteratively updating the selected coefficient β_j until the residual r is uncorrelated with all covariates $\mathbf{x}$ and thus achieves the minimal residual of sum squares (RSS) criteria. For data with a dichotomous response, GMIFS adopts the steepest descent optimization method by iteratively updating the coefficient β_j associated with the largest negative gradient to achieve the local minimum of the proposed function (Hastie et al., 2007). One common criticism of this greedy algorithm is that it may fail to produce the globally optimal solution since it neither searches exhaustively on all possible paths nor adjusts for previous paths. Because of this limited vision problem, the greedy algorithm is often implemented to perform variable selection rather than merely optimization.

In order to model high-dimensional data with an ordinal response, we modified and extended the GMIFS to an ordinal response under the framework of the cumulative logit ordinal model. Similar to data with a dichotomous response, the first-order derivative of the negative log-likelihood function $-\log L(\boldsymbol{\alpha}, \boldsymbol{\beta}; \mathbf{x})$ with respect to $\boldsymbol{\beta}$ is calculated to determine which coefficient β_j is to be updated. In each iteration, to determine the direction of β_j is computationally expensive as the calculation of the second-order derivative of the likelihood

is required. Hastie et al. (2007) showed by using the expanded representation $\tilde{\mathbf{X}} = (\mathbf{X}, -\mathbf{X})$ one can avoid this cumbersome computation which results in an efficient version of GMIFS. In the expanded representation, the coefficients are expanded to $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p, \beta_{p+1}, \dots, \beta_{2p})$ and the Karush-Kuhn-Tucker condition ensures β_j and β_{j+p} associated with the same covariate x_j cannot be updated simultaneously. To further accelerate the optimization

process, the estimate of intercept α_c remains the initial estimate $\hat{\alpha}_c = \log \frac{P(Y_i \leq c)}{1 - P(Y_i \leq c)}$ through the iterative procedure.

We now derive the gradient of the negative log-likelihood function in a cumulative logit ordinal model by starting with the general form

$$-\log L(\boldsymbol{\alpha}, \boldsymbol{\beta}; \mathbf{x}) = -\sum_{i=1}^n \log L_i(\boldsymbol{\alpha}, \boldsymbol{\beta}; \mathbf{x}_i) = -\sum_{i=1}^n \log \left(\prod_{c=1}^C \pi_c(\mathbf{x}_i)^{y_{ci}} \right) = -\sum_{i=1}^n \sum_{c=1}^C y_{ci} \log \pi_c(\mathbf{x}_i). \quad (7)$$

For a given β_j , the first-order derivative of $-\log L(\boldsymbol{\alpha}, \boldsymbol{\beta}; \mathbf{x})$ can be calculated as:

$$-\frac{\partial \log L(\boldsymbol{\alpha}, \boldsymbol{\beta}; \mathbf{x})}{\partial \beta_j} = -\sum_{i=1}^n \left(\sum_{c=1}^C y_{ci} \cdot \frac{\frac{\partial \pi_c(\mathbf{x}_i)}{\partial \beta_j}}{\pi_c(\mathbf{x}_i)} \right). \quad (8)$$

For the cumulative logit ordinal model, the probability for each category can be calculated using equation (3) and the kernel in equation (8) can be written as

$$\frac{\frac{\partial \pi_c(\mathbf{x}_i)}{\partial \beta_j}}{\pi_c(\mathbf{x}_i)} = x_{ij} \cdot (1 - P(Y_i \leq c-1 | \mathbf{x}_i) - P(Y_i \leq \mathbf{x}_i)) = x_{ij} \cdot \left(1 - \frac{\exp(\alpha_{c-1} + \mathbf{x}_i^T \boldsymbol{\beta})}{1 + \exp(\alpha_{c-1} + \mathbf{x}_i^T \boldsymbol{\beta})} - \frac{\exp(\alpha_c + \mathbf{x}_i^T \boldsymbol{\beta})}{1 + \exp(\alpha_{c-1} + \mathbf{x}_i^T \boldsymbol{\beta})} \right). \quad (9)$$

By plugging the exact form of $\frac{\partial \pi_c(\mathbf{x}_i)}{\partial \beta_j}$ into equation (8), the partial derivative of the negative log-likelihood with respect to β_j can be written as

$$\begin{aligned} -\frac{\partial \log L(\boldsymbol{\alpha}, \boldsymbol{\beta}; \mathbf{x})}{\partial \beta_j} &= \sum_{i=1}^n \left(\frac{y_{1i} \cdot \frac{\partial \pi_1(\mathbf{x}_i)}{\partial \beta_j}}{\pi_1(\mathbf{x}_i)} + \sum_{c=2}^{C-1} \frac{y_{ci} \cdot \frac{\partial \pi_c(\mathbf{x}_i)}{\partial \beta_j}}{\pi_c(\mathbf{x}_i)} + \frac{y_{Ci} \cdot \frac{\partial \pi_C(\mathbf{x}_i)}{\partial \beta_j}}{\pi_C(\mathbf{x}_i)} \right) \\ &= \sum_{i=1}^n x_{ij} \cdot \left(y_{1i} \frac{1}{1 + \exp(\alpha_1 + \mathbf{x}_i^T \boldsymbol{\beta})} + \sum_{c=2}^{C-1} y_{ci} \left(1 - \frac{\exp(\alpha_{c-1} + \mathbf{x}_i^T \boldsymbol{\beta})}{1 + \exp(\alpha_{c-1} + \mathbf{x}_i^T \boldsymbol{\beta})} \right) \right. \\ &\quad \left. - \frac{\exp(\alpha_c + \mathbf{x}_i^T \boldsymbol{\beta})}{1 + \exp(\alpha_{c-1} + \mathbf{x}_i^T \boldsymbol{\beta})} \right) - y_{Ci} \frac{\exp(\alpha_{C-1} + \mathbf{x}_i^T \boldsymbol{\beta})}{1 + \exp(\alpha_{C-1} + \mathbf{x}_i^T \boldsymbol{\beta})}. \end{aligned} \quad (10)$$

Thus, $-\frac{\partial \log L(\boldsymbol{\alpha}, \boldsymbol{\beta}; \mathbf{x})}{\partial \boldsymbol{\beta}} = \left(-\frac{\partial \log L(\boldsymbol{\alpha}, \boldsymbol{\beta}; \mathbf{x})}{\partial \beta_1}, \dots, -\frac{\partial \log L(\boldsymbol{\alpha}, \boldsymbol{\beta}; \mathbf{x})}{\partial \beta_{2p}} \right)$ is a vector of length $2p$ with each element having a form of (10). At each iteration, the β_j associated with the largest negative gradient is updated with a small incremental amount ε where $\beta_j^{(s+1)} \leftarrow \beta_j^{(s)} + \varepsilon$. When the iterative procedure converges according to certain criteria (see Algorithm 1), a sequence of penalized cumulative logit ordinal models are fitted at steps immediately

preceding the step where a new feature enters the active set using the penalized estimate of the features. The intercepts α_c are obtained by optimizing the penalized likelihood function, which only depends on the unknown parameter α .

