7-7-2017

Stagewise Estimating Equations

Greg Vaughan
University of Connecticut - Storrs, gregory.vaughan@uconn.edu

Stagewise Estimating Equations

Gregory Phillip Lucas Vaughan, Ph.D.
University of Connecticut, 2017

ABSTRACT

Stagewise estimation is a slow-brewing approach for model building that has recently experienced a revival due to its computational efficiency, its flexibility in handling complex data structures, and its intrinsic connections with penalized estimation. Synthesizing generalized estimating equations to handle correlated non-Gaussian data with stagewise techniques, this thesis proposes general stagewise estimation approaches that perform model selection in the presence of complex covariate structures.

First, the setting where there is a prior covariate grouping structure or hierarchy is considered. As the grouping structure in practice is often not ideal as even important groups may contain unimportant variables, the key is to simultaneously conduct group selection and within-group variable selection, or in other words, bi-level selection. This thesis presents two approaches to address the challenge. The first is the bi-level stagewise estimating equations (BiSEE) approach, which is shown to correspond to the sparse group lasso penalized regression. The second is the hierarchical stagewise estimating equations (HiSEE) approach that can handle a more general hierarchical grouping structure, in which each stagewise estimation step itself is executed as a hierarchical

selection process based on the grouping structure.

The second setting explored is regression with interaction terms. As it is often required that main effect terms be included when an interaction term is part of a model, the goal is to perform variable selection that maintains the variable hierarchy. Two approaches are proposed by this thesis. The first is a hierarchical lasso stagewise estimating equations approach, which is shown to directly correspond to the hierarchical lasso penalized regression. The second is a stagewise active set approach, which enforces the variable hierarchy by conforming the selection to a properly growing active set in each stagewise estimation step.

Simulation studies are presented to show the efficacy and superior computational efficiency of the proposed approaches. The approaches are also used to study the association between the suicide-related hospitalization rates among 15–19 year olds in Connecticut and the characteristics of the school districts in which they reside.

Stagewise Estimating Equations

Gregory Phillip Lucas Vaughan

B.S., Mathematics, Trinity College, CT, USA, 2012

B.S., Computer Science, Trinity College, CT, USA, 2012

A Dissertation
Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy
at the
University of Connecticut

2017

APPROVAL PAGE

Doctor of Philosophy Dissertation

Stagewise Estimating Equations

Presented by

Gregory Phillip Lucas Vaughan, B.S. Mathematics/ B.S. Computer Science

Major Co-Advisor _____

Dr. Kun Chen

Major Co-Advisor _____

Dr. Jun Yan

Associate Advisor _____

Dr. Yuping Zhang

University of Connecticut

2017

Dedication

To my family. You have always been there for me, and always made the juice worth the squeeze.

To Westley, who has been a constant comfort and reassurance.

To Bayleigh, who has seen me through all of the ups and downs of graduate school, and been my anchor throughout. The weight of writing a thesis was always that much lighter with you supporting me.

To my mother, Lucy, who made me everything that I am. You taught me to be kind, to work hard, and above all, how to endure. Without these lessons, I never would have made it to graduate school, let alone complete it.

Acknowledgements

"... *once you set your heart to movin' on, there ain't no road too long*"

-Follow that Bird, 1985

I first would like to acknowledge and thank the UConn statistics department as a whole. While our relationship was never a simple one, you were an integral part of my development as a student, an academic, a teacher, and as a person. They say it takes a village to raise a child, and I feel that the saying is true for making a PhD. You will always be a part of who I am.

I want to thank Dr. Aseltine for his guidance throughout my career at UConn as a graduate assistant and as a researcher. Additionally, he provided the data that this work is motivated by; this work is made invaluable by his contribution.

Thank you to Dr. Zhang for being my associate advisor to evaluate my work. Your unbiased critiques of my work are greatly appreciated.

Thank you to Megan Petsa and Tracy Burke. You were always kind and supportive. I am certain I would have floundered constantly without you.

To the faculty of the mathematics department at Trinity college, who knew I would want to teach before I did. Their constant support and dedication to me, and all of their students, has shaped so much of who I was and who I am, as both a student and a person.

To my fellow students, especially my cohort, I will always be indebted to you. Throughout all of the struggles of graduate school we have been there for each other. There is no doubt I would not have made it to this point without you. Thank you.

I want to thank Jen McGinniss and Gregory Matthews, those students who came before me, but continue to leave their mark on me. Full of wise guidance and always supportive, both of you have always helped me to put my graduate school woes in perspective, often while putting a smile on my face.

I want to thank Valerie Nazzaro for being my academic big sister. From my earliest days as a young graduate student studying for the qualifying exam, to working on my thesis, to developing my applications to academia, and beyond, you have supported me throughout. Your help and guidance have been essential to my development as a student and as a teacher, and I will never be able to thank you enough.

Finally, I would like to thank my advisors Dr. Jun Yan and Dr. Kun Chen. The heart of a truly great teacher is one that is full of an intense passion not simply for teaching, but for the students themselves; and both of you truly are demonstrations of this. Though at times I felt that I had let you down, I could always tell what appeared to be disapproval was in fact truly deep caring. You saw potential in me, and thus only ever expected the best of me. Your continued dedication to me and to your craft has made me into a better version of myself. Thank you for everything.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Overview

In contemporary scientific research, data of large size and variety are routinely collected in various fields such as genetics, medical imaging, health sciences, etc (Fan et al., 2014). In man cases, the data is categorized as high dimensional data; that is, the data set has a large number of variables to investigate. Consider an adolescent suicide risk study from the State of Connecticut (Chen and Aseltine, 2017). Adolescent suicide prevention is a major public health concern as suicide is one of the leading causes of death among adolescents. In this study, annual suicide-related hospitalization counts for the 15–19 age group in each of 119 school districts were obtained during 2010 to 2014 from all hospitals in the state. The research interest is the association between adolescent suicide risk proxied by hospitalization counts and school district characteristics. Several categories of covariates were collected, such as demographics, prosperity measures, academic measures, and time trend.

In high dimensional settings such as the suicide risk study, there are two common goals that the researcher may have. The first possible goal is to establish a general

relational structure between the different variables; this type of task is referred to as unsupervised learning. The second possible goal is characterized by a target that the researcher is aiming for; usually this target is one of the variables that the researcher would like to be able to predict in the future. This target is usually referred to as the dependent or response variable, while the other variables in the data set used to predict the response are called the independent variables or covariates. In the context of the suicide risk study, the response variable would be suicide related hospitalizations, and the covariates would be the all of the variables in the various categories describing the school districts. Trying to predict the response variable with the covariates creates a feedback mechanism that can be used to train the model. For this reason, this goal is referred to as supervised learning. In the statistics literature, this approach is usually referred to as regression, and will be a core focus of this thesis.

When performing regression, either to investigate inferential statements or to develop a predictive model, in the high dimensional setting, it is beneficial to perform a task that is called model selection, or variable selection. Model selection is the task of identifying the best subset of covariates to perform regression with in order to reach a particular goal. In the high dimensional setting this is beneficial because performing model selection will likely result in fewer covariates being used in the final model. This can greatly improve both a model's predictive capability if the covariates are strongly correlated, as well as ease the interpret-ability of the model. Additionally, in the high dimension setting, there may be so many covariates that performing direct regression without any model selection

is not physically possible; performing model selection can also address this issue.

The basic premise of most model selection techniques can be thought of as an optimization of what is known as the combinatorial approach. In the combinatorial approach, the research decides on a criterion by which he or she will evaluate different models. This criterion is not usually an absolute measure of goodness, but rather a comparative one; that is the criterion can only tell the research whether one model is better than another. With a chosen criterion, the researcher then performs some method of regression using all possible combinations of the covariates, and calculates that criterion for each resulting model. The model with the optimal criterion value is then selected.

In the case of high dimensional data however, this approach can become infeasible. If there are p possible covariates, where each one may either be in or out of the model, then there will be a total of p^2 different subsets of covariates. Even if there were only 100 covariates and it took on average 1 second to make each model, the comparison of all of these models would take over two and a half hours. If p were increased to 300 it would take more than a day. A wildly popular approach to performing model selection that addresses the computational issues of the combinatorial approach is called penalized regression.

The computational deficiencies of the combinatorial approach are addressed by penalized regression by reducing the number of possible subsets to be considered by disregarding clearly poor subsets. Typical regression generally takes the form where a loss

or objective function f is optimized with respect to a set of parameters, denoted as a $p \times 1$ vector β, that each reflect the importance of different covariates. The loss function describes how good a certain relationship, as described by the values of β, between the covariates and the response is. One common loss function is called the observed least squares, which is the sum of the squares of the difference between values predicted by a model and the observed response values.

Penalized regression instead uses a technique called regularization where the loss function is still optimized, but the possible values of β are constrained by what is called a penalty function. One common penalty function example is the ℓ_1 norm, $\phi(\beta) = \sum_{j=1}^{p} \beta_j$, where β_j is the jth element of β. This penalty function is constrained to be a pre-specified value t, and then f is optimized subject to the penalty of the value of β being less than t. The penalized regression approach then looks at several solutions calculated using a range of values for t. The resulting models are the candidate models from which the researcher selects using the desired criterion.

A different approach to model selection is the recently revitalized stagewise estimation approach. The main idea of a stagewise procedure is to build a model from scratch, gradually increasing the model complexity in a sequence of learning steps in a way that the computation in each step is kept cheap. Consider the familiar linear regression model

$$Y_i = X_i^\top \beta + e_i, \qquad i = 1, \ldots, n,$$

where Y_i is the ith response, X_i is a $p \times 1$ covariate vector, β is a $p \times 1$ coefficient vector, and e_i's are independent random errors of zero mean. For simplicity, we assume that the responses and the predictors are centered so that there is no intercept term. Starting with $\beta^{[0]} = 0$, a stagewise procedure determines a small increment $\delta^{[t]}$ in learning step t and updates the coefficients with $\beta^{[t]} = \beta^{[t-1]} + \delta^{[t]}$. Depending on the learning objective, there are different ways to determine the "optimal" $\delta^{[t]}$. For example, in forward stagewise regression, if the jth covariate is most correlated with the current residual vector $\hat{e}^{[t-1]}$ with correlation $r_j^{[t-1]}$, then, with a predetermined ϵ, the components in δ are set with $\delta_j^{[t]} = \epsilon \cdot \text{sign}(r_j^{[t-1]})$ and $\delta_i^{[t]} = 0$ for all $i \neq j$.

In general, a properly designed/implemented stagewise procedure can efficiently trace out a path of potential models with repeated simple calculations, making it attractive in complex statistical modeling problems. Under certain conditions, the classical forward stagewise regression path converges exactly to the solution path of the most popular penalized regression approach, lasso (Tibshirani, 1996), as the step size goes to zero (Efron et al., 2004; Zhao and Yu, 2007). Nonetheless, the merit of a stagewise method does not rely on the existence of such equivalency: even when the stagewise solution path deviates from its penalized estimation counterpart, its performance can remain competitive (Tibshirani, 2015).

Variable selection in modeling non-Gaussian clustered data, such as in the suicide risk study where the suicide related hospitalization counts are not well described by a Gaussian distribution, is an important but less studied problem. The term clustered data

refers to situations where there is a clear clustering of observations or rows within the data set. These clusters usually indicate potential correlation between the observations. For example the data may come from a longitudinal study where the observations are coming from the same person at different time points. In the suicide risk study, because the hospitalization counts for each district are collected over 5 consecutive years, we expect there to be some correlation or relationship between the measurements from year to year. In these cases, accounting for the within cluster correlation would improve estimation efficiency. When the response variable takes on non-Gaussian forms such as only non-negative integers, accounting for these correlations can be difficult. One popular approach is to use Generalized Estimating Equations (GEE) (Liang and Zeger, 1986) to perform standard regression. GEEs specify only the structure of the mean and the variance of the response, as opposed to the whole distribution, and propose an approximation of the correlation structure with what is called a working correlation matrix. Though GEEs are complicated to work with, the flexible nature of stagewise techniques makes the pair a good combination that together can perform model selection in the presence of non-Gaussian correlated data.

1.2 Literature Review

1.2.1 Penalized Regression

The task of penalized regression can be formulated as solving the following constrained optimization problem,

$$\min_{\theta} f(\theta) \text{ subject to } \phi(\beta) \leq s, \tag{1.1}$$

where f is a loss function reflecting lack of fit, ϕ is a penalty function controlling the complexity of model parameter vector β, usually a subset of θ, and $s \geq 0$ is a tuning parameter that determines the amount of regularization. When ϕ is the ℓ_1-norm function, the technique is referred to lasso (Tibshirani, 1996). The development of lasso is considered pioneering work, and the use of other penalty forms have been investigated. Using a pure ℓ_2-norm on all of the coefficients, the technique is called ridge regression (Miller, 2002; Draper and Smith, 1998), which performs estimation shrinkage that can address multicollinearity issues and reduce variation. A mixture of the lasso penalty and the ridge penalty gives elastic net (Zou and Hastie, 2005), which combines the benefits of both penalties. Techniques using penalties based on the ℓ_2 norm, the group lasso (Yuan and Lin, 2006) and the sparse group lasso (Friedman et al., 2010; Simon et al., 2013b), were developed to address grouped covariates. Non-convex penalties such as the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li, 2001), and the minimax concave penalty (MCP) (Zhang, 2010) reduce penalization for larger coefficient

estimates, thereby reducing bias from estimates.

Penalized regression approaches have also undergone other exciting developments in recent years. the concept of an adaptive lasso that uses pre-determined weights to reduce the bias and improve accuracy of the model selection of the penalized regression techniques was introduced by Zou (2006). Efficient algorithms the improve computation time have also been developed for lasso (Efron et al., 2004). Additionally, the development of post model selection inference has become a popular topic (Berk et al., 2013; Tibshirani et al., 2016; Lee et al., 2016). For a broader overview of penalized regression techniques, see Bühlmann and van de Geer (2011); Huang et al. (2012).

1.2.2 Generalized Estimating Equations

GEE has become an indispensable tool for analyzing clustered data when the marginal regression parameters are of primary interest. Efficiency can be gained if the working correlation structure is closer to the truth than working independence (Liang and Zeger, 1986). Commonly-used working correlation structures include independence, exchangeable, autoregressive, and unstructured. A major advantage of GEE is that the consistency of the estimator is not affected by misspecification of the correlation structure of the clusters. The sandwich variance estimator of the GEE estimator is asymptotically justified regardless of the working correlation structure. GEE has been extended in various ways, e.g., to allow a second estimating equations for covariance parameters (Prentice and Zhao, 1991), to model binary responses (Prentice, 1988), to model

categorical responses (Liang et al., 1992), to perform model comparison (Pan, 2001), and to incorporate covariates into nuisance parameters (Yan and Fine, 2004). Several implementations of GEE are available in major statistical software and packages (e.g., Halekoh et al., 2006).

1.2.3 Stagewise Estimation

In parallel to the developments in penalized regression, there has also been a revival of interest in some classical model selection techniques. In particular, the forward stagewise procedures, also known as the ϵ-boosting methods, have drawn much attention (e.g., Friedman et al., 2000; Wolfson, 2011; Tibshirani, 2015), for which an extensive body of literature exists in both statistics and machine learning (e.g., Bühlmann and Hothorn, 2007; Schapire and Freund, 2012; Efron et al., 2004; Breiman, 1998; Hastie et al., 2009).

In the context of linear regression, the stagewise selection procedure starts from a null model, at each step selects one predictor that can best explain the current model residuals, and then updates its corresponding coefficient by a small amount ϵ to partially adjust for its predictive effect. This process is repeated until a model with a desirable complexity level is reached or the model becomes excessively large. By directly linking the stagewise procedure to the regularized estimation problem, Tibshirani (2015) proposed a general framework for conducting stagewise estimation. Starting from $\beta^{[0]} = 0$,

the procedure at step $t = 1, 2, \ldots$ performs the following:

$$\delta^{[t]} = \arg\min_{\delta \in \mathbb{R}^p} f(\beta^{[t-1]} + \delta) - f(\beta^{[t-1]}) \text{ subject to } \phi(\beta^{[t-1]} + \delta) - \phi(\beta^{[t-1]}) \leq \epsilon,$$
$$\beta^{[t]} = \beta^{[t-1]} + \delta^{[t]},$$

where $\epsilon > 0$ is the step size. When the penalty function ϕ satisfies the triangular inequality $\phi(b + c) \leq \phi(b) + \phi(c)$ for all b and c, the constraints can be simplified to $\phi(\delta) \leq \epsilon$. For several commonly-used penalty forms including lasso and group lasso, the triangle inequality holds and the computation is very fast. Furthermore, to simplify and accelerate the minimization problem, $f(\beta^{[t-1]} + \delta) - f(\beta^{[t-1]})$ can be approximated, using Taylor expansion around $\beta^{[t-1]}$, by $\langle \nabla f(\beta^{[t-1]}), \delta \rangle$, where ∇ denotes the partial derivative and $\langle \cdot, \cdot \rangle$ denotes the inner product. With these substitutions, the resulting procedure becomes the following: start from $\beta^{[0]} = 0$, and for each step $t = 1, 2, \ldots$

$$\begin{aligned} \delta^{[t]} &= \arg\min_{\delta \in \mathbb{R}^p} \langle \nabla f(\beta^{[t-1]}), \delta \rangle \text{ subject to } \phi(\delta) \leq \epsilon, \\ \beta^{[t]} &= \beta^{[t-1]} + \delta^{[t]}. \end{aligned} \tag{1.2}$$

Tibshirani (2015) showed that the stagewise path produced by (1.2) well approximates its counterpart from regularized estimation.

The connections to regularized estimation, the impressive empirical performance, and the computational efficiency, all make stagewise estimation very attractive. In the context of GEE, Wolfson (2011) proposed an ϵ-boosting method (EEBoost). This method,

however, only considered individual variable selection corresponding to the ℓ_1 penalty, which is not suitable in the presence of complicated covariate structures. Additionally, neither Wolfson (2011) nor Tibshirani (2015) addressed the handling of nuisance parameters. The presence of such parameters are often inevitable, e.g., the intercept term β_0 in generalized linear models and the parameters of the working correlation structure in GEE.

1.2.4 Model Selection with Grouped Covariates

In many applications such as the suicide risk study, the predictors may have some prior grouping structure or more generally certain hierarchical structure, and it is important to incorporate such information into the selection procedure. Under the penalized estimation framework, the most popular approach is the group lasso (gLasso) (Yuan and Lin, 2006; Meier et al., 2008; Breheny and Huang, 2009), where each group of variables is either kept or removed from the model altogether. A stagewise estimation approach analogous to group lasso was proposed by Tibshirani (2015). In practice, however, some groups may be a mix of both important and irrelevant variables. Therefore, identification of the important variables within each of the selected groups is preferred. The sparse group lasso (sgLasso) (Friedman et al., 2010; Simon et al., 2013b) conducts such bi-level selection with a convex penalty. Non-convex approaches have been developed as well (Wang et al., 2007; Huang et al., 2009; Breheny and Huang, 2009; Chen et al., 2016). Bi-level selection however, has not yet been studied in stagewise estimation.

We propose two general forward stagewise approaches for variable selection, under the general framework of generalized estimating equations (GEE) (Liang and Zeger, 1986), that allows for flexible marginal modeling for clustered data without fully specifying the within-cluster dependence structure. While some versions of penalized GEE (Fu, 2003; Wang et al., 2012; Deshpande et al., 2016) or boosted GEE (Wolfson, 2011) have been proposed, no existing GEE approach addresses the group or bi-level selection problem.

1.2.5 Interaction Selection

In the interaction selection literature, a key concept is model hierarchy. There are two forms of hierarchy that are used with interaction models that allow for simpler model interpretation. Weak hierarchy requires that a particular interaction term may be included in the model only if at least one of its corresponding main effects is included. Strong hierarchy requires that both main effects be included. Liu et al. (2013) introduced a mixture of the minimax concave penalty (MCP) and group MCP, much like the sparse group lasso (Friedman et al., 2010) for gene-environment interactions that preserves strong hierarchy, but this technique only considers model selection with interaction terms where one of the corresponding main effects are unpenalized, and thus are always included in the model. Lim and Hastie (2015) suggested a novel way to use group lasso to induce strong hierarchy in a computationally efficient way when the covariates are categorical. Zhao et al. (2009), Jenatton et al. (2011), and Bach et al. (2012) proposed penalization techniques for imposing specific hierarchical structures that can

be adapted to the interaction setting. Bien et al. (2013) also proposed a penalization approach that extended the traditional lasso approach, which elucidates the effect of imposing a hierarchical structure. Though these penalization approaches do successfully impose the desired structure, they do so at a computational expense. Zhu et al. (2014) developed a stagewise approach for model selection that maintains these hierarchy structures that has both competitive performance compared to these penalization techniques and a computational advantage; but neither this technique nor the penalization approaches are able to handle non-Gaussian clustered data.

1.3 Outline

This thesis aims to present novel methods to perform model selection in the presence of non-Gaussian clustered data with overlaying covariate structures. To demonstrate the value of these techniques, they will be applied to the suicide risk study.

In the suicide risk study, the categories of the covariates, or groups of covariates, may indicate an underlying structure that would be useful in improving estimation. So, in Chapter 2, the concept of grouped covariates is explored, and new techniques are developed to harness this structure. Two new techniques, Bi-Level Stagewise Estimating Equations (BiSEE) and Hierarchical Stagewise Estimating Equations, are presented. Theses approaches utilize these potential structures while accounting for the non-Gaussian clustered form of the data. Illustrative examples and multiple numerical

studies are presented to demonstrate the efficacy of the techniques and demonstrate their advantages over other current techniques. The chapter concludes with an analysis of the suicide risk data using the new approaches.

Additionally, it may be beneficial to investigate the interactions between the various covariates in the study. So, in Chapter 3, the challenges of performing model selection when considering interaction terms is discussed and new techniques are presented to address these issues. Again, two new methods, Hierarchical Lasso Stagewise Estimating Equations (HiLa) and Active Set Stagewise Estimating Equations (ACTS), are presented. these approaches enforce typical structures associated with interaction models while accounting for the non-Gaussian clustered form of the data. As in Chapter 2, illustrative examples and simulation studies are presented to highlight the strengths of the new techniques. The chapter will conclude with a second analysis of the suicide risk study using the new techniques to evaluate possible interaction effects between the covariates and the response.

The thesis concludes in Chapter 4. A summary of the contributions of this work are presented and various new directions for further study of stagewise techniques, with or without the inclusion of GEEs is discussed. In the appendix the software implementation that was developed for these techniques in the `R` package `sgee` is covered.

Chapter 2

Stagewise Generalized Estimating Equations with Grouped Variables

2.1 Grouped Covariates

This chapter focuses on solutions to the *bi-level* selection problem, where the covariates have a grouping structure, but those groups may not perfectly separate important and un-important variables. This requires a group level selection in which important groups are identified, and an individual level selection in which important individual covariates are identified. Building upon Wolfson (2011) and Tibshirani (2015), we develop a bi-level stagewise estimating equations (BiSEE) approach that corresponds to the sparse group lasso penalized regression. The essence of forward stagewise estimation is to build a model by gradually adding well-chosen "weak learners", which motivates a general hierarchical stagewise estimating equations (HiSEE) approach. By properly designing the process of selecting weak learners to enter the model, HiSEE can flexibly take advantage of the hierarchical group structure.

2.2 Notation

Let Y_i be a $k_i \times 1$ response vector in cluster i, for $i = 1, ..., n$. Let $X_i = (\mathbf{1}_{k_i}, X_{i\mathcal{I}_1}, \ldots, X_{i\mathcal{I}_J})$, where $\mathbf{1}_{k_i}$ is a $k_i \times 1$ vector of $1's$ and $X_{i\mathcal{I}_j}$ is a $k_i \times p_j$ matrix of grouped covariates for Y_i where $\sum_{j=1}^{J} p_j = p$. It is assumed that the groups do not overlap. The conditional mean of Y_i given X_i is specified as $E[Y_i \mid X_i] = \mu_i = g^{-1}(\eta_i)$, where $\eta_i = \beta_0 \mathbf{1}_{k_i} + X_{i\mathcal{I}_1}\beta_{\mathcal{I}_1} + \ldots + X_{i\mathcal{I}_J}\beta_{\mathcal{I}_J}$, $\beta_{\mathcal{I}_j}$ is a $p_j \times 1$ coefficient vector for $j = 1, \ldots J$, β_0 is the scalar intercept, and g is a known link function. The regression coefficient vector $\beta = (\beta_{\mathcal{I}_\infty}^\top, \ldots, \beta_{\mathcal{I}_\mathcal{J}}^\top)^\top$ is of primary interest. The conditional variance of each component Y_{ij} of Y_i, $j = 1, \ldots, k_i$, is $V[Y_{ij} \mid X_{ij}] = \psi v(\mu_{ij})$, where ψ is a scalar, and $v(\cdot)$ is a variance function as in the exponential families.

The regression coefficients (β_0, β) given (ψ, α) are estimated by solving

$$U(\beta, \beta_0, \psi, \alpha) \equiv -\sum_{i=1}^{n} D_i^\top V_i^{-1}(Y_i - \mu_i) = 0,$$

where $D_i = (\partial\mu_i/\partial\eta_i^\top)X_i$, $V_i = A_i^{1/2} R_i(\alpha) A_i^{1/2}$, $A_i = \psi \operatorname{diag}\{v(\mu_{i1}), \ldots, v(\mu_{ik_i})\}$, and $R_i(\alpha)$ is an $k_i \times k_i$ working correlation matrix parameterized as a function of a parameter vector α. A major advantage of GEE is that the consistency of the estimator is not affected by misspecification of the correlation structure of the clusters (Liang and Zeger, 1986).

Given (β_0, β), estimates of (α, ψ) can be obtained by method of moments or additional estimating equations. The alternating updating continues until convergence.

In what follows, we write $U(\beta, \beta_0, \psi, \alpha) = U(\beta, \nu)$, where $\nu = (\beta_0, \psi, \alpha)$ collects all the nuisance parameters that are not directly related to variable selection. The estimating equations $U(\beta, \nu)$ can also be partitioned based on the group structure, i.e., $U(\beta, \nu) = (U_0(\beta, \nu), U_{\mathcal{I}_1}(\beta, \nu)^\top, \ldots, U_{\mathcal{I}_J}(\beta, \nu)^\top)^\top$ where $U_{\mathcal{I}_j}(\beta, \nu) \in \mathbb{R}^{p_j}$ for $j = 1, .., J$ and $U_0(\beta, \nu) \in \mathbb{R}$ pertains to the intercept term.

2.3 Stagewise Generalized Estimating Equations

2.3.1 Bi-level Stagewise Estimating Equation

From (1.2), only the gradient of the objective function being minimized is needed in computation. Although there is no explicit objective function in a GEE setting, it is helpful to view the estimating function $U(\beta, \nu)$ as the gradient of some convex and differentiable $f(\beta, \nu)$, possibly of no closed form. The stagewise estimation can still be carried out using U itself without knowing f (e.g., Wolfson, 2011). There remains, however, several challenges in applying the framework in (1.2) to our setup. Besides the regression coefficients β, the nuisance parameters have to be properly estimated/updated during the stagewise estimation. A main advantage of stagewise estimation is its computational efficiency, which requires that the optimization problem in each step is easy to solve. This is true for the stagewise procedures corresponding to either lasso or group lasso (Tibshirani, 2015). For more sophisticated bi-level or hierarchical variable selection, it is unclear what penalty permits efficient computation.

We consider the following BiSEE procedure: starting from $\beta^{[0]} = 0$, for $t = 1, 2, \ldots$

$(t.1)$ Given $\beta^{[t-1]}$, update the nuisance parameters to obtain $\nu^{[t]}$,

$(t.2)$ $\delta^{[t]} = \underset{\delta \in \mathbb{R}^p}{\arg\min} \langle U_{[0]}(\beta^{[t-1]}, \nu^{[t]}), \delta \rangle$ subject to $\phi(\delta) \leq \epsilon$,

$(t.3)$ $\beta^{[t]} = \beta^{[t-1]} + \delta^{[t]}$,

where $U_{[0]}(\beta^{[t-1]}, \nu^{[t]}) = (U_{\mathcal{I}_1}(\beta^{[t-1]}, \nu^{[t]})^\top, \ldots, U_{\mathcal{I}_J}(\beta^{[t-1]}, \nu^{[t]})^\top)^\top$. Here $\delta \in \mathbb{R}^p$ is partitioned in the same way as β; i.e. $\delta^\top = (\delta_{\mathcal{I}_1}^\top, \ldots, \delta_{\mathcal{I}_J}^\top)^\top$, where $\delta_{\mathcal{I}_i}^\top$ is a $p_i \times 1$ vector. We propose to use the sparse group lasso penalty (Friedman et al., 2010; Simon et al., 2013b)

$$\phi(\delta) = \lambda_1 \sum_{j=1}^{J} w_j \|\delta_{\mathcal{I}_j}\|_2 + \lambda_2 \|\delta\|_1, \tag{2.1}$$

where w_js are some group-level weights, $\|\cdot\|_k$ indicates the ℓ_k norm, and λ_1 and λ_2 are two tuning parameters controlling the relative degrees of group level and individual level penalization, respectively. The relationship between λ_1 and λ_2 is best described as $\lambda_1 + \lambda_2 = 1$ with $\lambda_1 \in [0, 1]$. If these parameters instead summed to some value c, then the step size could simply be scaled by c. Unless otherwise noted, we use $w_j = \sqrt{p_j}$ to adjust for the sizes of the groups. The nuisance parameters ν and the regression coefficients β are updated separately at each step. Following Tibshirani (2015), we have also simplified the constraint to $\phi(\delta) \leq \epsilon$ because the sparse group lasso penalty

satisfies the triangle inequality. Consequently, the simplified constraint guarantees that the increment in ϕ is at most ϵ.

The central task is to solve $(t.2)$. Define

$$B_{\mathcal{I}_{j(i)}}(\gamma) = \text{sign}\{-U_{\mathcal{I}_{j(i)}}(\beta^{[t-1]}; \nu^{[t]})\}\{|U_{\mathcal{I}_{j(i)}}(\beta^{[t-1]}; \nu^{[t]})| - \gamma\lambda_2\}_+/(\gamma\lambda_1 w_j), \tag{2.2}$$

as a function of scalar γ for $j = 1, \ldots, J$, where $B_{\mathcal{I}_{j(i)}}(\gamma)$ indicates the ith element of the $p_j \times 1$ vector $B_{\mathcal{I}_j}(\gamma)$, and $(x)_+ = \max(x, 0)$. The following theorem shows that $(t.2)$ can be solved efficiently, and as expected, at each step the update only changes a subset of coefficients within a particular group.

Theorem 2.1. *Consider the problem in (t.2), where ϕ is the sparse group lasso penalty in* (2.1) *with $\lambda_1 > 0$. Then the problem is solved as follows. First select k to be*

$$k = \underset{j:j\in\{1,\ldots,J\}}{\arg\max}\ \gamma_j$$

where γ_j is such that $\|B_{\mathcal{I}_j}(\gamma_j)\|_2 = 1$, with $B_{\mathcal{I}_j}(\gamma_j)$ as defined in (2.2)*. Then $\delta^{[t]}$ is given by*

$$\delta^{[t]}_{\mathcal{I}_k} = \frac{\epsilon B_{\mathcal{I}_k}(\gamma_k)}{\lambda_1 w_k + \lambda_2 \|B_{\mathcal{I}_k}(\gamma_k)\|_1},\ \textit{and}\ \delta^{[t]}_{\mathcal{I}_j} = 0,\ \forall j \neq k. \tag{2.3}$$

Theorem 2.1 is proven by using the Karush–Kuhn–Tucker conditions of $(t.2)$. Intuitively, γ_j evaluates the importance of a group as a whole. The group with the largest γ_j is selected and the configuration within the group that provides the most benefit is

determined. Theorem 2.1 ensures that BiSEE can achieve bi-level selection efficiently. Furthermore, the results encompass lasso with $\lambda_1 = 0$ and the group lasso with $\lambda_2 = 0$ as special cases. The proof of Theorem 2.1 and more specifics of the BiSEE method are available in the Web-based Supplementary Materials.

In practice, the tuning parameters involved in ϕ need to be selected. Because $\lambda_1 + \lambda_2 = 1$, we can choose a sequence of λ_1 values between $[0, 1]$ and fit BiSEE with each fixed λ_1 value. We then refit each of the unique models that appear in the paths generated by BiSEE using traditional GEE and use cross-validation to select the best model. The step size ϵ also needs to be selected with care. In general, too large a step size would produce inaccurate and unstable paths while too small a step size may cause unnecessary computation burden. We suggest examining the trace plot when selecting the step size. A sensitivity study using the illustrative example in Section 2.3.6 is reported in the Web-based Supplementary Materials.

2.3.2 Proof of Theorem 2.1

Proof. In the problem $(t.2)$, i.e.,

$$\min_{\delta \in \mathbb{R}^p} \langle U_{[0]}(\beta^{[t-1]}, \nu^{[t]}), \delta \rangle \text{ subject to } \phi(\delta) \leq \epsilon,$$

both $\langle U_{[0]}(\beta^{[t-1]}, \nu^{[t]}), \delta \rangle$ and $\phi(\delta)$ are convex functions of δ. Specifically, since the ℓ_1 and ℓ_2 norms are convex functions, it is clear that ϕ is convex as well. Since $\langle U_{[0]}(\beta^{[t-1]}, \nu^{[t]}), \delta \rangle$

is linear in δ, the convexity of the function is also clear. Since the problem is convex, the regularity conditions have thus been met and it only remains to show that the solution proposed by Theorem 2.1 satisfies the Karush–Kuhn–Tucker (KKT) conditions for $(t.2)$.