2.4 GMIFS for Random Coefficient Ordinal Response Model

To the best of our knowledge, there has not been methodology focused on variable selection, classification and prediction using longitudinal high-dimensional data with repeated ordinal outcomes. We therefore propose our method by further extending the GMIFS for variable selection and utilizing the classic random coefficient ordinal response model for classification and prediction. Note the proposed method should not be restricted to the application in high-dimensional longitudinal data, it can also be applied to the traditional longitudinal or clustered data with an ordinal outcome.

The transformation for GMIFS from traditional to longitudinal data with ordinal response is rather smooth as we will show it does not require estimation of the variance component in each iteration. For a random coefficient ordinal response model (5), the log-likelihood $\log L_i(\alpha, \beta, \mathbf{u}_i, \mathbf{G}_i; \mathbf{x}_i, \mathbf{z}_i)$ can be written as:

$$\log L_i(\alpha, \beta, \mathbf{u}_i, \mathbf{G}_i; \mathbf{x}_i, \mathbf{z}_i) = \log g(\mathbf{u}_i, \mathbf{G}_i) + \log \left(\prod_{j=1}^{n_i} \prod_{c=1}^C \pi_c(\mathbf{x}_{ij}, \mathbf{z}_i, \mathbf{u}_i)^{y_{ijc}} \right) \quad (11)$$

where $g(\mathbf{u}_i, \mathbf{G}_i) = \frac{1}{2\pi\sigma_{u1}\sigma_{u2}\sqrt{1-\rho^2}} \exp \left(-\frac{1}{2(1-\rho^2)} \left(\frac{u_{1i}^2}{\sigma_{u1}^2} + \frac{u_{2i}^2}{\sigma_{u2}^2} - \frac{2\rho u_{1i}u_{2i}}{\sigma_{u1}\sigma_{u2}} \right) \right)$ is the distribution of the random effects $\mathbf{u}_i$. As $g(\mathbf{u}_i, \mathbf{G}_i)$ is not a function of β_j , the partial derivative of negative log-likelihood with respect to β_j only depends on the second term on the right side of equation (11). The empirical Bayes estimates of random effects $\mathbf{u}_i$ and initial estimate of intercept α_c can be obtained from the null random coefficient model. Therefore, the gradient of the negative log-likelihood of the random coefficient ordinal response model is similar to that of the traditional ordinal model (10), which can be written as:

$$\begin{aligned} & - \frac{\partial \log L(\alpha, \beta, \mathbf{u}_i, \mathbf{G}_i; \mathbf{x}_i, \mathbf{z}_i)}{\partial \beta_j} \\ &= \sum_{i=1}^n \left(\frac{y_{1i} \cdot \frac{\partial \pi_1(\mathbf{x}_i, \mathbf{z}_i, \mathbf{u}_i)}{\partial \beta_j}}{\pi_1(\mathbf{x}_i, \mathbf{z}_i, \mathbf{u}_i)} + \sum_{c=2}^{C-1} \frac{y_{ci} \cdot \frac{\partial \pi_c(\mathbf{x}_i, \mathbf{z}_i, \mathbf{u}_i)}{\partial \beta_j}}{\pi_c(\mathbf{x}_i, \mathbf{z}_i, \mathbf{u}_i)} + \frac{y_{Ci} \cdot \frac{\partial \pi_C(\mathbf{x}_i, \mathbf{z}_i, \mathbf{u}_i)}{\partial \beta_j}}{\pi_C(\mathbf{x}_i, \mathbf{z}_i, \mathbf{u}_i)} \right) \\ &= \sum_{i=1}^n x_{ij} \cdot \left(y_{1i} \frac{1}{1 + \exp(\alpha_1 + \mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{z}_i \mathbf{u}_i)} + \sum_{c=2}^{C-1} y_{ci} \left(1 - \frac{\exp(\alpha_{c-1} + \mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{z}_i \mathbf{u}_i)}{1 + \exp(\alpha_{c-1} + \mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{z}_i \mathbf{u}_i)} \right) \right. \\ & \quad \left. - \frac{\exp(\alpha_c + \mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{z}_i \mathbf{u}_i)}{1 + \exp(\alpha_{c-1} + \mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{z}_i \mathbf{u}_i)} \right) - y_{Ci} \frac{\exp(\alpha_{C-1} + \mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{z}_i \mathbf{u}_i)}{1 + \exp(\alpha_{C-1} + \mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{z}_i \mathbf{u}_i)}. \end{aligned} \quad (12)$$

At each iteration, β_j associated with the largest negative gradient is updated with a small incremental amount, where $\beta_j^{(s+1)} \leftarrow \beta_j^{(s)} + \varepsilon$. A sequence of random coefficient ordinal response models is then fit to build a parsimonious model for prediction and classification,

which we will discuss in next section. We summarize GMIFS for longitudinal ordinal response data in Algorithm 1.

2.5 Fitting a Random Coefficient Ordinal Response Model

Fitting a random coefficient ordinal response model is nontrivial and challenging due to nonlinearities between fixed effects, random effects and the response. Here, we focus on fitting the model using the marginal likelihood approach discussed by Hedeker and Gibbons (1994), where the closed-form of the marginal likelihood can be approximated using the adaptive Gauss-Hermite quadrature method. For a random coefficient ordinal response model (5), the likelihood function for subject i can be written as:

$$L_i(\alpha, \beta, \mathbf{G}_i, \mathbf{u}_i | \mathbf{x}_i, \mathbf{y}_i) = f_{\mathbf{x}_i, \mathbf{y}_i | \mathbf{u}_i}(\mathbf{x}_i, \mathbf{y}_i | \mathbf{u}_i, \beta, \mathbf{G}_i) f_{\mathbf{u}_i}(\mathbf{u}_i | \mathbf{G}_i) \quad (13)$$

Algorithm 1

- 1 Create a negative version $-x_j$ of each predictor x_j and expand the predictor space to $\tilde{\mathbf{X}} = (\mathbf{X}, -\mathbf{X})$. Set the initial values at step $s = 0$ for the coefficients $\beta^{(0)} = (\beta_1, \dots, \beta_{2p}) = \mathbf{0}$. The initial estimates of the intercept α_c and random effects $\mathbf{u}_i$ are obtained from the null random coefficient model.
- 2 Find the predictor x_j , $j = 1, \dots, 2p$ with the largest negative gradient of the log-likelihood $\log L(\alpha, \beta, \mathbf{u}_i, \mathbf{G}_i; \mathbf{x}_i, \mathbf{z}_i)$ evaluated at the current estimate $\beta^{(s)}$.

$$-\frac{\log L(\alpha, \beta, \mathbf{u}_i, \mathbf{G}_i; \mathbf{x}_i, \mathbf{z}_i)}{\beta_j}$$
- 3 Update the coefficient estimate of the selected predictor x_j in step 2 with $\beta_j^{(s+1)} \leftarrow \beta_j^{(s)} + \varepsilon$, where ε is a small positive amount; a rational choice is $\varepsilon = 1 \times 10^{-4}$.
- 4 Repeat steps 2 and 3 many times until it meets either criteria: 1) the difference between two successive log-likelihood is smaller than a given value δ ; 2) the number of features having a nonzero coefficient estimates is less than a specified value.
- 5 Fit a sequence of random coefficient ordinal response models using features selected from GMIFS at steps immediately preceding the step where a new feature enters the active set. The parsimonious model is based on model fitting criteria, e.g. AIC, BIC.