For notational simplicity, $U_{[0]}(\beta^{[t-1]}; \nu^{[t]})$ from Equation $(t.2)$ will be shortened to just U, and the scaling factor $\epsilon/(\lambda_1 w_k + \lambda_1 \|B_{\mathcal{I}_k}(\gamma_k)\|_1)$ from Equation (2.3) will be represented as c.

Let $W = \operatorname{diag}\{\operatorname{diag}(w_1 \mathbf{1}_{p_1}^\top), \ldots, \operatorname{diag}(w_J \mathbf{1}_{p_J}^\top)\}$ be a $p \times p$ block diagonal matrix, $S_{\mathcal{I}_j}(\delta^{[t]})$ be an element of the subgradient of the ℓ_2 norm evaluated at $\delta^{[t]}_{\mathcal{I}_j}$, and $Q_{\mathcal{I}_{j(i)}}(\delta^{[t]})$ be an element of the subgradient of the absolute value function evaluated at $\delta^{[t]}_{\mathcal{I}_{j(i)}}$. The forms of the ℓ_2 norm and the absolute value function sub-gradients are

$$\partial \|y\|_2 \Big|_{y=x} = \begin{cases} \frac{x}{\|x\|_2} & x \neq \mathbf{0} \\ \mathcal{L} \in \{\mathcal{L} : \|\mathcal{L}\|_2 \leq 1\} & x = \mathbf{0} \end{cases}, \text{ and } \quad \partial |y| \Big|_{y=x} = \begin{cases} \operatorname{sign}(x) & x \neq 0 \\ \mathcal{L} \in [-1, 1] & x = 0 \end{cases}.$$

The KKT conditions that correspond to Equation $(t.2)$ evaluated at $\delta = \delta^{[t]}$ are

$$-U = \gamma_k \{\lambda_1 W S(\delta^{[t]}) + \lambda_2 Q(\delta^{[t]})\}, \tag{2.4}$$

$$\gamma_k \geq 0,$$

$$\gamma_k (\lambda_1 \sum_{j=1}^{J} w_j \|\delta^{[t]}_{\mathcal{I}_j}\|_2 + \lambda_2 \sum_{j=1}^{J} \sum_{i=1}^{p_j} |\delta^{[t]}_{\mathcal{I}_{j(i)}}| - \epsilon) = 0,$$

$$\lambda_1 \sum_{j=1}^{J} w_j \|\delta^{[t]}_{\mathcal{I}_j}\|_2 + \lambda_2 \sum_{j=1}^{J} \sum_{i=1}^{p_j} |\delta^{[t]}_{\mathcal{I}_{j(i)}}| - \epsilon \leq 0.$$

By construction, $\gamma_k \geq 0$ and $\phi(\delta^{[t]}) = \epsilon$, so only (2.4) remains to be demonstrated.

We proceed by showing that each equality $-U_{\mathcal{I}_j(i)} = \gamma_k\{\lambda_1 WS(\delta^{[t]}) + \lambda_2 Q(\delta^{[t]})\}_{\mathcal{I}_j(i)}$ for $j \in \{1, \ldots, J\}$ and $i \in \{1, \ldots, p_j\}$ from Equation (2.4), holds. Each equation is demonstrated in one of the following cases: 1) when $j = k$ and i is such that $|U_{\mathcal{I}_k(i)}| > \gamma_k\lambda_2$, 2) when $j = k$ and i is such that $|U_{\mathcal{I}_k(i)}| \leq \gamma_k\lambda_2$, and 3) when $j \neq k$.

Consider case 1,) which implies that $\delta^{[t]}_{\mathcal{I}_k(i)} = c\,\text{sign}(-U_{\mathcal{I}_k(i)})(|U_{\mathcal{I}_k(i)}| - \gamma_k\lambda_2)/\gamma_k\lambda_1 w_k$. Then

$$
\begin{aligned}
&\gamma_k\{\lambda_1 WS(\delta^{[t]}) + \lambda_2 Q(\delta^{[t]})\}_{\mathcal{I}_k(i)} \\
&= \gamma_k \left\{\lambda_1 w_k S_{\mathcal{I}_k(i)}(\delta^{[t]}) + \lambda_2 Q_{\mathcal{I}_k(i)}(\delta^{[t]})\right\} \\
&= \gamma_k \left\{\lambda_1 w_k \frac{cB_{\mathcal{I}_k(i)}(\gamma_k)}{\|cB_k(\gamma_k)\|_2} + \lambda_2\,\text{sign}(cB_{\mathcal{I}_k(i)}(\gamma_k))\right\} \\
&= \gamma_k \left\{\lambda_1 w_k \frac{c}{c}\frac{\text{sign}(-U_{\mathcal{I}_k(i)})[|U_{\mathcal{I}_k(i)}| - \gamma_k\lambda_2]}{\gamma_k\lambda_1 w_k\|B_k(\gamma_k)\|_2} + \lambda_2\,\text{sign}\left(c\frac{\text{sign}(-U_{\mathcal{I}_k(i)})[|U_{\mathcal{I}_k(i)}| - \gamma_k\lambda_2]}{\gamma_k\lambda_1}\right)\right\} \\
&= -U_{\mathcal{I}_k(i)} - \text{sign}(-U_{\mathcal{I}_k(i)})\gamma_k\lambda_2 + \lambda_2\,\text{sign}(-U_{\mathcal{I}_k(i)})\gamma_k \\
&= -U_{\mathcal{I}_k(i)}.
\end{aligned}
$$

So, the equations hold in this case.

Now consider case 2), which implies that $\delta^{[t]}_{\mathcal{I}_k(i)} = 0$ and that $Q_{\mathcal{I}_k(i)}(\delta^{[t]})$ can be selected to be $-U_{\mathcal{I}_k(i)}/\gamma_k\lambda_2$. We see that the equations hold in this case:

$$\gamma_k\{\lambda_1 WS(\delta^{[t]}) + \lambda_2 Q(\delta^{[t]})\}_{\mathcal{I}_k(i)} = \gamma_k\{\lambda_1 w_k S_{\mathcal{I}_k(i)}(\delta^{[t]}) + \lambda_2 Q_{\mathcal{I}_k(i)}(\delta^{[t]})\}$$

$$= \gamma_k\{\lambda_2 Q_{\mathcal{I}_k(i)}(\delta^{[t]})\} = \gamma_k\lambda_2\frac{-U_{\mathcal{I}_k(i)}}{\gamma_k\lambda_2} = -U_{\mathcal{I}_k(i)}.$$

Consider now case 3). Since $\delta^{[t]}_{\mathcal{I}_j} = \mathbf{0}$, we must select $S_{\mathcal{I}_k}(\delta^{[t]})$ from the set $\{\mathcal{L} : \|\mathcal{L}\|_2 \leq 1\}$ and $Q_{\mathcal{I}_k(i)}(\delta^{[t]})$ from $[-1, 1]$ for all i. Since $\gamma_k \geq \gamma_j \ \forall j$ and $\|B_{\mathcal{I}_j}(\gamma_j)\|_2 = 1$ by construction, we can conclude that $\|B_{\mathcal{I}_j}(\gamma_k)\|_2 \leq \|B_{\mathcal{I}_j}(\gamma_j)\|_2 = 1$. Thus, it is valid to select

$$S_{\mathcal{I}_j(i)}(\delta^{[t]}) = B_{\mathcal{I}_j(i)}(\gamma_k) = \frac{-\operatorname{sign}(U_{\mathcal{I}_j(i)})[|U_{\mathcal{I}_j(i)}| - \gamma_k\lambda_2]}{\gamma_k\lambda_1 w_j},$$

and we do so. We now proceed in two sub-cases: 3a) when i is such that $|U_{\mathcal{I}_j(i)}| > \gamma_k\lambda_2$, and 3b) when i is such that $|U_{\mathcal{I}_j(i)}| \leq \gamma_k\lambda_2$.

For the first sub-case we can let $Q_{\mathcal{I}_k(i)}(\delta^{[t]}) = \operatorname{sign}(-U_{\mathcal{I}_j(i)})$ and we see that

$$\begin{aligned}
&\gamma_k(\lambda_1 W S(\delta^{[t]}) + \lambda_2 Q(\delta^{[t]}))_{\mathcal{I}_j(i)} \\
&= \gamma_k\left\{\lambda_1 w_j S_{\mathcal{I}_j(i)}(\delta^{[t]}) + \lambda_2 Q_{\mathcal{I}_j(i)}(\delta^{[t]})\right\} \\
&= \gamma_k\left\{\lambda_1 w_j \frac{\operatorname{sign}(-U_{\mathcal{I}_j(i)})[|U_{\mathcal{I}_j(i)}| - \gamma_k\lambda_2]}{\gamma_k\lambda_1 w_j} + \lambda_2 \operatorname{sign}(-U_{\mathcal{I}_j(i)})\right\} \\
&= -U_{\mathcal{I}_j(i)} - \operatorname{sign}(-U_{\mathcal{I}_j(i)})\gamma_k\lambda_2 + \lambda_2 \operatorname{sign}(-U_{\mathcal{I}_j(i)})\gamma_k \\
&= -U_{\mathcal{I}_j(i)}.
\end{aligned}$$

Finally, consider the second sub-case. If we select $Q_{\mathcal{I}_j(i)}(\delta^{[t]}) = -U_{\mathcal{I}_j(i)}/\lambda_2\gamma_k$ then we

see that the KKT conditions continue to hold:

$$\gamma_k(\lambda_1 WS(\delta^{[t]}) + \lambda_2 Q(\delta^{[t]}))_{\mathcal{I}_j(i)} = \gamma_k\{\lambda_1 w_j S_{\mathcal{I}_j(i)}(\delta^{[t]}) + \lambda_2 Q_{\mathcal{I}_j(i)}(\delta^{[t]})\}$$
$$= \gamma_k\{\lambda_2 Q_{\mathcal{I}_j(i)}(\delta^{[t]})\} = \gamma_k\lambda_2\frac{-U_{\mathcal{I}_j(i)}}{\gamma_k\lambda_2} = -U_{\mathcal{I}_j(i)}.$$

□

2.3.3 Lasso and Group Lasso As Special Cases

Our results in Theorem 2.1 generalize both the special case of lasso that corresponds to $\lambda_1 = 0$ and the special case of group lasso which corresponds to $\lambda_2 = 0$.

When $\lambda_2 = 0$, it is clear that $\gamma_j = \|U_{\mathcal{I}_j}(\beta^{[t-1]}; \nu^{[t]})\|_2/w_j$ for all j which in turn yields an update of $\delta^{[t]}_{\mathcal{I}_k} = -\epsilon U_{\mathcal{I}_k}(\beta^{[t-1]}; \nu^{[t]})/w_k\|U_{\mathcal{I}_k}(\beta^{[t-1]}; \nu^{[t]})\|_2$ when k is selected to be $\underset{j:j\in\{1,\ldots,J\}}{\arg\max}\ \gamma_j$.

In the case of $\lambda_1 = 0$, Theorem 2.1 does not directly provide a solution, but as λ_1 goes to 0, the solution given in Theorem 2.1 does converge to the solution obtained when $\lambda_1 = 0$. Specifically, as $\lambda_1 \to 0$, we see that at some point, for all $j \in \{1, \ldots, J\}$, $B_{\mathcal{I}_j}(\gamma_j)$ contains only one non-zero element, $B_{\mathcal{I}_j(l_k)}(\gamma_j)$, where l_k is such that $|U_{\mathcal{I}_j(l_k)}(\beta^{[t-1]}; \nu^{[t]})| \geq |U_{\mathcal{I}_j(i)}(\beta^{[t-1]}; \nu^{[t]})|$ for all $i \in \{1, \ldots, p_j\}$. At this point $\|B_{\mathcal{I}_j}(\gamma_j)\|_2 = |B_{\mathcal{I}_j(l_k)}(\gamma_j)| = 1$, which implies that $\gamma_j = |U_{\mathcal{I}_j(l_k)}(\beta^{[t-1]}; \nu^{[t]})|/(\lambda_2 + \lambda_1 w_j)$. Therefore, we have $\gamma_j \to |U_{\mathcal{I}_j(l_k)}(\beta^{[t-1]}; \nu^{[t]})|$, and thus to select $k = \underset{j:j\in\{1,\ldots,J\}}{\arg\max}\ \gamma_j$ is to select the group with the largest individual contributor, which yields the update

$\delta^{[t]}_{\mathcal{I}_k(l_k)} = -\epsilon \,\mathrm{sign}\{U_{\mathcal{I}_k(l_k)}(\beta^{[t-1]}; \nu^{[t]})\}$.

For completeness, we present these two special cases in the following corollaries.

Corollary 2.2. *When* $\phi(\delta) = \sum_{j=1}^{J} w_j \|\delta_{\mathcal{I}_j}\|_2$, *the solution to (t.2) is*

$$\delta^{[t]}_{\mathcal{I}_k} = -\epsilon U_{\mathcal{I}_k}(\beta^{[t-1]}; \nu^{[t]})/w_k \|U_{\mathcal{I}_k}(\beta^{[t-1]}; \nu^{[t]})\|_2,$$
$$\delta^{[t]}_{\mathcal{I}_j} = 0, \quad \forall j \neq k,$$

where $k = \underset{j:j\in\{1,\ldots,J\}}{\arg\max} \|U_{\mathcal{I}_j}(\beta^{[t-1]}; \nu^{[t]})\|_2/w_j$.

Corollary 2.3. *When* $\phi(\delta) = \|\delta\|_1$, *the solution to (t.2) is*

$$\delta^{[t]}_{\mathcal{I}_k(l_k)} = -\epsilon \,\mathrm{sign}\{U_{\mathcal{I}_k(l_k)}(\beta^{[t-1]}; \nu^{[t]})\},$$
$$\delta^{[t]}_{\mathcal{I}_j(i)} = 0, \quad \forall (j,i) \neq (k, l_k),$$

where $(k, l_k) = \underset{(j,i):j\in\{1,\ldots,J\},i\in\{1,\ldots,p_j\}}{\arg\max} |U_{\mathcal{I}_j(i)}(\beta^{[t-1]}; \nu^{[t]})|$.

2.3.4 Hierarchical Stagewise Estimating Equation

Despite the attractive equivalency to sparse group lasso, BiSEE has some limitations. The relative weights of the group-level and the individual level regularization need to be tuned, which increases computation cost. A simpler and more direct way to achieve bi-level selection is to treat the update as a hierarchical selection process, according to the prior grouping structure of the variables. The most important group based on a

certain criterion can be identified first, and then the important variables within that group can be identified.

We thus propose the following HiSEE procedure. Starting from $\beta^{[0]} = 0$, for $t = 1, 2, \ldots$

(t.a) Given $\beta^{[t-1]}$, update the nuisance parameters to obtain $\nu^{[t]}$.

(t.b) $\delta^{(g)} = \underset{\delta \in \mathbb{R}^p}{\arg\min} \langle U_{[0]}(\beta^{[t-1]}, \nu^{[t]}), \delta \rangle$ subject to $\phi_1(\delta) \leq \epsilon_1$, and let $\mathcal{K}^{[t]} = \{j : \delta^{(g)}_{\mathcal{I}_j} \neq 0\}$.

(t.c) $\delta^{[t]} = \underset{\delta : \delta_{\mathcal{I}_j} = 0, \forall j \notin \mathcal{K}^{[t]}}{\arg\min} \langle U_{[0]}(\beta^{[t-1]}, \nu^{[t]}), \delta \rangle$ subject to $\phi_2(\delta) \leq \epsilon_2$.

(t.d) $\beta^{[t]} = \beta^{[t-1]} + \delta^{[t]}$.

Here ϕ_1 is a group-level sparsity-inducing penalty while ϕ_2 is an individual-level sparsity-inducing penalty. We propose to use the group lasso penalty for $\phi_1(\delta) = \sum_{j=1}^{J} w_j \|\delta_{\mathcal{I}_j}\|_2$ and lasso penalty for $\phi_2(\delta) = \|\delta\|_1$. This pair of penalties leads to very simple updating rules, enabling efficient computation. For simplicity, we set the group and the individual step sizes to be equal, i.e., $\epsilon_1 = \epsilon_2 = \epsilon$. The details of the procedure are given in the Web-based Supplementary Materials.

Although we mainly focus on bi-level selection here, HiSEE can be extended to hierarchical variable selection with more than two levels. In each step, at each level of the hierarchy one or more subgroups are selected until the lowest level of the structure is reached.

2.3.5 Algorithm Details

Specifics of the BiSEE and HiSEE methods are summarized in Algorithms 1 & 2. In both Algorithms, the model is initialized with an empty model, indicated by $\beta^{[0]} = 0$. In the initial step, the intercept is updated to be $\beta_0^{[t]}$, the root of $U_0(\beta^{[t-1]}, \beta_0, \psi^{[t-1]}, \alpha^{[t-1]})$ with respect to β_0. Other nuisance parameters ψ and α are updated with the method of moments from the Pearson residuals evaluated at $\beta^{[t-1]}$ (Liang and Zeger, 1986). Our algorithm can be extended to incorporate covariates in ψ and α (Yan and Fine, 2004). In the next step the optimal update is determined using the corresponding penalty (penalties). Finally, the update is applied to yield $\beta^{[t]}$, which is the used in the next iteration.