where $f_{\mathbf{x}_i, \mathbf{y}_i | \mathbf{u}_i}(\cdot) = \prod_{j=1}^{n_i} \prod_{c=1}^C \pi_c(\mathbf{u}_i, \mathbf{x}_{ij})^{y_{ijc}}$ denotes the probability mass function of a multinomial distribution, where $\pi_c(\mathbf{u}_i, \mathbf{x}_{ij})$ represents the probability observation i at timepoint j falls into the c^{th} category. For the cumulative logit model, $\pi_c(\mathbf{u}_i, \mathbf{x}_{ij})$ can be written as $\pi_c(\mathbf{u}_i, \mathbf{x}_{ij}) = P(Y_{ij} \leq c | \mathbf{x}_{ij}, \mathbf{u}_i) - P(Y_{ij} \leq c-1 | \mathbf{x}_{ij}, \mathbf{u}_i)$. Here, $f_{\mathbf{u}_i}(\mathbf{u}_i | \mathbf{G}_i)$ is a probability density function of the bivariate normal distribution. According to its definition, the marginal likelihood (or integrated likelihood) $L_i^*(\alpha, \beta, \mathbf{G}_i | \mathbf{x}_i, \mathbf{y}_i)$ is constructed by marginalizing the random effects $\mathbf{u}_i$:

$$L_i^*(\alpha, \beta, \mathbf{G}_i | \mathbf{x}_i, \mathbf{y}_i) = \int \int L_i(\alpha, \beta, \mathbf{G}_i, \mathbf{u}_i | \mathbf{x}_i, \mathbf{y}_i) du_{1i} du_{2i}. \quad (14)$$

Pinheiro and Bates (1995) incorporated the importance sampling idea in Gauss-Hermite quadrature and introduced the adaptive Gauss-Hermite quadrature procedure where the abscissas z_i^* and weights w_i were constructed according to both the Hermite polynomial (Abramowitz et al., 1972) as well as the observed samples. To approximate (14), we denote

$$u_{1i} = \hat{u}_{1i} + \sqrt{2} |f''_{1i}(\cdot, u_{1i})|^{-1/2} z_i^*$$

$$u_{2i} = \hat{u}_{2i} + \sqrt{2} |f''_{2i}(\cdot, u_{2i})|^{-1/2} z_i^*$$

and it then follows

$$du_1 = \sqrt{2} |f''_{1i}(\cdot, u_{1i})|^{-1/2} dz_i^*$$

$$du_2 = \sqrt{2} |f''_{2i}(\cdot, u_{2i})|^{-1/2} dz_i^*.$$

Let

$$g(\cdot, \mathbf{u}_i) = -\frac{u_{1i}^2}{2\sigma_{u_1}^2} - \frac{u_{2i}^2}{2\sigma_{u_2}^2} + \log \prod_{j=1}^{n_j} \prod_{c=1}^C \pi_c^{y_{ijc}}(\cdot, \mathbf{u}_i)$$

$$f(\cdot, \mathbf{u}_i) = -\log(\exp(g(\cdot, \mathbf{u}_i)))$$

The empirical Bayes estimates of u_{1i} and u_{2i} can be obtained from

$$\hat{\mathbf{u}}_i = \text{argmin} -\log(\exp(g(\cdot, \mathbf{u}_i))),$$

where $f''_{1i}(\cdot, u_{1i})$ is the second derivative of $f(\cdot, \mathbf{u}_i)$ with respect to u_{1i} and $f''_{2i}(\cdot, u_{2i})$ is the second derivative of $f(\cdot, \mathbf{u}_i)$ with respect to u_{2i} . $f''_{1i}(\cdot, u_{1i})$ and $f''_{2i}(\cdot, u_{2i})$ are the diagonal elements of the Hessian matrix of function $f(\cdot, \mathbf{u}_i)$. In the context of GLMM, the exact form of the second-order derivative of function $f(\cdot, u_{1i}, u_{2i})$ can be quite complex, an approximate form of the Hessian matrix is often applied in the optimization process. The marginal likelihood can be approximated as:

$$\begin{aligned}
L_i^*(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{G}_i | \mathbf{x}_i, \mathbf{y}_i) &\simeq \int \int \exp \left(\log \prod_{j=1}^{n_j} \prod_{c=1}^C (\pi_c^{y_{ijc}} (\hat{u}_1 + \sqrt{2} |f_{1i}''(\cdot, u_{1i})|^{-1/2} z_i^*, u_{2i}) \right. \\
&\quad \left. - h(\hat{u}_1 + \sqrt{2} |f_{1i}''(\cdot, u_{1i})|^{-1/2} z_i^*, u_{2i}) \right) dz_i^* du_{2i} \\
&= \int \sum_i^{N_{GQ}} \exp \left(\log \prod_{j=1}^{n_j} \prod_{c=1}^C \pi_c^{y_{ijc}} (\hat{u}_1 + \sqrt{2} |f_{1i}''(\cdot, u_{1i})|^{-1/2} z_i^*, u_{2i}) \right. \\
&\quad \left. - h(\hat{u}_1 + \sqrt{2} |f_{1i}''(\cdot, u_{1i})|^{-1/2} z_i^*, u_{2i}) \right) \mathbf{W}_i du_{2i} \\
&= \sum_i^{N_{GQ}} \left(\sum_i^{N_{GQ}} \exp \left(\log \prod_{j=1}^{n_j} \prod_{c=1}^C \pi_c^{y_{ijc}} (\hat{u}_1 + \sqrt{2} |f_{1i}''(\cdot, u_{1i})|^{-1/2} z_i^*, \right. \right. \\
&\quad \left. \left. \hat{u}_2 + \sqrt{2} |f_{2i}''(\cdot, u_{2i})|^{-1/2} z_i^* \right) - h(\hat{u}_1 + \sqrt{2} |f_{1i}''(\cdot, u_{1i})|^{-1/2} z_i^*, \right. \\
&\quad \left. \left. \hat{u}_2 + \sqrt{2} |f_{2i}''(\cdot, u_{2i})|^{-1/2} z_i^* \right) \right) \mathbf{W}_i \mathbf{W}_i
\end{aligned} \tag{15}$$

where $\mathbf{W}_i = \exp(\|\mathbf{z}_i^*\|) \prod_i w_i$ and $h(u_{1i}, u_{2i}) = \frac{1}{2(1-\rho^2)} \left(\frac{u_{1i}^2}{\sigma_{u1}^2} + \frac{u_{2i}^2}{\sigma_{u2}^2} - \frac{2\rho u_{1i}u_{2i}}{\sigma_{u1}\sigma_{u2}} \right)$. Correspondingly, the log-likelihood for all subjects can be written as

$$\begin{aligned}
\log L^*(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{G}_i | \mathbf{x}_i) &= -N \log \pi - \frac{N}{2} \log(\sigma_{u1}^2 \sigma_{u2}^2 (1-\rho^2)) - \frac{1}{2} \sum_i^N \log |f_{1i}''(\cdot, u_{1i})| - \\
&\quad \frac{1}{2} \sum_i^N \log |f_{2i}''(\cdot, u_{2i})| + \sum_i^N \log L_i^*(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{G}_i | \mathbf{x}_i, \mathbf{y}_i).
\end{aligned} \tag{16}$$

To evaluate the predictability of the random coefficient ordinal response model, the predicted ordinal membership is calculated for each observation. The fixed effects $\boldsymbol{\alpha}, \boldsymbol{\beta}$ and variance $\mathbf{G}_i$ can be obtained by maximizing the marginal log-likelihood (16) using general-purpose optimization methods, *e.g.*, Newton-Raphson, quasi-Newton or Fisher's scoring. The random effects estimates $\mathbf{u}_i$ can be obtained by empirical Bayes (EB) method (Efron and Morris, 1973). As discussed by Hedeker and Gibbons (2006), the EB estimate of individual random effects (also known as EAP, 'Expected a Posterior') is derived as the mean of the posterior distribution of $\mathbf{u}_i$ given the data, which is proportional to the likelihood with estimated $\boldsymbol{\alpha}, \boldsymbol{\beta}$ and $\mathbf{G}_i$. An important property of the EB estimator is it is said to be shrunken to the population mean compared to the ordinary least squares (OLS) estimator. The amount of shrinkage depends on the number of repeated measurements within each subject. By borrowing the information from the data, the EB estimate approaches the mean of the prior distribution when the number of measurements is small, while it is very similar to OLS estimate when the number of measurements is large. After obtaining the EB estimate $\hat{\mathbf{u}}_i$, for each category c , the predicted ordinal value of observation i at time j can be obtained by plugging all estimates into equation (6) and denoted as $\hat{\pi}_{c,ij}$. The category associated with the largest probability $\hat{\pi}_{c,ij}$ is the predicted category for observation i at time j .

3 Simulation Results

3.1 Simulation Study 1

The first simulation aims to validate the assumption that by tentatively ignoring the within-subject correlation in the dimension reduction process, GMIFS can still accurately detect the active set that contains the truly important feature. Since it is difficult to simulate a high-dimensional data with a time-dependent ordinal outcome, we modified a small existing time-dependent microarray dataset called **tccl.10** collected from a study to investigate the response of human T cells to PMA and ionomycin treatment (available in R package *longitudinal* (Opgen-Rhein and Strimmer, 2014)). The dataset contains samples from 10 patients collected at 10 unequally spaced time points, where expression levels of 58 genes were obtained for each patient per time point. We simulated two scenarios: 1 or 2 out of 58 genes were significantly associated with the ordinal outcome and the corresponding $\beta_j \sim \text{Uniform}(-5, 5)$. The ordinal outcome was generated using a latent variable approach. We define the latent response $\mathbf{y}^*$ to be $\mathbf{y}^* = \mathbf{x}^T \boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon} \sim N(0, 1)$. The ordinal outcome $\mathbf{y}$ with three ordered categories was created by using the sample quantile, where

$$y_i = \begin{cases} 1 & \text{if } \text{Prob}(y_i^*) \leq 1/3 \\ 2 & \text{if } \text{Prob}(y_i^*) \leq 2/3 \\ 3 & \text{if } \text{Prob}(y_i^*) > 2/3. \end{cases}$$

For each scenario, we ran the simulation 1000 times. The accuracy was high for both scenarios with 87.7% and 97.8% of the time GMIFS was able to select the active set that contained the predefined important gene(s), separately.

3.2 Simulation Study 2

The second simulation aims to compare the proposed method to the existing method that is capable of classifying clustered high-dimensional data. As aforementioned, there is no other method designed to classify high-dimensional data with a clustered ordinal outcome. Therefore, we simulated high-dimensional clustered data with a binary outcome, which should be comparable to simulation results for cumulative logit ordinal response models. We compared the performances of: GMIFS logistic model with random effects, LASSO logistic model with random effects, glmm-Lasso and LASSO alone. Efron et al. (2004) discussed the striking similarity between Forward Stagewise and LASSO in linear regression and it is of great interest to explore the behaviors of these two methods for data with a discrete response, in particular, clustered discrete response. For the GMIFS and LASSO logistic models with random effects, after the variable selection, a sequence of random effects logistic regression models was fit using R package *lme4* (Bates et al., 2012). The optimal model was selected according to the ‘elbow criterion’ in BIC. For LASSO alone, a sequence of logistic models was fit and the parsimonious model was selected according to the ‘elbow criterion’ in BIC. In addition, for the glmmLasso algorithm proposed by Schelldorfer et al. (2013), we used the R package *glmmixedlasso* (Schelldorfer, 2014) and the tuning parameter was selected according to the minimum BIC.

3.2.1 Balanced Design—We modified the dataset from Schellldorfer et al. (2013) and simulated dataset containing $N = 40$ subjects and $p = 1000$ features, where each subject was repeatedly measured 10 times. Among the $p = 1000$ features, we assumed the first four features $x_1, \dots, x_4$ were important and set their corresponding coefficients to $\beta^T = (1, -1, 1, -1, 0, \dots, 0)$. Each feature x_i was randomly generated from a normal distribution with mean 0 and standard deviation 1. Suppose the intercept and x_1 have a subject-specific structure where the corresponding random effect u_1 and u_2 are generated from normal distribution $N(0, 2)$ and $N(0, 1)$, respectively. The linear predictor can be constructed by $\eta = \mathbf{x}^T \beta + \mathbf{Z} \mathbf{b}_i$ and the probability that observation i at measurement j falls into a given category is

$\pi_{ij} = \frac{\exp(\eta_{ij})}{1 + \exp(\eta_{ij})}$. Correspondingly, the binary response y_{ij} was simulated from a binomial distribution $B(1, \pi_{ij})$.

3.2.2 Unbalanced Design—We simulated a dataset containing $N = 40$ subject and $p = 1000$ features, where the number of times subject i was measured was randomly generated as either 8, 10, or 12 times. The other settings were exactly the same as in the balanced design.

3.3 Results

Table 1 compares the parameter estimates and model fitting results when applying the GMIFS logistic model with random effects, LASSO logistic model with random effects, glmmLasso, and LASSO alone to the two simulated datasets. In the balanced design, all methods are capable of detecting features associated with nonzero coefficients. GMIFS logistic model with random effects and LASSO logistic model with random effects result in identical results with the smallest squared error $\|\hat{\beta} - \beta\|^2$. The glmmLasso returns the minimum prediction error. The LASSO alone has seemingly reasonable parameter estimates but the prediction error is high, which highlights the importance of including random effects for capturing the within-subject correlation as it enhances model predictability. In the unbalanced design, GMIFS logistic model with random effects and LASSO logistic model with random effects have slightly different results in including the noise feature while the parameter estimates stand out in comparison with the other methods. In both scenarios, our proposed GMIFS has competitive results to LASSO while LASSO is extremely efficient in computational time.