The algorithm can be terminated in several ways. For the stagewise estimation framework in (1.2), a general approach is to stop the algorithm when the change in f falls below certain threshold. In the GEE setup, as there is no such loss function, the algorithm can be terminated when $|\langle U_{[0]}(\beta^{[t-1]}, \nu^{[t]}), \delta^{[t]} \rangle|$ is below certain threshold, e.g., 10^{-4}. The number of maximum iterations can also be roughly estimated, e.g., based on the ratio between the value of ϕ evaluated at a regular GEE solution (when available) and the step size ϵ. An advantage of the stagewise estimation is that if a given number of iterations is not adequate, the algorithm can be restarted from its last step.

2.3.6 An Illustration

To illustrate the efficacy of BiSEE and HiSEE, we consider a Poisson regression with a simulated data set of 50 clusters of size 4. There are 7 covariate groups of size 3. Only the first three groups have non-zero regression coefficients: $\beta_{\mathcal{I}_1} = (0, 0, .2)^\top$, $\beta_{\mathcal{I}_2} = (-.25, 0, .15)^\top$, and $\beta_{\mathcal{I}_3} = (.2, .15, -.15)^\top$. The covariates within each group are correlated. See Section 2.4 for the details of the data generation process.

The solution paths of β generated by lasso, group lasso, BiSEE ($\lambda_1 = \lambda_2 = 0.5$), and HiSEE are presented in Figure 1. Both lasso and group lasso bring in unimportant covariates before all the important ones enter the model. As a consequence, neither of them produces the correct model structure on their paths. In contrast, both BiSEE

Algorithm 1 Bi-level Stagewise Estimating Equations (BiSEE)

Initialize: $\beta^{[0]} = 0$, $\nu^{[0]}$, $w_j > 0$, $\lambda_1 > 0$, $\lambda_2 > 0$, and $\epsilon > 0$.
  for $t = 1, 2, \ldots$ **do**
    $(t.1)$ Given $\beta^{[t-1]}$, update the nuisance parameters to obtain $\nu^{[t]}$.
    $(t.2)$ Solve $\|B_{\mathcal{I}_j}(\gamma_j)\|_2 = 1$ to obtain γ_j and let $k = \underset{j:j\in\{1,\ldots,J\}}{\arg\max}\ \gamma_j$.
    $(t.3)$ $\beta^{[t]}_{\mathcal{I}_k} = \beta^{[t-1]}_{\mathcal{I}_k} + \epsilon B_{\mathcal{I}_k}(\gamma_k)/(\lambda_1 w_k + \lambda_2 \|B_{\mathcal{I}_k}(\gamma_k)\|_1)$.
  end for

Algorithm 2 Hierarchical Stagewise Estimating Equations (HiSEE)

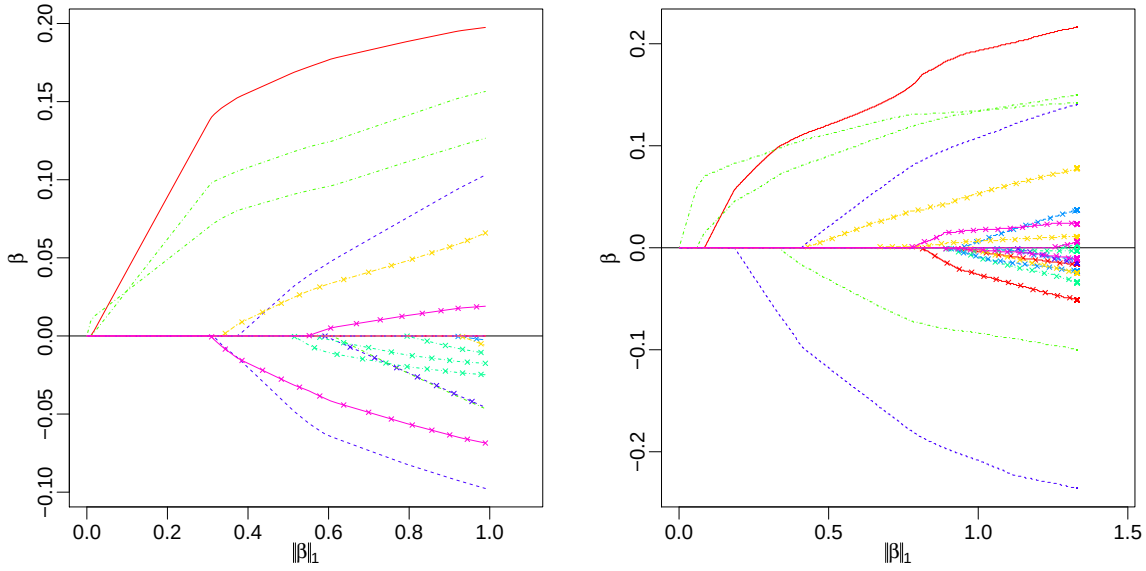
Initialize: $\beta^{[0]} = 0$, $\nu^{[0]}$, $w_j > 0$, and $\epsilon > 0$.
  for $t = 1, 2, \ldots$ **do**
    $(t.\text{a})$ Given $\beta^{[t-1]}$, update the nuisance parameters to obtain $\nu^{[t]}$.
    $(t.\text{b})$ $k = \underset{j:j\in\{1,\ldots,J\}}{\arg\max}\ \|U_{\mathcal{I}_j}(\beta^{[t-1]}; \nu^{[t]})\|_2 / w_j$.
    $(t.\text{c})$ $l_k = \underset{i:i\in\{1,\ldots,p_k\}}{\arg\max}\ |U_{\mathcal{I}_k(i)}(\beta^{[t-1]}; \nu^{[t]})|$.
    $(t.\text{d})$ $\beta^{[t]}_{\mathcal{I}_k(l_k)} = \beta^{[t-1]}_{\mathcal{I}_k(l_k)} + \epsilon\, \text{sign}\{U_{\mathcal{I}_k(l_k)}(\beta^{[t-1]}; \nu^{[t]})\}$.
  end for

and HiSEE are able to distinguish all the important variables from the irrelevant ones. The lasso and group lasso methods are outperformed by the proposed methods mainly because they fail to utilize the bi-level variable grouping structure. First, many of the groups contain only irrelevant predictors. Considering the predictors individually, as lasso does, is wasteful and creates a greater risk of false discovery. Second, the groups that do contain useful predictors may also contain irrelevant ones. However, group lasso can only include or exclude a group as a whole.

Figure 2 presents the corresponding paths of the mean squared prediction errors, based on 1000 replications. At first, all paths are comparable, but later on, the stagewise approaches, especially HiSEE, achieve lower mean predictive errors than lasso and group lasso.

2.4 Numerical Studies

We consider a longitudinal setting with cluster size $k_i = k = 4$ and covariates in groups of size $p_j = p_0 = 24$. Both low and high dimension settings are considered, with $(n, p, J) = (100, 72, 3)$ and $(n, p, J) = (50, 216, 9)$, respectively. Each set of p_0 covariates in the same group is generated from a multivariate normal distribution with mean zero and covariance matrix Σ_x of an exchangeable correlation structure, with off diagonal elements $\rho_x = 0.4$. Three patterns of group sparsity are investigated: (I) no sparsity, (II) moderate sparsity, and (III) high sparsity. In each setting, the number of important covariates is



(a) lasso (b) HiSEE

(c) Group lasso (d) BiSEE

Figure 1: The illustration example: paths of individual coefficient estimates against the ℓ_1 norm of the estimates, generated by lasso (a), HiSEE (b), group lasso (c), and BiSEE (d). All the paths are plotted against the ℓ_1 norm of the solution, e.g., $\|\widehat{\beta}\|_1$, along the path. Each grouped coefficients share the same line style. Paths of irrelevant predictors are marked with "x" and those of important predictors are left unmarked.

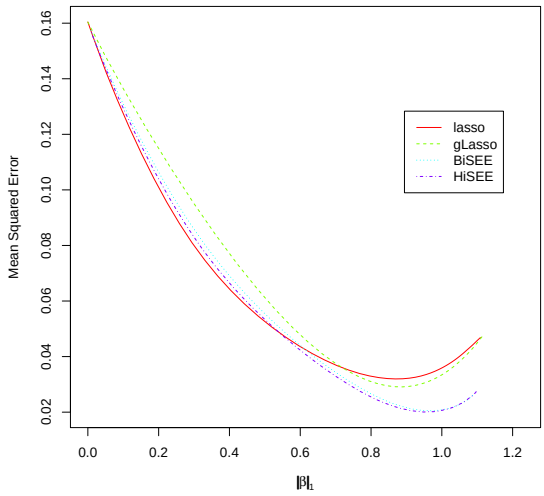


Figure 2: The illustration example: the path of mean prediction error as a function of the ℓ_1 norm of the coefficient estimates, generated by lasso, group lasso, HiSEE, and BiSEE, averaged over 1000 replicates.

always set to be $p_0 = 24$, with all the corresponding regression coefficients set to be 1; all other covariates have zero regression coefficients. In (I) the p_0 covariates in the first group are set as the important covariates, in (II) there are $p_0/2 = 12$ important covariates in each of the first two groups, and in (III), there are $p_0/3 = 8$ important covariates in each of the first three groups. Finally, each $k \times 1$ response vector for each cluster is generated from the multivariate normal distribution with mean $X_i\beta$ (where the intercept value is $\beta_0 = 1$) and covariance matrix Σ_y of an exchangeable correlation structure, i.e., Σ_y has diagonal elements σ_y^2 and off diagonal elements $\sigma_y^2\rho_y$. The variance σ_y^2 is chosen to fix the signal to noise ratio (SNR) to be 2, where SNR $= \sum_{j=1}^{J} \beta_{\mathcal{I}_j}^\top \Sigma_x \beta_{\mathcal{I}_j}/\sigma_y^2$. We consider $\rho_y \in \{.3, .6\}$, corresponding to moderate and high within-cluster correlations. Each configuration is replicated 100 times.

The methods to be compared with BiSEE and HiSEE include group lasso (gLasso) and sparse group lasso (sgLasso), which have been implemented in R package `SGL` (Simon et al., 2013a). In BiSEE and HiSEE, we use exchangeable working correlation, and set the number of iterations as $N = 2000$ and the step size as $\epsilon = 0.05$. For BiSEE and sgLasso, the tuning parameters are set to be $\lambda_2 = 1 - \lambda_1$, where $\lambda_1 \in \{0, 0.1, \ldots, 0.9, 1\}$. To compare all methods fairly, they are tuned based on independently generated testing data of large size. Specifically, the prediction error of a given solution on a given solution path/surface is defined as $\sum_{i=1}^{\tilde{n}} (\tilde{Y}_i - \hat{\mu}_i)^\top \mathcal{V}_i^{-1} (\tilde{Y}_i - \hat{\mu}_i)/\tilde{n}$, where $\{\tilde{Y}_i, i = 1, \ldots, \tilde{n} = 10n\}$ denotes the testing data, $\hat{\mu}_i$ denotes the prediction of $\tilde{Y}_i$ based on the fitted model, and $\mathcal{V}_i = \Sigma_y$. The solution with the lowest prediction error in each solution path/surface is then selected as the final solution for that path/surface. We also use the lowest prediction error as a predictive measure for comparing different methods. The prediction error of an oracle estimator, obtained by fitting GEEs with the true set of important covariates, is also computed. To evaluate the variable selection performance, we report both the false positive rate and the false negative rate. The false positive rate is the percent of true zero coefficients that were identified as non-zero. The false negative rate is the percent of true non-zero coefficients that were identified as zero.

Table 1 reports the simulation results when $\rho_y = 0.3$; results when $\rho_y = 0.6$ are provided in Table 2. Figure 3 shows the boxplots of the predictive measure. Across all simulation settings, BiSEE has the lowest predictive measure among all competing techniques. HiSEE's predictive performance is close to, or better than, that of either

Table 1: Simulation results with $\rho_y = 0.3$ from 100 replicates. Reported are the mean and standard deviations (sd) of the predictive measure (Msr), the false positive rate (FP), and the false negative rate (FN).

				No Sparsity			Mod. Sparsity			High Sparsity		
				Msr	FP	FN	Msr	FP	FN	Msr	FP	FN
Gaussian	High Dimension	Oracle	mean	1.15			1.15			1.14		
			sd	0.06			0.06			0.06		
		gLasso	mean	1.05	0.16	0.00	1.14	0.45	0.00	1.22	0.60	0.00
			sd	0.04	0.15	0.00	0.05	0.16	0.00	0.05	0.15	0.00
		sgLasso	mean	1.05	0.16	0.00	1.14	0.43	0.00	1.21	0.46	0.02
			sd	0.04	0.15	0.00	0.05	0.16	0.01	0.05	0.15	0.04
		BiSEE	mean	1.03	0.27	0.00	1.13	0.52	0.00	1.21	0.53	0.01
			sd	0.04	0.18	0.00	0.04	0.20	0.00	0.05	0.16	0.03
		HiSEE	mean	1.13	0.02	0.22	1.20	0.07	0.17	1.26	0.11	0.16
			sd	0.05	0.01	0.07	0.05	0.02	0.07	0.06	0.03	0.06
	Low Dimension	Oracle	mean	1.06			1.06			1.07		
			sd	0.03			0.03			0.03		
		gLasso	mean	1.03	0.21	0.00	1.08	0.88	0.00	1.14	1.00	0.00
			sd	0.03	0.29	0.00	0.03	0.22	0.00	0.03	0.00	0.00
		sgLasso	mean	1.03	0.21	0.00	1.08	0.72	0.01	1.12	0.50	0.02
			sd	0.03	0.29	0.00	0.03	0.26	0.02	0.04	0.17	0.02
		BiSEE	mean	1.01	0.53	0.00	1.07	0.83	0.00	1.11	0.69	0.01
			sd	0.03	0.34	0.00	0.03	0.23	0.01	0.03	0.28	0.02
		HiSEE	mean	1.06	0.04	0.08	1.09	0.23	0.04	1.12	0.39	0.02
			sd	0.03	0.04	0.05	0.03	0.08	0.04	0.04	0.10	0.03
Poisson	High Dimension	Oracle	mean	1.08			1.08			1.08		
			sd	0.05			0.05			0.05		
		lasso	mean	1.16	0.06	0.04	1.20	0.08	0.09	1.24	0.11	0.14
			sd	0.07	0.02	0.05	0.07	0.02	0.07	0.08	0.03	0.09
		gLasso	mean	1.18	0.41	0.00	1.27	0.54	0.00	1.36	0.63	0.00
			sd	0.08	0.15	0.00	0.09	0.13	0.00	0.09	0.13	0.00
		BiSEE	mean	1.04	0.17	0.00	1.12	0.26	0.00	1.16	0.32	0.00
			sd	0.05	0.11	0.00	0.06	0.13	0.02	0.06	0.15	0.01
		HiSEE	mean	1.12	0.01	0.04	1.15	0.03	0.09	1.19	0.04	0.13
			sd	0.06	0.01	0.05	0.06	0.01	0.07	0.07	0.02	0.09
	Low Dimension	Oracle	mean	1.04			1.03			1.03		
			sd	0.03			0.03			0.03		
		lasso	mean	1.06	0.32	0.00	1.06	0.37	0.01	1.07	0.46	0.01
			sd	0.03	0.14	0.02	0.04	0.10	0.02	0.03	0.11	0.02
		gLasso	mean	1.07	0.96	0.00	1.08	1.00	0.00	1.11	1.00	0.00
			sd	0.04	0.14	0.00	0.04	0.00	0.00	0.04	0.00	0.00
		BiSEE	mean	1.02	0.49	0.00	1.04	0.54	0.00	1.05	0.46	0.00
			sd	0.03	0.37	0.00	0.04	0.30	0.01	0.03	0.27	0.01
		HiSEE	mean	1.05	0.05	0.01	1.05	0.19	0.01	1.06	0.29	0.01
			sd	0.03	0.06	0.02	0.04	0.10	0.02	0.04	0.10	0.03

Table 2: High response correlation: simulation results with $\rho_y = 0.6$. Reported are the average values and standard deviations of the predictive measure (Msr), the false positive rate (FP), and the false negative rate (FN).