4 Application

4.1 Description of the Glue Grant Data

The Inflammation and the Host Response to Injury is a large-scale collaborative research program supported by the National Institute of General Medical Sciences which began in 1998. It aims to better understand the human body's response to serious injury using a discovery-driven approach. One goal was to identify gene sets having high predictability of multiple organ failure. We demonstrate the application of our proposed model using 869 buffy coat samples collected from burn injury patients and hybridized to Affymetrix Human Genome U133 Plus 2.0 Arrays with each sample having 54675 probe sets. The samples were normalized and summarized using the dChip method (Li and Wong, 2001). After

removing observations with missing outcomes, 657 samples from 169 burn injury patients were used in the analysis. That is, on average, each patient has approximate 4 repeated measurements with the severity of illness possibly fluctuating over time. We further reduced the dimensionality by filtering probe sets that were absent on all arrays, leaving 48093 probe sets for statistical analysis. The severity of illness of the burn injury patients was assessed using the Marshall Multiple Organ Dysfunction Score (Marshall, 1995), which is considered to be a comprehensive and effective measurement system for critical illness condition and has been demonstrated to be strongly associated with risk of ICU and hospital mortality. The Marshall Score evaluates the dysfunction level of six systems: 1) the respiratory system (Po2/FIO2 ratio); 2) the renal system (serum creatinine concentration); 3) the hepatic system (serum bilirubin concentration); 4) the hematologic system (platelet count); 5) the central nervous system (Glasgow Coma Scale) and 6) the cardiovascular system (pressure-adjusted heart rate). The assessment of each organ results in an ordinal outcome ranging from 0 to 4 with 0 indicating normal and 4 indicating severe dysfunction. We applied the proposed model to classify the severity of illness according to all six Marshall score measurements. The results from Marshall score assessed on the renal system and central nervous system will be mainly discussed. The original time covariate in the data was recorded on the hour scale which we converted and standardized to a day scale for convenience purposes. Figure 1 illustrates the distribution of sample collection times where the majority of samples were collected at four scheduled time points: baseline, day 4, day 7 and day 14. However, a substantial number of samples were also collected at unscheduled times which induces a complex correlation structure among observations. The performance of the proposed model was evaluated using prediction error. The consistency of variable selection was evaluated using cross-validation. There are few publications that discuss cross-validation for longitudinal high-dimensional data. We conducted a quasi K -fold cross-validation by dividing the subjects, instead of observations into K partitions where the $K - 1$ partitions formed a training set and the remaining 1 part formed a test set. Since the number of repeated measurements per subject varied, the number of observations in the test set also varied for each fold.

All analyses were performed using R 3.1.0 on a Linux Beowulf cluster at Department of Biostatistics, Virginia Commonwealth University.

4.2 Results: Marshall score on renal system

According to Ibrahim et al. (2013), acute kidney injury (AKI) is known to be a major complication leading to mortality in burn injury patients. However, the treatment for this condition is still not well defined. Therefore, early diagnosis and prevention are of ultimate importance for preventing aggressive progression that incurs irreversible tissue damage. In the burn injury data, the original Marshall score assessed on the renal system has five ordered categories representing: 0(normal), 1(mild), 2(moderate), 3(markedly) and 4(severe) illness. We combined the moderate, markedly and severe illness groups to create a modified ordinal outcome with three ordered categories and summarized the overall frequencies by level (Table 2 and Figure 2 top). We applied the GMIFS with increment amount $\varepsilon = 1 \times 10^{-4}$, convergence criteria $\delta = 1 \times 10^{-4}$, and the maximal proportion of important features to be detected 0.0030 (which corresponds to including a maximum of 144 probe sets of the

48093). Upon convergence, 144 probe sets with nonzero coefficients entered the active set. We then fit a sequence of random intercept ordinal response models at steps immediately preceding the step where a new feature entered the active set using R package `ordinal` (Christensen, 2013). The best parsimonious model associated with the minimal BIC was at step 6865 which included 6 probe sets having a non-zero coefficient estimate in the active set. By fitting a random intercept and a random coefficient ordinal response model with 6 selected features and time, we calculated the predicted ordinal response. By comparing the predicted and observed ordinal response, high classification accuracies: 86.8% for the random intercept ordinal response model (Table 3) and 93.0% for the random coefficient ordinal response model (Table 4) were obtained, respectively. In this case, the random coefficient model had a better performance, which could be due to reasons the selected features enhance predicability as they are essentially monotonically changing with time. However, the computational time required for fitting a random coefficient model can be much longer compared with random intercept model (total of approximately 68 hours), though we are working on optimizing our code and reduce run time using alternative code as well.

Figure 3 illustrates the feature selection and model fitting process. The 6 probe sets in active set were: 203234_at, 203932_at, 214090_s_at, 216336_x_at, 224414_s_at and 235568_at. By matching the Affymetrix HG-U133A2 array annotation to gene symbol identifier, three genes: UPP1, HLA-DMB and DDAH2 were identified and found to be monotonically associated with the ordinal response. DDAH2 is known to be associated with hypertension, which is a leading cause of kidney disease. A large amount of research has concentrated on understanding the biological mechanism behind hypertension. For example, Pullamsetti et al. (2005) investigated the role of the metabolizing enzyme DDAH in the course of Idiopathic pulmonary arterial hypertension (IPAH). Two isoforms of DDAH (DDAH1 and DDAH2) have been found in mammals. When comparing the tissue samples from healthy donors and IPAH patients, a significant reduction of DDAH2 immunoreactivity was observed while no significant difference in DDAH1 immunostaining intensity, which demonstrated the change in expression of DDAH in IPAH lungs. UPP1 is known to be predominantly expressed in certain types of cancer, *e.g.*, pancreatic cancer (Sahin et al., 2005) but its relation to kidney failure is yet to be discovered. The consistency of variable selection is evaluated using a 10-fold cross-validation. All three genes were accurately identified in the cross-validation process (Supplementary Table S1).

4.3 Results: Marshall score on central nervous system

Mental health and quality of life of severe burn injury patients are also of great concern. There is growing evidence that psychological health problems during the acute care setting have long-term consequences and influence outcome of burn injury (Renneberg et al., 2013). Post-traumatic stress disorder (PTSD) is a common mental disorder that has been seen in up to 43% of burn injury patients 1 or more years after hospitalization (McKibben et al., 2008). This aftermath psychological depression, if not identified and treated properly at early stages, could markedly change the survivors' quality of life. In this burn injury data, the original Marshall score assessed on the central nervous system used five ordered categories. We aggregated it to a more balanced three-category ordinal scale and summarized the overall frequency by level (Table 5 and Figure 2 bottom). We applied the GMIFS with

increment amount $\varepsilon = 1 \times 10^{-4}$, convergence criteria $\delta = 1 \times 10^{-4}$ and the maximal proportion of important features to be detected 0.0030 (which corresponds to including a maximum of 144 probe sets of the 48093). Upon convergence, 144 probe sets with nonzero coefficients entered the active set. We then fit a sequence of random intercept ordinal response models at steps immediately preceding the step where a new feature entered the active set using R package *ordinal* (Christensen, 2013). The best parsimonious model associated with the minimal BIC was at step 23830 with 60 probe sets having a non-zero coefficient estimate. By fitting a random intercept and a random coefficient ordinal response model with 60 selected features and time, we calculated the predicted ordinal response. By comparing the predicted and observed ordinal response, high classification accuracies: 83.4% for the random intercept ordinal response model (Table 6) and 72.9% for the random coefficient ordinal response model (Table 7) were obtained, respectively. In this case, conversely, the random intercept model has a better performance, which reveals the difficulty in accurately estimating the variances in the presence of a moderate number of features with small effect in a random coefficient model. The total computational time is approximately 83 hours.