			No Sparsity			Mod. Sparsity			High Sparsity		
			Msr	FP	FN	Msr	FP	FN	Msr	FP	FN
Gaussian	High Dimension	Oracle	mean	1.16			1.16			1.15		
			sd	0.06			0.06			0.05		
		gLasso	mean	1.08	0.18	0.00	1.23	0.45	0.00	1.36	0.64	0.00
			sd	0.05	0.15	0.00	0.07	0.16	0.00	0.09	0.17	0.00
		sgLasso	mean	1.08	0.18	0.00	1.23	0.43	0.00	1.36	0.50	0.02
			sd	0.05	0.15	0.00	0.07	0.16	0.02	0.09	0.17	0.04
		BiSEE	mean	1.03	0.31	0.00	1.18	0.47	0.00	1.30	0.52	0.01
			sd	0.03	0.21	0.00	0.04	0.18	0.00	0.07	0.20	0.03
		HiSEE	mean	1.16	0.02	0.13	1.25	0.07	0.11	1.33	0.11	0.09
			sd	0.05	0.02	0.06	0.06	0.02	0.06	0.09	0.02	0.05
	Low Dimension	Oracle	mean	1.07			1.06			1.06		
			sd	0.03			0.03			0.03		
		gLasso	mean	1.05	0.17	0.00	1.14	0.88	0.00	1.22	1.00	0.00
			sd	0.03	0.29	0.00	0.04	0.22	0.00	0.04	0.00	0.00
		sgLasso	mean	1.05	0.17	0.00	1.14	0.72	0.01	1.19	0.50	0.02
			sd	0.03	0.29	0.00	0.04	0.25	0.02	0.05	0.18	0.03
		BiSEE	mean	1.01	0.53	0.00	1.09	0.83	0.00	1.12	0.47	0.00
			sd	0.02	0.33	0.00	0.03	0.28	0.01	0.04	0.20	0.01
		HiSEE	mean	1.08	0.05	0.04	1.11	0.27	0.01	1.12	0.40	0.00
			sd	0.03	0.04	0.04	0.04	0.10	0.02	0.04	0.09	0.01
Poisson	High Dimension	Oracle	mean	1.09			1.09			1.07		
			sd	0.07			0.05			0.07		
		lasso	mean	1.25	0.07	0.04	1.31	0.09	0.09	1.35	0.11	0.13
			sd	0.11	0.03	0.06	0.11	0.03	0.08	0.11	0.03	0.08
		gLasso	mean	1.29	0.39	0.00	1.42	0.52	0.00	1.53	0.61	0.00
			sd	0.12	0.12	0.00	0.14	0.09	0.00	0.14	0.12	0.00
		BiSEE	mean	1.05	0.19	0.00	1.17	0.29	0.00	1.21	0.29	0.00
			sd	0.06	0.14	0.00	0.06	0.16	0.01	0.08	0.14	0.01
		HiSEE	mean	1.15	0.01	0.02	1.19	0.03	0.04	1.22	0.05	0.07
			sd	0.09	0.01	0.04	0.07	0.02	0.05	0.09	0.02	0.07
	Low Dimension	Oracle	mean	1.04			1.03			1.04		
			sd	0.04			0.04			0.04		
		lasso	mean	1.09	0.32	0.00	1.10	0.38	0.01	1.11	0.45	0.01
			sd	0.06	0.14	0.01	0.05	0.12	0.02	0.06	0.11	0.02
		gLasso	mean	1.11	0.98	0.00	1.15	1.00	0.00	1.17	1.00	0.00
			sd	0.06	0.09	0.00	0.06	0.00	0.00	0.06	0.00	0.00
		BiSEE	mean	1.02	0.56	0.00	1.06	0.50	0.00	1.07	0.39	0.00
			sd	0.04	0.35	0.00	0.04	0.29	0.00	0.04	0.22	0.00
		HiSEE	mean	1.05	0.08	0.00	1.06	0.19	0.00	1.07	0.25	0.00
			sd	0.04	0.09	0.01	0.04	0.10	0.01	0.04	0.11	0.00

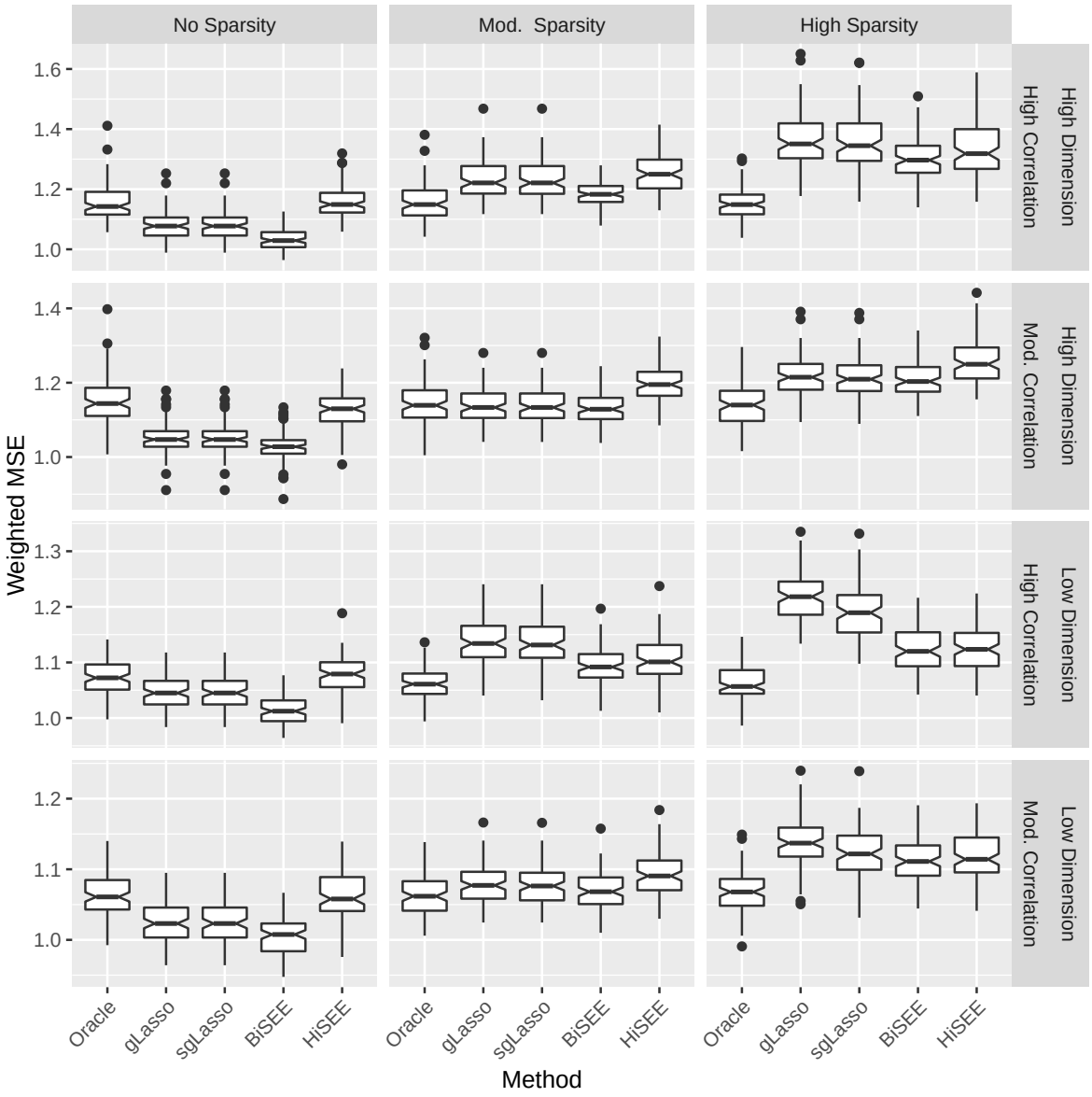


Figure 3: Gaussian example: boxplots of the predictive measures over 100 replicates.

group lasso or sparse group lasso. Also, the advantage of BiSEE and HiSEE becomes more visible when the response correlation increases. This is largely due to the fact that both BiSEE and HiSEE are based on GEEs, which accounts for within cluster dependence. In the no sparsity setting, we notice that the lowest predictive measure is not necessarily produced by the oracle estimator. Further investigation shows that this is because the stagewise estimators and the penalized estimators are all capable of inducing shrinkage estimation, which can be beneficial to deal with multi-collinearity. For variable selection, BiSEE yields comparable false positive rate compared to gLasso and sgLasso, and all of them have very low false negative rate. This is expected as it is known that convex penalization methods tend to select more variables when tuned based on predictive performance. HiSEE has the lowest false positive rate across all settings, at the expense of a higher false negative rate.

The study with Poisson response is similar to that with Gaussian response. The differences are described as follows. The within-cluster dependence of Poisson responses is set to be a normal copula with an exchangeable correlation structure whose off-diagonal values are $\rho_y \in \{.3, .6\}$. The marginal Poisson distributions are set to have mean $g^{-1}(X_{ij}\beta)$ for the jth observation in the ith cluster, where g is the log link function. The group size of the covariates is set to be $p_0 = 12$, and the model dimensions in low and high dimension settings are set as $(n, p, J) = (100, 36, 3)$ and $(n, p, J) = (50, 216, 18)$, respectively. The three group sparsity patterns are set in the same fashion as in the

Gaussian study, with the coefficients of $p_0 = 12$ important variables set to be 0.1 distributed in the first three groups. Since sgLasso implementation for Poisson regression is not available, we only consider lasso and gLasso in the comparison, using implementations from R package `grpreg` (Breheny and Huang, 2009). HiSEE and BiSEE use step size of $\epsilon = 0.025$. Each configuration is replicated 100 times. The performance of different methods is still compared in terms of prediction error, false positive rate, and false negative rate, as defined earlier. In the prediction error, however, $\mathcal{V}_i$, the variance matrix of $\tilde{Y}_i$, is approximated as $A_i^{1/2} R(\hat{\alpha}) A_i^{1/2}$, where $\hat{\alpha}$ is estimated from an independent data set of size $10n$.

The simulation results are summarized in Tables 1 and 2. Figure 4 presents the boxplots of the predictive measures of different methods. Most observations in the Gaussian case remain. BiSEE and HiSEE outperform both lasso and gLasso in prediction. The variable selection performance of BiSEE is in between those of lasso and gLasso. HiSEE in general yields the smallest false positive rates among all methods, with well-controlled false negative rates. The advantages of BiSEE and HiSEE become more visible as the within-cluster dependence increases.

2.4.1 Between Group Correlation

We conducted an additional simulation study to examine the effect of between group correlation. All of the same simulation settings described in Section 2.4 are used with correlation induced between groups in a manner similar to that which is described in

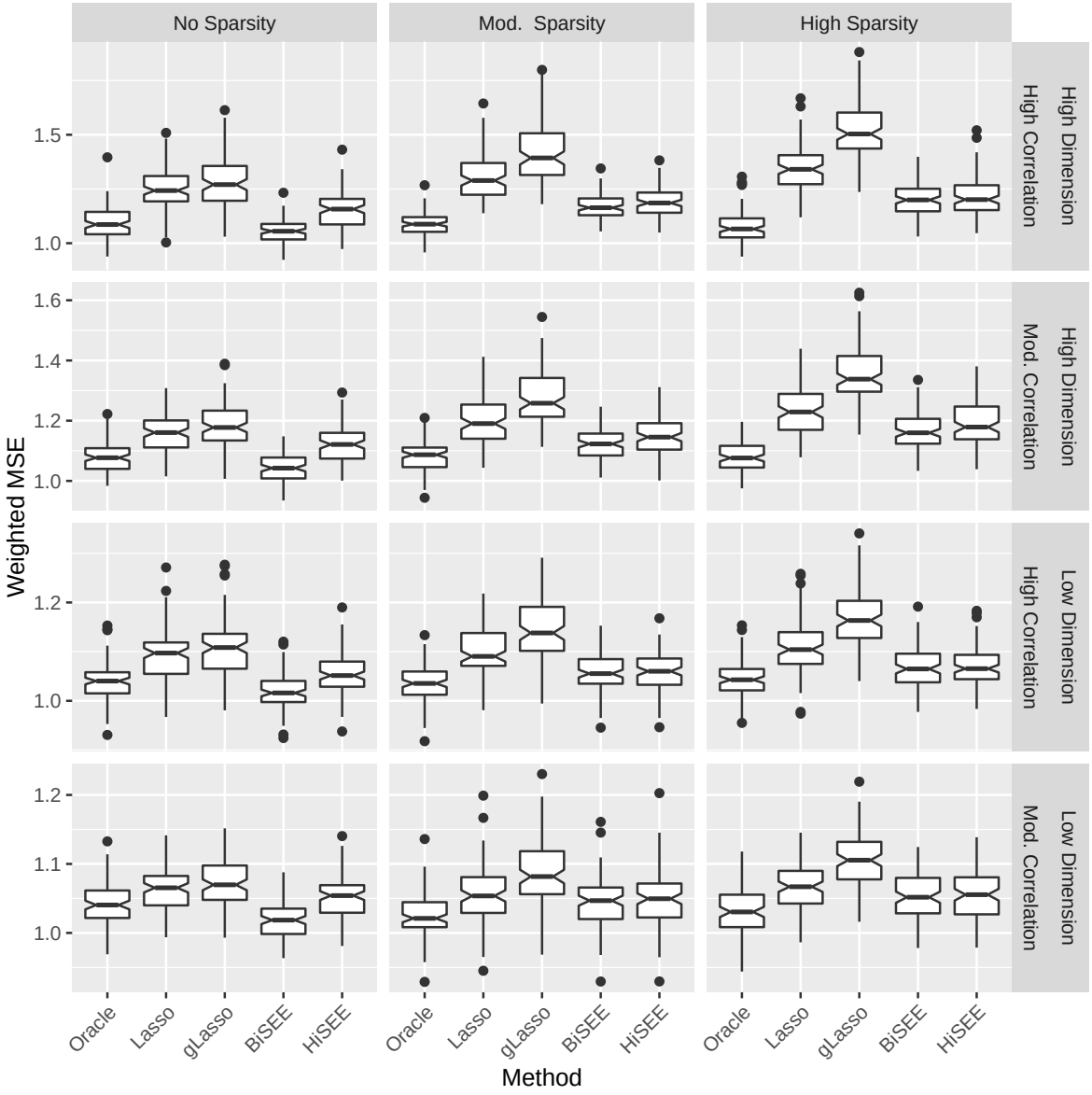


Figure 4: Poisson example: boxplots of the prediction errors over 100 replicates.

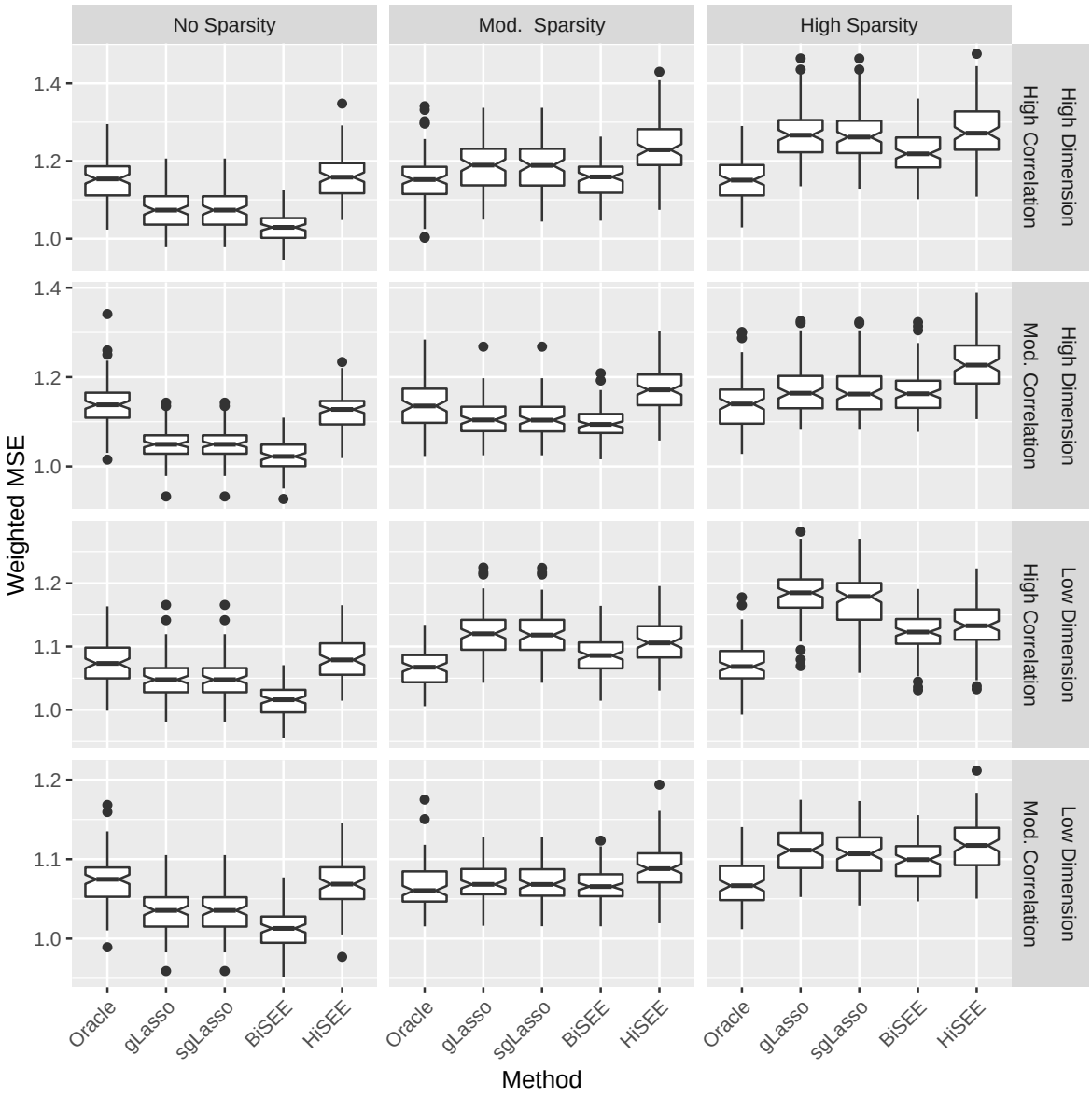


Figure 5: Gaussian example: boxplots of the predictive measures over 100 replicates, with a between different covariate correlation of $(.4)^{|i-j|+1}$, where i and j are group indices.

Huang et al. (2009) such that the pairwise correlation between two different covariates in groups i and j, respectively, is $(0.4)^{|i-j|+1}$. We present these results in Tables 3 and 4, with boxplots in Figures 5 and 6.

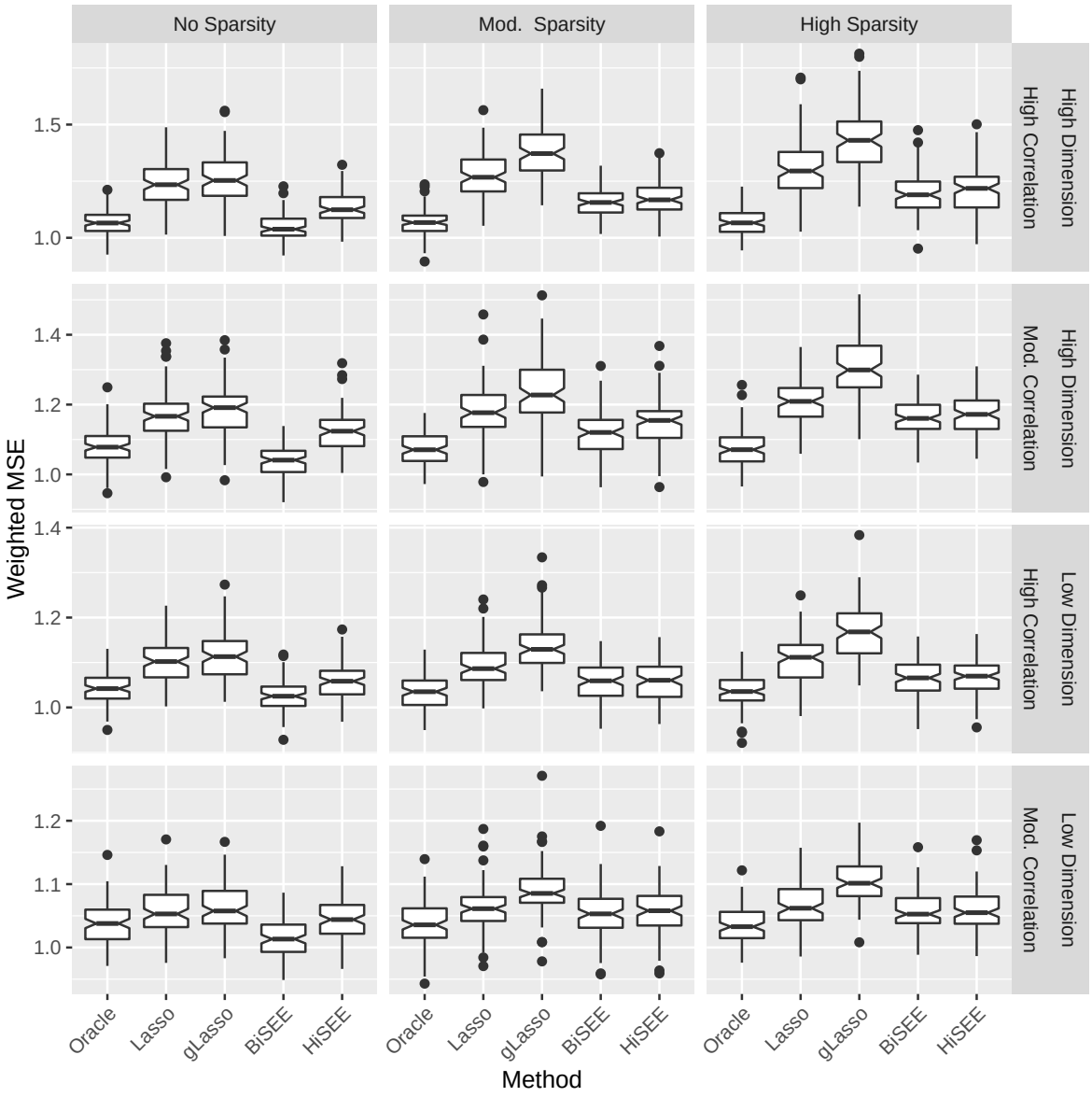


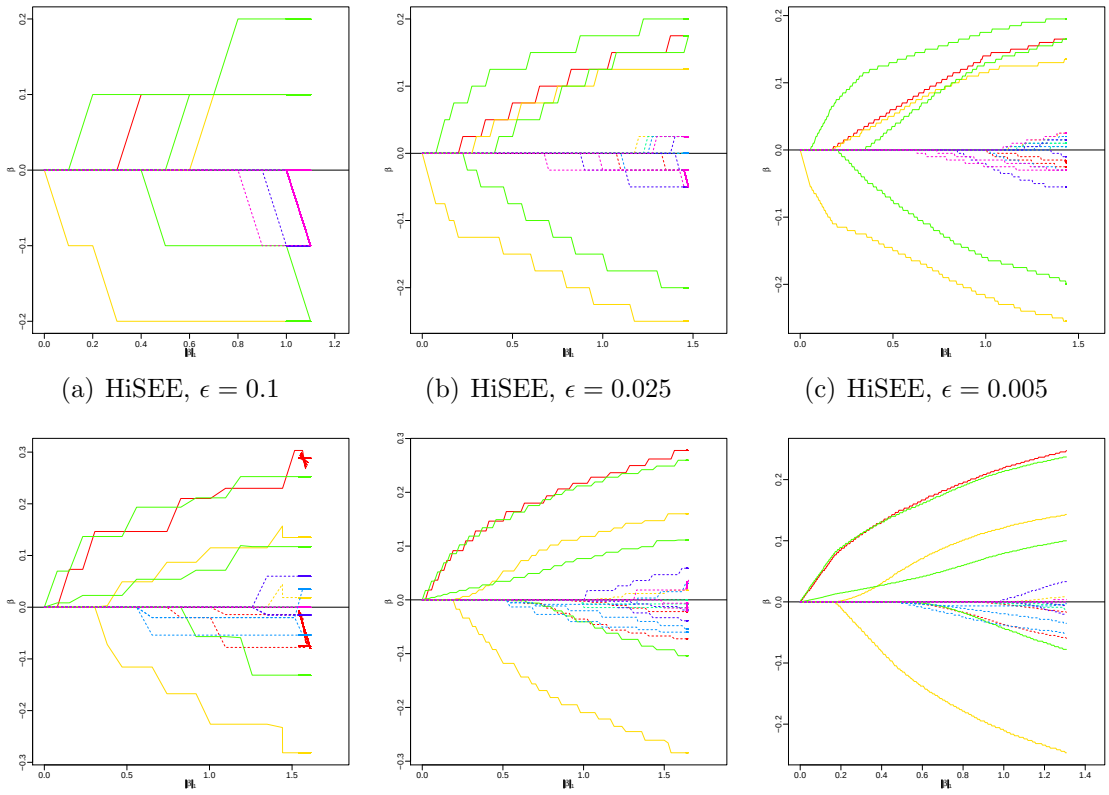
Figure 6: Poisson example: boxplots of the predictive measures over 100 replicates, with a between different covariate correlation of $(.4)^{|i-j|+1}$, where i and j are group indices.

Table 3: Moderate response correlation: simulation results with $\rho_y = 0.3$, and with a between different covariate correlation of $(.4)^{|i-j|+1}$, where i and j are group indices. Reported are the average values and standard deviations of the predictive measure (Msr), the false positive rate (FP), and the false negative rate (FN).

				No Sparsity			Mod. Sparsity			High Sparsity		
				Msr	FP	FN	Msr	FP	FN	Msr	FP	FN
Gaussian	High Dimension	Oracle	mean	1.14			1.14			1.14		
			sd	0.05			0.05			0.06		
		gLasso	mean	1.05	0.18	0.00	1.10	0.39	0.00	1.17	0.54	0.00
			sd	0.04	0.14	0.00	0.04	0.15	0.00	0.05	0.15	0.00
		sgLasso	mean	1.05	0.18	0.00	1.10	0.38	0.00	1.17	0.47	0.01
			sd	0.04	0.14	0.00	0.04	0.16	0.02	0.06	0.15	0.03
		BiSEE	mean	1.03	0.34	0.00	1.10	0.48	0.00	1.17	0.57	0.00
			sd	0.03	0.21	0.00	0.04	0.18	0.00	0.05	0.18	0.01
		HiSEE	mean	1.12	0.02	0.21	1.17	0.07	0.23	1.23	0.10	0.25
			sd	0.04	0.02	0.07	0.05	0.02	0.07	0.06	0.02	0.07
	Low Dimension	Oracle	mean	1.07			1.07			1.07		
			sd	0.03			0.03			0.03		
		gLasso	mean	1.03	0.28	0.00	1.07	0.79	0.00	1.11	1.00	0.00
			sd	0.03	0.34	0.00	0.02	0.25	0.00	0.03	0.00	0.00
		sgLasso	mean	1.03	0.28	0.00	1.07	0.70	0.01	1.11	0.68	0.03
			sd	0.03	0.34	0.00	0.02	0.25	0.02	0.03	0.22	0.04
		BiSEE	mean	1.01	0.60	0.00	1.07	0.86	0.00	1.10	0.87	0.01
			sd	0.02	0.35	0.00	0.02	0.20	0.00	0.03	0.23	0.03
		HiSEE	mean	1.07	0.05	0.09	1.09	0.25	0.08	1.12	0.40	0.07
			sd	0.03	0.04	0.05	0.03	0.08	0.05	0.03	0.08	0.05
Poisson	High Dimension	Oracle	mean	1.08			1.07			1.07		
			sd	0.05			0.05			0.05		
		Lasso	mean	1.17	0.06	0.05	1.18	0.08	0.08	1.21	0.10	0.10
			sd	0.07	0.02	0.06	0.07	0.03	0.07	0.07	0.03	0.08
		gLasso	mean	1.18	0.41	0.00	1.24	0.48	0.00	1.31	0.56	0.00
			sd	0.08	0.13	0.00	0.09	0.13	0.00	0.09	0.13	0.00
		BiSEE	mean	1.04	0.16	0.00	1.12	0.23	0.00	1.16	0.25	0.01
			sd	0.05	0.11	0.00	0.06	0.13	0.01	0.06	0.12	0.04
		HiSEE	mean	1.12	0.01	0.05	1.15	0.03	0.10	1.17	0.04	0.13
			sd	0.06	0.01	0.06	0.07	0.02	0.07	0.06	0.02	0.09
	Low Dimension	Oracle	mean	1.04			1.04			1.04		
			sd	0.03			0.03			0.03		
		Lasso	mean	1.06	0.33	0.00	1.06	0.39	0.01	1.07	0.42	0.01
			sd	0.04	0.14	0.02	0.04	0.13	0.02	0.03	0.11	0.03
		gLasso	mean	1.06	0.95	0.00	1.09	1.00	0.00	1.11	1.00	0.00
			sd	0.04	0.17	0.00	0.04	0.00	0.00	0.04	0.00	0.00
		BiSEE	mean	1.01	0.54	0.00	1.05	0.57	0.00	1.06	0.44	0.01
			sd	0.03	0.34	0.00	0.04	0.33	0.01	0.03	0.24	0.02
		HiSEE	mean	1.04	0.07	0.00	1.06	0.22	0.01	1.06	0.28	0.01
			sd	0.03	0.08	0.01	0.04	0.12	0.02	0.03	0.12	0.03

Table 4: High response correlation: simulation results with $\rho_y = 0.6$, and with a between different covariate correlation of $(.4)^{|i-j|+1}$, where i and j are group indices. Reported are the average values and standard deviations of the predictive measure (Msr), the false positive rate (FP), and the false negative rate (FN).

				No Sparsity			Mod. Sparsity			High Sparsity		
				Msr	FP	FN	Msr	FP	FN	Msr	FP	FN
Gaussian	High Dimension	Oracle	mean	1.15			1.16			1.15		
			sd	0.06			0.06			0.06		
		gLasso	mean	1.08	0.15	0.00	1.19	0.37	0.00	1.27	0.54	0.00
			sd	0.05	0.14	0.00	0.06	0.16	0.00	0.07	0.15	0.00
		sgLasso	mean	1.08	0.15	0.00	1.19	0.36	0.00	1.27	0.46	0.02
			sd	0.05	0.14	0.00	0.06	0.15	0.01	0.07	0.13	0.04
		BiSEE	mean	1.03	0.27	0.00	1.15	0.42	0.00	1.22	0.48	0.01
			sd	0.04	0.18	0.00	0.05	0.17	0.01	0.05	0.17	0.03
		HiSEE	mean	1.16	0.02	0.14	1.23	0.07	0.16	1.28	0.11	0.15
			sd	0.06	0.01	0.07	0.07	0.02	0.07	0.07	0.03	0.07
	Low Dimension	Oracle	mean	1.07			1.07			1.07		
			sd	0.03			0.03			0.03		
		gLasso	mean	1.05	0.23	0.00	1.12	0.81	0.00	1.18	1.00	0.00
			sd	0.03	0.30	0.00	0.04	0.25	0.00	0.04	0.00	0.00
		sgLasso	mean	1.05	0.23	0.00	1.12	0.71	0.01	1.17	0.64	0.03
			sd	0.03	0.30	0.00	0.03	0.26	0.02	0.04	0.21	0.03
		BiSEE	mean	1.01	0.60	0.00	1.09	0.87	0.00	1.12	0.69	0.01
			sd	0.03	0.34	0.00	0.03	0.23	0.01	0.04	0.30	0.02
		HiSEE	mean	1.08	0.06	0.04	1.11	0.26	0.04	1.13	0.42	0.03
			sd	0.03	0.04	0.03	0.03	0.08	0.04	0.04	0.11	0.03
Poisson	High Dimension	Oracle	mean	1.07			1.07			1.07		
			sd	0.06			0.06			0.06		
		Lasso	mean	1.24	0.06	0.04	1.28	0.08	0.07	1.31	0.09	0.11
			sd	0.10	0.02	0.05	0.11	0.02	0.07	0.13	0.03	0.08
		gLasso	mean	1.26	0.40	0.00	1.38	0.47	0.00	1.44	0.53	0.00
			sd	0.11	0.13	0.00	0.12	0.11	0.00	0.15	0.14	0.00
		BiSEE	mean	1.04	0.20	0.00	1.16	0.26	0.00	1.20	0.24	0.01
			sd	0.06	0.13	0.00	0.07	0.17	0.01	0.09	0.15	0.02
		HiSEE	mean	1.13	0.01	0.01	1.17	0.03	0.04	1.21	0.04	0.06
			sd	0.07	0.01	0.03	0.08	0.01	0.05	0.10	0.02	0.07
	Low Dimension	Oracle	mean	1.04			1.04			1.04		
			sd	0.04			0.04			0.04		
		Lasso	mean	1.10	0.32	0.01	1.09	0.37	0.01	1.10	0.42	0.01
			sd	0.05	0.12	0.02	0.05	0.11	0.02	0.05	0.11	0.03
		gLasso	mean	1.12	0.98	0.00	1.13	1.00	0.00	1.17	1.00	0.00
			sd	0.05	0.10	0.00	0.05	0.00	0.00	0.06	0.00	0.00
		BiSEE	mean	1.03	0.56	0.00	1.06	0.46	0.00	1.07	0.41	0.00
			sd	0.03	0.34	0.00	0.04	0.30	0.00	0.04	0.24	0.00
		HiSEE	mean	1.06	0.05	0.00	1.06	0.20	0.00	1.07	0.26	0.00
			sd	0.04	0.05	0.00	0.04	0.12	0.00	0.04	0.12	0.01



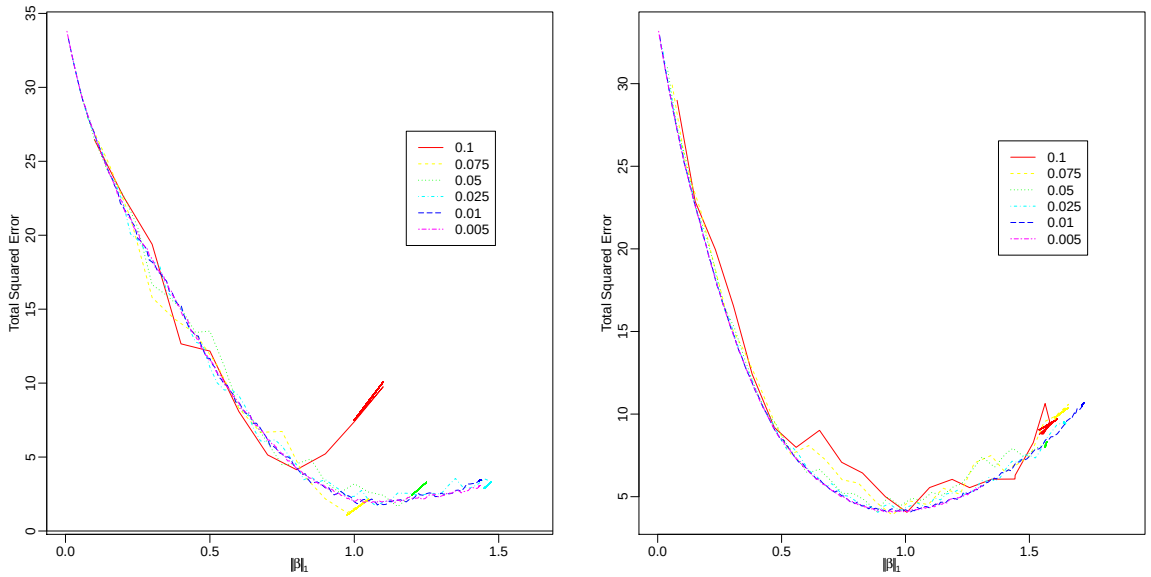
(a) HiSEE, $\epsilon = 0.1$ (b) HiSEE, $\epsilon = 0.025$ (c) HiSEE, $\epsilon = 0.005$

(d) BiSEE, $\epsilon = 0.1$ (e) BiSEE, $\epsilon = 0.025$ (f) BiSEE, $\epsilon = 0.005$

Figure 7: Coefficient Trace plots: paths of individual coefficient estimates against the ℓ_1 norm of the estimates, generated by HiSEE (a) – (c) and BiSEE (d) – (f). Each grouped coefficients share the same line style and color. Paths of dashed lines represent irrelevant predictors and those of solid lines represent important predictors.