Figure 4 illustrates the feature selection and model fitting process. By matching the Affymetrix HG-U133A2 array annotation to gene symbol identifier, 36 genes were identified. A full list and detailed information of these genes can be found in Supplementary Table S2. Among them, several genes have been reported to be associated with mental disorders and psychosocial impairment, such as Alzheimer's disease, schizophrenia, and neuritis. For example, RGS10 protein is known to have a predominant location in the cytosol (Rivero et al., 2010) and its movement between the cytoplasm and the nucleus is considered as a possible mechanism of regulating intra-cellular signaling (Burgon et al., 2001). Rivero et al. (2013) designed an experiment to evaluate the treatment effect of an antipsychotic or antidepressant drugs on schizophrenic, non-diagnosed suicide, and control groups where cytosolic RGS10 protein immunoreactivity has been involved. Another gene that has been previously linked to mental disorders is LRP8, which may be related to two receptors in reelin signaling pathway (Fatemi, 2001). The reelin signaling is involved in the etiology of neurodevelopmental and psychiatric disorders such as: schizophrenia, bipolar, depression and autism.

5 Discussion and Conclusions

In this paper, we extended the novel statistical learning algorithm GMIFS for feature selection in high-dimensional data with a longitudinal ordinal response. The classical GLMM is then used to obtain the predicted response and estimate the prediction error. Computationally, we demonstrate good performance of the proposed method in terms of consistency in feature selection and accuracy in prediction using three simulated datasets and one time-course microarray gene expression dataset. Similar to most penalization models, this Boosting-type method follows the 'Bet on Sparsity' principle for high-dimensional problems, where it has superior performance when the degree of sparseness in data is high (Hastie et al., 2009). For ordinal response specifically, some previous theoretical and simulation work has been done to show the ordered response with 3 or 4 categories often carries the maximal information. If the number of categories goes beyond 5, it is

preferred to be treated as continuous when modeling (Pasta, 2009). By exploring things computationally, we also conclude the predictive power is closely related to the sparsity in data. When applying GMIFS to the Glue Grant dataset, substantially more features were selected and a lower predictive accuracy was obtained at the optimal step when using Marshall score on central nervous system as the ordinal outcome compared to when using renal system as the ordinal outcome. The high denseness of ‘-omics’ data in neurology has been discussed (e.g., Karssen et al. (2006)), where a relatively large set of modest changes in gene expression rather than a small set of strong signals are often considered to incur mental disorders with high variability, complexity, and heterogeneity. A very interesting question then arises is whether GMIFS can correctly select all the important covariates in the context of GLM and GLMM. Previously, for forward stagewise in high-dimensional linear models, Bühlmann (2006) has proved its consistency which is defined as $\lim_{n \rightarrow \infty} P(\hat{M}_n = M) = 1$, where M and $\hat{M}_n$ are the active set of variables in true and estimated models, respectively. As far as we know, little has been done regarding the theoretic property of GMIFS in a more generalized framework.

In the original paper wherein the GMIFS algorithm for a binary outcome was introduced (Hastie et al., 2007), there was no stopping criteria and tuning parameters were needed as the Boosting-type algorithm had its own resistance to over-fitting (Zhao and Yu, 2004). In our modified GMIFS, we implemented two convergence criteria to control the total number of covariates in the active set. We argue it would be unnecessary for GMIFS to accumulate a large number of covariates that are noisy and unimportant in the active set, as that contradicts our goal of detecting a few features to build a parsimonious model. The two convergence criteria we implemented are: 1) The difference between successive log-likelihoods is smaller than a given tolerance, e.g., $\delta = 10^{-6}$; 2) The proportion of features with nonzero coefficients reaches a pre-specified number. From our experience, criterion 1) can often be satisfied in data where the correlation between observations is negligible. However, criterion 1) is seldom met in longitudinal data as the likelihood is fluctuating due to heterogeneities among subjects. Instead, the longitudinal high-dimensional data often converges according to criterion 2). Although the choice of cutoff value in criterion 2) is somewhat arbitrary, based on the ‘Bet on Sparsity’ principle discussed above, only a small number of features with nonzero coefficients should be included the final model, thus a large enough proportion should be able to cover all the important features if the variable screening property holds (see below). After feature selection, we fit a sequence of random coefficient ordinal response models and selected the best parsimonious model based on model fitting criteria BIC. This procedure can be computationally expensive without the help of parallel computing. In addition, the random coefficient model may be at a risk of not converging as the number of features increases. An improved procedure is desired to shorten the computational time while maintaining the precision of model selection.

In addition to the future research direction discussed above, there are more things to consider. The connection between the penalization model and Boosting-type methods is of great interest at both computational and theoretical levels. Most of the work has been done in the context of linear regression. Notably, Hastie et al. (2007) defines the L_1 arc-length of

$\beta(t)$ be $\int_0^t \left\| \frac{\partial \beta(s)}{\partial s} \right\| ds$, where $\beta(t)$ is a one-dimensional differentiable curve. Hastie et al.

(2007) shows the L_1 arc-length is equal to the L_1 norm in the expanded representation where the forward stagewise computes the solution to the monotone LASSO. For a general convex loss function, the connection between L_1 -regularized GLM path (Park and Hastie, 2007) and the forward stagewise path has not been explicitly studied. Zhao and Yu (2004) proposed a very interesting algorithm called `boosted_lasso` algorithm to link LASSO and forward stagewise. It allows backward steps in forward stagewise fitting when it deviates from the lasso regularization path. Under the negative log-likelihood loss function, this approach approximates the L_1 -regularized GLM path. This could serve as a starting point for our future research.

The other interesting aspect is to explore theoretically how penalization models or Boosting-type methods behave when fit using high-dimensional longitudinal or clustered data. In fact, this area of research is still at a very early stage and has caught wide attention in fields of statistics as well as computer science. Some pioneering theoretical and computational work of L_1 penalization in the context of linear mixed-effects models has been done by Schellldorfer et al. (2011). In addition, Chen et al. (2014) published a review paper on recent progress in statistical learning methods in longitudinal high-dimensional data, where methods such as Longitudinal Support Vector Classifier (LSVC)(Chen and Bowman, 2011), a penalized joint log-likelihood with an adaptive penalty for selection of both fixed and random effects simultaneously (Bondell et al., 2010) and a class of nonconcave penalized profile likelihood with proxy matrix (Fan and Li, 2012) have been discussed and compared. For generalized linear mixed models, the penalized quasi-likelihood approach for binary and count data have been proposed by (Groll and Tutz, 2012) and (Schellldorfer et al., 2011). However, much still remains unexplored and some of the interesting questions remain unanswered, such as whether the variable screening property (Bühlmann and van de Geer, 2011) still holds in longitudinal high-dimensional data with a discrete response.

In a conclusion, we provide currently the only solution for analyzing longitudinal ordinal response for high-dimensional data, which brings a new broad of perspective for a refined classification system in data with high volume and large variability. The application of statistical learning to clustered and time-dependent data in translational medicine field is unprecedented and exciting, it is expected to utilize the complex ‘-omics’ data and other health information to monitor the disease progression and help physicians make real-time treatment decisions.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Research reported in this publication was supported by the National Library Of Medicine of the National Institutes of Health under Award Number R01LM011169. The authors would like to thank the contribution of the Inflammation and the Host Response to Injury Investigators and the support of the Large-scale Collaborative Project Award (5U54GM062119) from the National Institute of General Medical Sciences for providing the data. The content is solely the responsibility of the authors and does not necessarily represent the official view of the National Institute of Health.