2.4.2 Sensitivity Analysis on Step Size

The choice of step size ϵ is not arbitrary but should be done with care. In our numerical studies, we made sure that the step size was small enough to ensure a sufficiently smooth solution path. For Gaussian responses, we used $\epsilon = 0.05$. For Poisson responses, since the coefficient values were smaller in magnitude, we used $\epsilon = 0.025$. In the real data example, we used $\epsilon = 0.001$ to ensure precision, though larger values would have been



(a) HiSEE (b) BiSEE

Figure 8: Predictive Error: the path of prediction error as a function of the ℓ_1 norm of the coefficient estimates, generated by HiSEE (a) and BiSEE (b) using different values of ϵ.

sufficient.

The determination of the step size is an important issue. If the step size is too large, the stagewise estimation may produce inaccurate and unstable path; on the other hand, a too-small step size would cause unnecessary computational effort. In practice, we suggest the step size be checked based on some diagnostics plots. One can either examine the coefficient trace plot, i.e., the coefficient paths against the ℓ_1 norm of the coefficients (or the iteration number), or the prediction trace plot, i.e., the predictive error (by cross validation) against ℓ_1 norm of the coefficients. When the step size is too large, we will see jagged, "fluttering" paths that bounce back and forth between similar points.

To show the impact of step sizes on BiSEE and HiSEE, we use our illustrative example from Section 2.3.6. Web Figure 7 shows the coefficient trace plots for $\epsilon \in \{0.1, 0.025, 0.005\}$, and Web Figure 8 shows the prediction trace plots for

$$\epsilon \in \{0.1, 0.075, 0.05, 0.025, 0.01, 0.005\}.$$

Web Figure 7 shows that the smaller the step size, the smother the solution path. Too large step sizes may lead to "fluttering" paths. Web Figure 8 confirms that as long as the step size is below certain value, the paths become stable and similar to each other.

2.5 Connecticut Adolescent Suicide Risk Study

Suicide among youth is one of the most serious public health problems in the United States (e.g., Chen and Aseltine, 2017). According to the Centers for Disease Control (CDC), the suicide rate in 2013 was 11.1 per $100,000$ among youth aged 15–24, making it the third leading cause of death of this age group. Suicide among adolescents aged 15–19 years tripled between 1950 and 2011, exhibiting an alarming increasing trend. Effectively preventing youth suicide is an extremely challenging task which requires development of reliable metrics for assessing suicide risk, identification of areas of greater risk for effective resource allocation, and thorough understanding of important risk factors and their interactions.

We used five years of hospitalization data from 2010–2014 from the Connecticut Hospital Inpatient Discharge Database, to gain insight about the association between the suicide risk of the 15–19 age group at the school district level and the socio-economic, demographic, and academic characteristics of the schools districts. Annual counts of inpatient hospitalizations due to intentional self-injury for the 15–19 age group were obtained for each of 119 public schools districts with high schools in Connecticut. Although a suicide attempt does not always lead to a hospitalization, the suicide-related hospitalization rate, i.e., the ratio between the suicide-related hospitalization count and the population size of the 15–19 age group, can serve as a valid proxy for indicating the relative risk levels of the school districts. For comparison, the annual total counts of inpatient hospitalizations of the 15–19 age group were also obtained, which can be used as a proxy for assessing the overall well-being of the youth in different school districts. One school district (Thomaston School District) was removed from the analysis for having missing values, leaving a final sample size of $n = 118$.

Several variables about the characteristics of the school districts were collected, which fall into multiple categories: 1) demographic measures, including average household size, proportion of population under the age of 18, and the proportion of population that is white; 2) academic measures, including average score on the Connecticut Academic Performance Test (CAPT) and the average attendance rate of the high schools in the district; 3) incidence measures, including the incidence rate, defined as the ratio between the number of incidences and the total enrollment; 4) prosperity measures, including the

median income of the district; 5) grant status, indicating whether a school district has ever received a state grant related to suicide prevention. As the covariates (excluding grant status) may have nonlinear effects on the hospitalization counts, we computed the orthogonal polynomials of orders 1 and 2 of each covariate, and included both terms in its corresponding variable group. We also set the linear and the quadratic terms of time as another variable group, to capture the trend over time. As such, there were 6 variables groups, with their group sizes ranging from 2 to 6.

We then used BiSEE and HiSEE to conduct a Poisson regression analysis, in which the annual hospitalization counts served as the clustered response, the logarithm of the district population of the 15–19 age group as the offset, and the aforementioned groups of variables as the predictors. The step size was set to $\epsilon = 0.001$. The model selection was conducted by 10-fold cross-validation, using the deviance to measure the predictive performance of the models. As a result, while BiSEE and HiSEE resulted in similar solution paths, BiSEE produced the best model for modeling the overall hospitalization counts, and HiSEE produced the best model for modeling the suicide-related hospitalization counts. (To ease the interpretation, if the quadratic term of a variable was selected, its linear term was then also included.) We then refitted the data using regular GEEs with the selected model structures. For comparison, the group lasso approach implemented in the R package `grplasso`, which does not account for correlation within the school districts, was also applied in the same manner.

Table 5 reports the refitting results using GEE. The median income, the proportion of

Table 5: Suicide study: the fitted Poisson regression models for the overall hospitalization counts and the suicide-related hospitalization counts. BiSEE, HiSEE, and group lasso were used for model selection, and the estimation results were from refitted models using GEE. Between BiSEE and HiSEE, BiSEE produced the best model for the overall hospitalizations, whereas HiSEE produced the best model for the suicide-related hospitalizations. Models selected using group lasso are presented in the gLasso columns. The linear terms are presented with superscript ^{1}, and the quadratic terms are presented with ^{2}.

	Overall		Suicidal	
	BiSEE	gLasso	HiSEE	gLasso
(Intercept)	−3.486	−3.466	−6.686	−6.694
Demographic measures				
Proportion of white^1	−0.072			
Prosperity measures				
Median income^1	−0.174	−0.228	−0.176	
Median income^2		0.068		
Academic measures				
Average CAPT^1	−0.033		0.122	−0.057
Average CAPT^2			−0.042	−0.137
Attendance rate^1			0.003	−0.026
Attendance rate^2			−0.105	−0.060
Incidence measures				
Incidence rate^1			0.055	
Incidence rate^2			−0.087	
Grant status				
Grant			−0.062	
Time trend				
Time^1	−0.089		0.061	
Time^2	−0.023			

white and the average CAPT are all negatively associated with the overall hospitalization rate, indicating that the general well-being of the 15-19 age group tends to be better in wealthier communities with a larger white population and better academic performance. Also, there is a downward trend of the overall hospitalization rate over time, indicating the general well-being of the youth is improving over time. These results are as expected. The results from group lasso only included the prosperity measure group (linear and quadratic); academic measure and time trend were not selected as whole groups.

In contrast, the analysis of the suicidal hospitalization counts shows a different picture. The median income level is still negatively associated with the suicidal risk. However, while the proportion of the population that is White is negatively associated with the overall hospitalization, it is no longer selected for modeling the suicide-related hospitalization; this may be because the Whites have a much higher suicide attempt rate than all other race groups, which may offset the negative effect seen in the analysis of overall hospitalization. The incidence rate appears to have a positive effect on the suicide risk and the association may be nonlinear. Also, after adjusting for the other terms, academic performance, as measured by average CAPT and attendance rate, appears to have a positive linear effect and a negative quadratic effect on the suicide risk. This suggests that after accounting for the income level and other terms in the model, the better the academic performance, the higher the suicide attempt risk, and this positive association diminishes as academic performance continues to improve. This conditional positive association may be attributed to the fact that students in school districts of

better academic performance also tend to be under higher pressure, which may induce psychological distress. Having received a suicide prevention grant is negatively associated with the suicide risk. While the overall hospitalization rate has been decreasing, there appears to be increases in hospitalizations for suicide attempts over time. These findings agree with existing studies (Chen and Aseltine, 2017). On the other hand, the results from group lasso only included the academic measure group; the signs of the linear terms of average CAPT and attendance rate are opposite to those from the HiSEE approach; this is mainly caused by not adjusting the effects from other important factors including median income, incidence measures, grant status, and time trend. Overall, incorporating variable hierarchy and data correlation via the proposed approach allows us to gain several insights regarding adolescent suicide risk, which can be very useful in guiding future suicide monitoring and prevention efforts.

Chapter 3

Efficient Interaction Selection via Stagewise Generalized Estimating Equations

3.1 Interaction Selection

In regression analysis, interaction terms can potentially lead to a more precise and insightful model reflecting the underlying relationships being studied. Model selection with interaction terms is, however, difficult as the dimensionality of the problem is exponential in the number of covariates. Furthermore, it is necessary to maintain certain hierarchical structure in the model where higher order terms are included only if lower order terms are also included to make interpretation straightforward.

We propose two techniques to facilitate model selection with non-Gaussian clustered data while preserving interaction hierarchy. Our techniques make use of the general framework of generalized estimating equations (GEE) (Liang and Zeger, 1986), that

allows for flexible marginal modeling for clustered data without fully specifying the within-cluster dependence structure. Building upon previous stagewise techniques for GEEs (Wolfson, 2011; Vaughan et al., 2017), we develop a stagewise counterpart to the hierarchical lasso proposed by Bien et al. (2013) called hierarchical lasso (HiLa) stagewise estimating equations that makes use of the same penalty function in a stagewise context. Additionally, we develop a stagewise active set (ACTS) technique that extends the work by Zhu et al. (2014) to GEEs.

3.2 Notation

Let Y_i be a $k_i \times 1$ response vector in cluster i with cluster size k_i, $i = 1, ..., n$. Let $X_i = (X_{i[11]}, \ldots, X_{i[pp]}, X_{i[12]}, \ldots, X_{i[1p]}, X_{i[23]}, \ldots, X_{i[(p-1)p]})$, where $X_{i[jk]}$ is a k_i covariate vector for Y_i where $j = k$ corresponds to a main effect, and $j \neq k$ corresponds to an interaction effect. That is, $X_{i[jk]} = X_{i[jj]}X_{i[kk]}$ for $j \neq k$. The conditional mean of Y_i given X_i is specified as $E[Y_i \mid X_i] = \mu_i = g^{-1}(\eta_i)$, where $\eta_i = \beta_0 \mathbf{1}_{k_i} + X_i\beta$, β is a $(p^2+p)/2 \times 1$ coefficient vector, β_0 is the scalar intercept, $\mathbf{1}_{k_i}$ is a $k_i \times 1$ vector of 1's, and g is a known link function. The regression coefficient vector $\beta = (\beta_{11}, \ldots, \beta_{pp}, \beta_{12}, \ldots, \beta_{(p-1)p})^\top$, where β_{jj} corresponds to the jth main effect and β_{jk} with $j \neq k$ corresponds to the interaction effect between the jth and kth main effects, is of primary interest. The conditional variance of each component Y_{ij} of Y_i, $j = 1, \ldots, k_i$, is $V[Y_{ij} \mid X_{ij}] = \psi v(\mu_{ij})$, where X_{ij} is the jth row of X_i, ψ is a scalar, and $v(\cdot)$ is a variance function as in the exponential families.

The correlation structure is approximated by $R_i(\alpha)$, a $k_i \times k_i$ working correlation matrix parameterized as a function of a parameter vector α.

The regression coefficients (β_0, β) given (ψ, α) are estimated by solving

$$U(\beta, \beta_0, \psi, \alpha) \equiv -\sum_{i=1}^{n} D_i^\top V_i^{-1}(Y_i - \mu_i) = 0,$$

where $D_i = (\partial\mu_i/\partial\eta_i^\top)X_i$, $V_i = A_i^{1/2}R_i(\alpha)A_i^{1/2}$, and $A_i = \psi \operatorname{diag}\{v(\mu_{i1}), \ldots, v(\mu_{ik_i})\}$. A major advantage of GEE is that the consistency of the estimator of β is not affected by misspecification of the correlation structure of the clusters (Liang and Zeger, 1986). Given (β_0, β), estimates of (ψ, α) can be obtained by method of moments (Liang and Zeger, 1986). The alternating updating continues until convergence.

For notational simplicity, we write $U(\beta, \beta_0, \psi, \alpha) = U(\beta, \nu)$ from here forward, where $\nu = (\beta_0, \psi, \alpha^\top)^\top$ is a vector of the nuisance parameters that are not considered for variable selection. The estimating equations $U(\beta, \nu)$ can also be partitioned based on the interaction structure, i.e., $U(\beta, \nu) = (U_0(\beta, \nu), U_{11}(\beta, \nu)^\top, \ldots, U_{pp}(\beta, \nu)^\top, U_{12}(\beta, \nu)^\top, \ldots, U_{(p-1)p}(\beta, \nu)^\top)^\top$ where $U_{jk}(\beta, \nu) \in \mathbb{R}$ pertains to the jth main effect if $j = k$, the interaction effect between the jth and kth main effects if $j \neq k$, and $U_0(\beta, \nu) \in \mathbb{R}$ pertains to the intercept term. To ease notation, we may interchange the subscripts of interaction terms to refer to the same object; i.e. $\beta_{jk} = \beta_{kj}$ and $U_{jk} = U_{kj}$.

3.3 Stagewise GEE for Interaction Selection

3.3.1 HiLa Stagewise Estimating Equations

Bien et al. (2013) introduced a penalization approach to model selection with interaction terms that preserves the desired strong hierarchical structure. First, a convex relaxation of the constrained optimization that adds the restriction requiring

$$\|(\beta_{j1}, \ldots, \beta_{j(j-1)}, \beta_{j(j+1)}, \ldots, \beta_{jp})\|_1 \leq |\beta_{jj}|$$

to the original lasso problem is introduced. This convex relaxation is done by splitting the main effects β_{jj} into two parts. That is, β_{jj} is re-written as $\beta_{jj} = \beta_{jj}^+ - \beta_{jj}^-$, where β_{jj}^+ and β_{jj}^- are both non-negative. This is different from a usual splitting into "positive" and "negative" terms as $\beta_{jj}^{\pm} \neq \max(\pm\beta_{jj}, 0)$ necessarily; both terms may be positive. β_{jj}^+ and β_{jj}^- are each the jth element of the $p \times 1$ vectors β^+ and β^-, respectively. Due to the parity between U and β, we also have U_{jj}^+ and U_{jj}^- for all main effects j, where $U_{jj} = U_{jj}^+ = -U_{jj}^-$. The resulting constrained optimization problem is represented

as

$$\begin{aligned} &\underset{\beta^{\pm} \in \mathbb{R}^p,\, \beta_{\mathcal{I}} \in \mathbb{R}^{p(p-1)/2}}{\text{minimize}} && L(\beta^+ - \beta^-, \beta_{\mathcal{I}}) \\ &\text{subject to} \\ &\mathbf{1}_p^\top(\beta^+ + \beta^-) + \|\beta_{\mathcal{I}}\|_1 \le \lambda, \\ &\|(\beta_{j1}, \ldots, \beta_{j(j-1)}, \beta_{j(j+1)}, \ldots, \beta_{jp})\|_1 \le \beta_j^+ + \beta_j^-, && j = 1, \ldots, p, \\ &\beta_j^+ \ge 0, && j = 1, \ldots, p, \\ &\beta_j^- \ge 0, && j = 1, \ldots, p, \end{aligned} \tag{3.1}$$

where L is a convex loss function, $\|\cdot\|_1$ indicates the ℓ_1 norm, $\beta_{\mathcal{I}}$ indicates the $p(p-1)/2 \times 1$ sub-vector of interaction terms of the original β vector, and λ is a tuning parameter.