References

- Abramowitz, M.; Stegun, IA., et al. Handbook of Mathematical Functions. Vol. 1. Dover New York: 1972.
- Bates D, Maechler M, Bolker B. lme4: Linear mixed-effects models using eigen and jags. 2012
- Bondell HD, Krishna A, Ghosh SK. Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics*. 2010; 66:1069–1077. [PubMed: 20163404]
- Bühlmann P. Boosting for high-dimensional linear models. *The Annals of Statistics*. 2006;559–583.
- Bühlmann, P.; van de Geer, S. *Statistics for High-dimensional Data: Methods, Theory and Applications*. Heidelberg, Germany: Springer; 2011.
- Burgon P, Lee W, Nixon A, Peralta E, Casey P. Phosphorylation and nuclear translocation of a regulator of g protein signaling (rgs10). *Journal of Biological Chemistry*. 2001; 276:32828–32834. [PubMed: 11443111]
- Chen S, Bowman FD. A novel support vector classifier for longitudinal high-dimensional data and its application to neuroimaging data. *Statistical Analysis and Data Mining*. 2011; 4:604–611. [PubMed: 25309639]
- Chen S, Grant E, Wu TT, Bowman FD. Some recent statistical learning methods for longitudinal high-dimensional data. *Wiley Interdisciplinary Reviews: Computational Statistics*. 2014; 6:10–18. [PubMed: 25285184]
- Christensen RHB. R package ordinal: Regression models for ordinal data. 2013 Available at <http://cran.r-project.org/web/packages/ordinal/index.html>.
- Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. *Annals of Statistics*. 2004; 32:407–499.
- Efron B, Morris C. Stein's estimation rule and its competitors: an empirical bayes approach. *Journal of the American Statistical Association*. 1973; 68:117–130.
- Ezzet F, Whitehead J. A random effects model for ordinal responses from a crossover trial. *Statistics in Medicine*. 1991; 10:901–907. [PubMed: 1876780]
- Fan Y, Li R. Variable selection in linear mixed effects models. *Annals of statistics*. 2012; 40:2043–2068. [PubMed: 24850975]
- Fatemi S. Reelin mutations in mouse and man: from reeler mouse to schizophrenia, mood disorders, autism and lissencephaly. *Molecular Psychiatry*. 2001; 6:129–133. [PubMed: 11317213]
- Groll A. R package glmmlasso: Variable selection for generalized linear mixed models by 11-penalized estimation. 2014 Available at <http://cran.r-project.org/web/packages/glmmlasso/index.html>.
- Groll A, Tutz G. Variable selection for generalized linear mixed models by 11-penalized estimation. *Statistics and Computing*. 2012;1–18.
- Harville D, Mee R. A mixed-model procedure for analyzing ordered categorical data. *Biometrics*. 1984; 40:393–408.
- Hastie T, Taylor J, Tibshirani R, Walther G. Forward stagewise regression and the monotone lasso. *Electronic Journal of Statistics*. 2007; 2007:1–29.
- Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. 2nd edition. New York, NY: Springer; 2009.
- Heagerty PJ, Zeger SL. Marginal regression models for clustered ordinal measurements. *Journal of the American Statistical Association*. 1996; 91:1024–1036.
- Hedeker D, Gibbons R. A random-effects ordinal regression model for multilevel analysis. *Biometrics*. 1994; 50:933–944. [PubMed: 7787006]
- Hedeker, D.; Gibbons, R. *Longitudinal Data Analysis*. Hoboken, NJ: John Wiley & Sons, Inc.; 2006.
- Ibrahim, A.; Sarhane, K.; Fagan, S.; Gorman, J. Technical report. Massachusetts General Hospital; 2013. Renal dysfunction in burns: A review.
- Karssen A, Li J, Her S, Patel P, Meng F, Evans S, Vawter M, Tomita H, Choudary P, Bunney WE, Jones EG, Watson S, Akil H, Myers RM, Schatzberg AF, Lyons DM. Application of microarray technology in primate behavioral neuroscience research. *Methods*. 2006; 38:227–234. [PubMed: 16469505]

- Li C, Wong WH. Model-based analysis of oligonucleotide arrays: Model validation, design issues and standard error application. *Genome Biology*. 2001; 2:1–11.
- Marshall J. Multiple organ dysfunction score: A reliable descriptor of a complex clinical outcome. *Critical Care Medicine*. 1995; 23:1638–1652. [PubMed: 7587228]
- McCullagh P. Regression models for ordinal data(with discussion). *Journal of the Royal Statistical Society; Series B*. 1980; 42:109–142.
- McKibben J, Bresnick M, Askay S, Fauerbach J. Acute stress disorder and posttraumatic stress disorder: a prospective study of prevalence, course, and predictors in a sample with major burn injuries. *Journal of Burn Care & Research*. 2008; 29:22–35. [PubMed: 18182894]
- Opgen-Rhein R, Strimmer K. R package longitudinal: Analysis of multiple time course data. 2014 Available at <http://cran.r-project.org/web/packages/longitudinal/index.html>.
- Park MY, Hastie T. L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2007; 69:659–677.
- Pasta, D. SAS Global Forum. Washington, DC: 2009. Learning when to be discrete: Continuous vs. categorical predictors.
- Pinheiro JC, Bates DM. Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics*. 1995; 4:12–35.
- Pullamsetti S, Kiss L, Ghofrani H, Voswinckel R, Haredza P, Klepetko W, Aigner C, Fink L, Muiyal J, Weissmann N, Grimminger F, Seeger W, Schermuly R. Increased levels and reduced catabolism of asymmetric and symmetric dimethylarginines in pulmonary hypertension. *The FASEB Journal*. 2005; 19:1175–1177.
- Renneberg B, Ripper S, Schulze J, Seehausen A, Weiler M, Wind G, Hartmann B, Germann G, Liedl A. Quality of life and predictors of long-term outcome after severe burn injury. *Journal of Behavioral Medicine*. 2013:1–10.
- Rivero G, Gabilondo A, García-Sevilla J, Callado L, Harpe RL, Morentin B, Meana J. Brain RGS4 and RGS10 protein expression in Schizophrenia and depression effect of drug treatment. *Psychopharmacology*. 2013; 226:1–12. [PubMed: 23344556]
- Rivero G, Gabilondo A, García-Sevilla J, Harpe RL, Morentín B, Meana J. Characterization of regulators of g-protein signaling rgs4 and rgs10 proteins in the postmortem human brain. *Neurochemistry International*. 2010; 57:722–729. [PubMed: 20816714]
- Sahin F, Qiu W, Wilentz RE, Iacobuzio-Donahue CA, Grosmark A, Su GH. Rpl38, fosl1, and upp1 are predominantly expressed in the pancreatic ductal epithelium. *Pancreas*. 2005; 30:158. [PubMed: 15714138]
- Schapire R, Freund Y, Bartlett P, Lee W. Boosting the margin: A new explanation for the effectiveness of voting methods. *Annals of Statistics*. 1998; 26:1651–1686.
- Schellldorfer J. R package glmmlmixedlasso: Generalized linear mixed models with lasso. 2014 Available at https://r-forge.r-project.org/R/?group_id=984.
- Schellldorfer J, Bühlmann P, van de Geer S. Estimation for high-dimensional linear mixed-effects models using l1-penalization. *Scandinavian Journal of Statistics*. 2011; 38:197–214.
- Schellldorfer J, Meier L, Bühlmann P, Winterthur A, Zürich E. Glmmllasso: An algorithm for high-dimensional generalized linear mixed models using l1-penalization. *Journal of Computational and Graphical Statistics*. 2013
- Tai Y, Speed T. A multivariate empirical bayes statistic for replicated microarray time course data. *Annals of Statistics*. 2006; 34:2387–2412.
- Yang H. Variable selection procedures for generalized linear mixed models in longitudinal data analysis. 2007
- Yang, M. Multinomial regression. In: Leyland, A.; Goldstein, H., editors. *Multilevel Modeling of Health Sciences*. New York, NY: John Wiley & Sons; 2001.
- Yuan M, Kendziorski C. Hidden markov models for microarray time course data under multiple biological conditions (with discussion). *Journal of the American Statistical Association*. 2006; 101:1323–1340.
- Zhang Y, Tibshirani R, Davis R. Classification of patients from time-course gene expression. *Biostatistics*. 2013; 14:87–98. [PubMed: 22926914]

- Zhao, P.; Yu, B. Technical report. Berkeley: University of California; 2004. Boosted lasso.
- Zhou B, Xu W, Herndon D, Tompkins R, Davis R, Xiao W, Wong WH. Analysis of factorial time-course microarrays with application to a clinical study of burn injury. *Proceedings of the National Academy of Sciences*. 2010; 107:9923–9928.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 1.

Frequency by day when buffy coat samples were collected and hybridized to Affymetrix HG-U133 Plus 2.0 arrays from 169 burn injury patients during their hospitalization.

Figure 2.
Stacked barplot for Marshall Score on the renal system (top) and the central nervous system (bottom) for individuals at time points: baseline, days 4, 7, 14, and 21 since hospitalization.

Figure 3.

Top panel: The regularization profile from GMIFS for the burn injury data with Marshall score on the renal system as the ordinal outcome with three ordered categories. The horizontal axis is the step GMIFS has undertaken and the vertical axis represents the penalized estimate for the coefficients. Bottom panel: The model fitting criteria BIC from the random intercept ordinal response model for the burn injury data with Marshall score on the renal system as the ordinal outcome with three ordered categories. The horizontal axis is the number of features in the active set and the vertical axis is the model fitting criteria (BIC).

Figure 4.

Top panel: The regularization profile from GMIFS for the burn injury data with Marshall score on the central nervous system as the ordinal outcome with three ordered categories. The horizontal axis is the step GMIFS has undertaken and the vertical axis represents the penalized estimate for the coefficients. Bottom panel: The model fitting criteria BIC from the random intercept ordinal response model for the burn injury data with Marshall score on the central nervous system as the ordinal outcome with three ordered categories. The horizontal axis is the number of features in the active set and the vertical axis is the model fitting criteria (BIC).

Table 1

Simulation results for high-dimensional data with clustered binary outcome generated from balanced and unbalanced designs.

Balanced Design					
Estimate	TRUE	GMIFS/glmer	LASSO/glmer	gmmLasso	LASSO
σ_{21}	2	2.46	2.46	1.14	-
σ_{22}	1	1.11	1.11	0.52	-
(Intercept)	0	-0.60(0.30)	-0.60(0.30)	-0.45	-0.36(0.13)
β_2	1	0.97(0.27)	0.97(0.27)	0.67	0.53(0.22)
β_3	-1	-1.08(0.19)	-1.08(0.19)	-0.41	-0.70(0.09)
β_4	1	1.14(0.19)	1.14(0.19)	0.46	0.77(0.05)
β_5	-1	-0.95(0.18)	-0.95(0.18)	-0.32	-0.60(0.16)
Prediction Error		0.123	0.123	0.118	0.310
True Positives		5	5	5	5
Time(sec)		5710.93	1887.87	6495.28	22.43
Unbalanced Design					
Estimate	TRUE	GMIFS/glmer	LASSO/glmer	gmmLasso	LASSO
σ_{21}	2	3.78	3.43	1.86	-
σ_{22}	1	1.17	1.15	0.56	-
(Intercept)	0	-0.63(0.35)	-0.63(0.40)	-0.52	-0.41(0.17)
β_2	1	-	-	0.30	-
β_3	-1	-0.77(0.18)	-0.77(0.05)	-0.19	-0.48(0.27)
β_4	1	1.18(0.20)	1.10(0.01)	0.49	0.68(0.10)
β_5	-1	-0.82(0.17)	-0.78(0.05)	-0.27	-0.47(0.28)
β_{415}	0	-	-0.46(0.21)	-0.05	-0.29(0.08)
β_{440}	0	-	-	-0.02	-
β_{751}	0	0.58(0.17)	-	0.09	-
Prediction Error		0.137	0.127	0.149	0.310
True Positives		4	4	5	4
Time(sec)		5121.20	1303.38	9559.24	25.82

TRUE indicates the underlying parameter value; GMIFS logistic model with random effects (GMIFS/glmer), LASSO logistic model with random effects (LASSO/glmer), glmmLasso, LASSO indicate the parameter estimates (and standard error if possible) in the optimal model using four approaches, respectively. The optimal model for GMIFS logistic model with random effects glmer, LASSO logistic model with random effects and LASSO were selected according to the 'elbow criterion' in BIC and the optimal model for glmmLasso was selected according to the minimal BIC.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2
Overall frequencies across all timepoints for the original and modified Marshall score for the renal system.

Original	Normal(0)	Mild(1)	Moderate(2)	Markedly(3)	Severe(4)
	236	395	17	4	5
Modified	Normal(0)	Mild(1)	Moderate+(2+)		
	236	395	26		

Table 3

Contingency table of the observed and predicted three-category Marshall score assessed on renal system. The predicted ordinal outcome is calculated using the random intercept ordinal response model.

Pred/Obs	Normal(0)	Mild(1)	Moderate+(2+)
Normal(0)	196	27	1
Mild(1)	40	366	17
Moderate+(2+)	0	2	8

Table 3

Contingency table of the observed and predicted three-category Marshall score assessed on renal system. The predicted ordinal outcome is calculated using the random coefficient ordinal response model.

Pred/Obs	Normal(0)	Mild(1)	Moderate+(2+)
Normal(0)	214	9	0
Mild(1)	22	385	14
Moderate+(2+)	0	1	12

Table 5

Overall frequencies across all timepoints for the original and modified Marshall score for the central nervous system.

Original	Normal(0)	Mild(1)	Moderate(2)	Markedly(3)	Severe(4)
	210	48	125	84	190
Modified	Normal/Mild(0.1)	Moderate/Markedly(2.3)	Severe (4)		
	258	209	190		

Table 6

Contingency table of the observed and predicted three-category Marshall score assessed on central nervous system. The predicted ordinal outcome was calculated using the random intercept ordinal response model.

Pred/Obs	Mild(0,1)	Moderate(2,3)	Severe(4)
Mild(0,1)	222	28	0
Moderate(2,3)	34	161	25
Severe(4)	2	20	165

Table 7

Contingency table of the observed and predicted three-category Marshall score assessed on central nervous system. The predicted ordinal outcome was calculated using the random coefficient ordinal response model.

Pred/Obs	Mild(0,1)	Moderate(2,3)	Severe(4)
Mild(0,1)	217	60	42
Moderate(2,3)	39	128	14
Severe(4)	2	21	134