We propose the following algorithm as a stagewise analog to the hierarchical lasso.

(t.1) Given $\beta^{[t-1]}$, update the nuisance parameters to obtain $\nu^{[t]}$.

(t.2) Solve for

$$\begin{aligned} \delta^{[t]} = &\underset{\delta^{\pm} \in \mathbb{R}^p_+,\, \delta_{\mathcal{I}} \in \mathbb{R}^{p(p-1)/2}}{\arg\min} && \langle U_{[0]}(\beta_0^{[t-1]}, \beta^{+[t-1]} - \beta^{-[t-1]}, \beta_{\mathcal{I}}^{[t-1]}, \nu^{[t]}), \delta \rangle \\ &\text{subject to} \\ &\mathbf{1}_p^\top(\delta^+ + \delta^-) + \|\delta_{\mathcal{I}}\|_1 \le \epsilon, \\ &\|(\delta_{j1}, \ldots, \delta_{jj-1}, \delta_{jj+1}, \ldots, \delta_{jp})\|_1 \le \delta_j^+ + \delta_j^-, && j = 1, \ldots, p, \\ &\delta_j^+ \ge 0, && j = 1, \ldots, p, \\ &\delta_j^- \ge 0, && j = 1, \ldots, p, \end{aligned}$$

where δ is subset and decomposed in a manner similar to β; that is, $\delta = (\delta_{11}^{+} - \delta_{11}^{-}, \ldots, \delta_{pp}^{+} - \delta_{pp}^{-}, \delta_{12}, \ldots, \delta_{(p-1)p})^{\top}$.

(t.3) $\beta^{[t]} = \beta^{[t-1]} + \delta^{[t]}$.

In the general framework proposed by Tibshirani (2015), for a desired step size ϵ, $\phi(\delta) \leq \epsilon$ is used to constrain $\phi(\beta^{[t-1]} + \delta) - \phi(\beta^{[t-1]})$ to be less than ϵ, but this does not consider addition constraints as we have in Equation (3.1). If we consider the original form of the penalty, then what is desired is

$$\delta^{[t]} = \underset{\delta^{\pm} \in \mathbb{R}^{p}_{+}, \delta_{\mathcal{I}} \in \mathbb{R}^{p(p-1)/2}}{\arg\min} \langle U_{[0]}(\beta_0^{[t-1]}, \beta^{+[t-1]} - \beta^{-[t-1]}, \beta_{\mathcal{I}}^{[t-1]}, \nu^{[t]}), \delta \rangle$$

subject to

$$\mathbf{1}_p^{\top}(\theta^{+} + \theta^{-}) + \|\theta_{\mathcal{I}}\|_1 - \mathbf{1}_p^{\top}(\beta^{+[t-1]} + \beta^{-[t-1]}) - \|\beta_{\mathcal{I}}^{[t-1]}\|_1 \leq \epsilon,$$

$$\|(\theta_{(jj)})\|_1 \leq \theta_j^{+} + \theta_j^{-}, \quad j = 1, \ldots, p,$$

$$\theta_j^{+} \geq 0, \qquad j = 1, \ldots, p,$$

$$\theta_j^{-} \geq 0, \qquad j = 1, \ldots, p,$$

where, $\theta = \beta^{[t-1]} + \delta$ and $(\theta_{(jj)}) = (\beta_{j1}^{[t-1]} + \delta_{j1}, \ldots, \beta_{jj-1}^{[t-1]} + \delta_{jj-1}, \beta_{jj+1}^{[t-1]} + \delta_{jj+1}, \ldots, \beta_{jp}^{[t-1]} + \delta_{jp})$. Satisfying these conditions in each iteration would ensure that $\beta^{[t]}$ satisfies the constraints for the hierarchical lasso in Equation (3.1), and therefore satisfies the strong hierarchy structure. However, the essence of stagewise estimation is simple computation, and determining the optimal δ for these conditions does not meet that standard.

To simplify the computation, we instead force δ itself to satisfy the hierarchical lasso conditions for strong hierarchy. This in turn guarantees that $\beta^{[t]}$ will satisfy the strong hierarchy conditions at each iteration. This tighter constraint on the updates may result in a slightly more constrained path, but the computation becomes much simpler.

One of the biggest benefits of stagewise techniques is their computational efficiency, so it is critical that the update determined in $(t.2)$ can be easily calculated. The following theorem shows that $(t.2)$ can be solved efficiently, while ensuring a strong hierarchy in each iteration.

Theorem 3.1. *The resulting update for* $(t.2)$ *is constructed as follows: First, identify*

$$(k,l) = \arg\max_{i \neq j} |U_{ij}| + |U_{ii}| + |U_{jj}|, \text{ and,}$$

$$z = \arg\max |U_{ii}|.$$

If $|U_{kl}| + |U_{kk}| + |U_{ll}| > 3|U_{zz}|$*, then the update will have the form*

$$\begin{aligned}
\delta_{kl}^{[t]} &= \epsilon \operatorname{sign}(-U_{kl})/3, \\
\delta_{kk}^{\operatorname{sign}(-U_{kk})[t]} &= \epsilon/3, \\
\delta_{ll}^{\operatorname{sign}(-U_{ll})[t]} &= \epsilon/3, \\
\delta_{ij}^{(\pm)[t]} &= 0 \quad \forall (i,j) \notin \{(k,l),(k,k),(l,l)\}.
\end{aligned}$$

Otherwise, the update will have the form

$$\delta_{zz}^{\text{sign}(-U_{zz})[t]} = \epsilon,$$

$$\delta_{ij}^{(\pm)[t]} = 0, \quad \forall (i,j) \neq (z,z).$$

Theorem 3.1 is proven using the Karush–Kuhn–Tucker conditions of $(t.2)$. The general idea is that there are two possible updates that preserve the strong hierarchical structure, an update of a single main effect, or an update of an interaction effect along with its corresponding main effects. All possible updates of either type are considered and compared. When comparing the optimal main effect update to the optimal interaction update, a multiplicative factor of three is needed because when updating an interaction effect and its two main effects, the total step size must be evenly divided between the three effects.

In practice, the step size ϵ needs to carefully chosen; too large a step size would result in the generation of an unstable path, but too small a step size would result in excess computation to fully develop the path. Vaughan et al. (2017) suggest relying on using preliminary trace plots to evaluate the quality of the step size chosen. To make this approach more robust, we propose an adaptive step size that reduces in size as the path develops to produce more finely tuned steps as the change in the loss function decreases.

When working with GEE however, it must be understood that there is no closed

form loss function; instead, the assumption is that the estimating equations are approximations to the gradient of a loss function that is inaccessible. Thus, when using the SEE framework, monitoring the changes in the loss function is impossible; instead we opt to monitor the subsequent updates. When the current iteration's update will effectively undo the previous iteration's update, this indicates that the step size is too large for the updates to get the estimating equations closer to zero. If this is the case, the algorithm returns to the previous iteration, and repeats its update with a halved step size. The resulting re-update will produce current estimating function values that are closer to zero than before and the algorithm will be able to proceed. Should a similar situation occur again, the same steps may be taken.

3.3.2 Proof of Theorem 3.1

Proof. We show that the update $\delta^{[t]}$ specified in Theorem 3.1 is the solution to the optimization problem in $(t.2)$ by showing that it satisfies the corresponding KKT conditions. To simplify conceptualization, we will use the subscripting notation presented earlier to index vectors based on their correspondence with either interaction terms or with main effect terms. Additionally, because we are now minimizing with respect to both δ_{ii}^{+} and δ_{ii}^{-} terms for all i, we will include superscript notations of plus and minus symbols to further indicate an indexing referring to the positive and negative components of main effects, i.e. U_{ii}^{+} refers to the estimating equation pertaining to β_{ii}^{+}, which is not to be confused with the positive component of U_{ii}. Rather, $U_{ii}^{+} = U_{ii}$, where U_{ii} refers to the

classical estimating equations evaluation for the component β_{ii}.

Let $Q^{(1)}$ be a $(p^2+3p)/2\times 1$ vector such that $Q^{(1)}_{ij}(\delta^{[t]})$ is an element of the subgradient of the absolute value function evaluated at $\delta^{[t]}_{ij}$ for all i and j such that $i \neq j$ and zero elsewhere. Similarly, let $Q^{(2)}_{(i)}$ be a $(p^2+3p)/2\times 1$ vector such that $Q^{(2)}_{(i)ij}(\delta^{[t]})$ is an element of the subgradient of the absolute value function evaluated at $\delta^{[t]}_{ij}$ for all $j \neq i$ and zero elsewhere. Additionally, let $1^{(+)}_{(i)}$ and $1^{(-)}_{(i)}$ be $(p^2+3p)/2\times 1$ vectors such that $1^{(+)+}_{(i)ii} = 1$, $1^{(-)-}_{(i)ii} = 1$, and all other entries for both vectors are zero.

The KKT conditions corresponding to the optimization problem in $(t.2)$ evaluated at $\delta = \delta^{[t]}$ are

Stationarity,

$$
\begin{aligned}
-U = \gamma\{Q^{(1)}(\delta^{[t]}) + \sum_{i\in\{1,\ldots,p\}} 1^{(+)}_{(i)} + 1^{(-)}_{(i)}\} \\
+ \sum_{i\in\{1,\ldots,p\}} \gamma^{\mathcal{I}}_i\{Q^{(2)}_{(i)}(\delta^{[t]}) - 1^{(+)}_{(i)} - 1^{(-)}_{(i)}\} \\
+ \sum_{i\in\{1,\ldots,p\}} \gamma^{+}_i(-1^{(+)}_{(i)}) + \gamma^{-}_i(-1^{(-)}_{(i)});
\end{aligned}
\tag{3.2}
$$

Dual Feasibility,

$$\left.\begin{array}{l} \gamma \geq 0, \\ \gamma_i^{\mathcal{I}} \geq 0, \\ \gamma_i^{+} \geq 0, \\ \gamma_i^{-} \geq 0, \end{array}\right\} i \in \{1, \ldots, p\}; \tag{3.3}$$

Complementary Slackness,

$$\left.\begin{array}{l} \gamma(1^\top\{\delta^+ + \delta^-\} + \|\delta_{\mathcal{I}}\|_1 - \epsilon) = 0 \\ \gamma_i^{\mathcal{I}}(\|(\delta_{i1}, \ldots, \delta_{ii-1}, \delta_{ii+1}, \ldots, \delta_{ip})\|_1 - \delta_{ii}^+ - \delta_{ii}^-) = 0 \\ \gamma_i^+(-\delta_{ii}^+) = 0, \gamma_i^-(-\delta_{jj}^-) = 0 \end{array}\right\} i \in \{1, \ldots, p\}; \text{ and} \tag{3.4}$$

Primal Feasibility,

$$\left.\begin{array}{l} 1^\top\{\delta^+ + \delta^-\} + \|\delta_{\mathcal{I}}\|_1 - \epsilon \leq 0 \\ \|(\delta_{i1}, \ldots, \delta_{ii-1}, \delta_{ii+1}, \ldots, \delta_{ip})\|_1 - \delta_i^+ - \delta_i^- \leq 0 \\ -\delta_{ii}^+ \leq 0, -\delta_{ii}^- \leq 0 \end{array}\right\} i \in \{1, \ldots, p\}. \tag{3.5}$$

As there are two different possible updates that are applied, we will identify the appropriately selected values of $\gamma, \gamma_i^{\mathcal{I}}, \gamma_i^+, \gamma_i^-, Q^{(1)}$, and $Q_i^{(2)}$ that satisfy conditions represented in Equations (3.2)–(3.5) for each possible update.

The first update form considered is

$$\begin{aligned}
\delta_{kl}^{[t]} &= -\epsilon \operatorname{sign}(U_{kl})/3, \\
\delta_{kk}^{\operatorname{sign}(-U_{kk})[t]} &= \epsilon/3, \\
\delta_{ll}^{\operatorname{sign}(-U_{ll})[t]} &= \epsilon/3,
\end{aligned}$$

where $(k,l) = \arg\max_{i \neq j} |U_{ij}| + |U_{ii}| + |U_{jj}|$ and $|U_{kl}| + |U_{kk}| + |U_{ll}| > 3|U_{zz}|$, with $z = \arg\max |U_{ii}|$. In this case, we set $\gamma = (|U_{kl}| + |U_{kk}| + |U_{ll}|)/3, \gamma_i^{\mathcal{I}} = \gamma - |U_{ii}|, \gamma_i^{\operatorname{sign}(-U_{ii})} = 0$, and $\gamma_i^{-\operatorname{sign}(-U_{ii})} = 2|U_{ii}|$, for all i. Note that $|U_{zl}| + |U_{zz}| + |U_l| > 3|U_{zz}|$ implies $\gamma > 3|U_{zz}|/3 \geq |U_{ii}|$ for all i, thus $\gamma, \gamma_i^{\mathcal{I}}, \gamma_i^{\pm}$ are all non-negative for all i.

Let $Q_{ij}^{(1)}(\delta^{[t]}) = Q_{(i)ij}^{(2)}(\delta^{[t]}) = -U_{ij}/(\gamma + \gamma_i^{\mathcal{I}} + \gamma_j^{\mathcal{I}})$ for all (i,j) and zero elsewhere. Note that

$$\begin{aligned}
(k,l) &= \arg\max_{i \neq j} |U_{ij}| + |U_{ii}| + |U_{jj}| \\
\implies |U_{kl}| + |U_{kk}| + |U_l| &\geq |U_{ij}| + |U_{ii}| + |U_{jj}| \\
\implies |U_{kl}| + |U_{kk}| + |U_l| - |U_{ii}| - |U_{jj}| &\geq |U_{ij}| \\
\implies \gamma + \gamma_i^{\mathcal{I}} + \gamma_j^{\mathcal{I}} &\geq |U_{ij}| \\
\implies 1 &\geq |U_{ij}|/(\gamma + \gamma_i^{\mathcal{I}} + \gamma_j^{\mathcal{I}})
\end{aligned}$$

for all (i,j) and in the case of $(i,j) = (k,l)$ implies $Q_{kl}^{(1)}(\delta^{[t]}) = Q_{(i)ij}^{(2)}(\delta^{[t]}) = \operatorname{sign}(-U_{kl})$.

By construction, the Equations (3.3)–(3.5) are satisfied. Only the conditions specified

by Equation (3.2) remain.

We proceed by first showing that each equality

$$-U_{ii}^{\pm} = \Bigg[\gamma\{Q^{(1)}(\delta^{[t]}) + \sum_{a\in\{1,\ldots,p\}} 1_{(a)}^{(+)} + 1_{(a)}^{(-)}\} + \sum_{a\in\{1,\ldots,p\}} \gamma_a^{\mathcal{I}}\{Q_{(a)}^{(2)}(\delta^{[t]}) - 1_{(a)}^{(+)} - 1_{(a)}^{(-)}\} + \sum_{a\in\{1,\ldots,p\}} \gamma_a^{+}(-1_{(a)}^{(+)}) + \gamma_i^{-}(-1_{(a)}^{(-)})\Bigg]_{ii}^{\pm}$$

holds for all i, which pertains to the main effects, and then further showing that each equality

$$-U_{ij} = \Bigg[\gamma\{Q^{(1)}(\delta^{[t]}) + \sum_{a\in\{1,\ldots,p\}} 1_{(a)}^{(+)} + 1_{(a)}^{(-)}\} + \sum_{a\in\{1,\ldots,p\}} \gamma_a^{\mathcal{I}}\{Q_{(a)}^{(2)}(\delta^{[t]}) - 1_{(a)}^{(+)} - 1_{(a)}^{(-)}\} + \sum_{a\in\{1,\ldots,p\}} \gamma_i^{+}(-1_{(a)}^{(+)}) + \gamma_i^{-}(-1_{(a)}^{(-)})\Bigg]_{ij}$$

holds for all $i \neq j$, which pertain to the interaction effects.

For the equations pertaining to the main effects, we see that

$$\begin{aligned}
-U_{ii}^{\pm} &= \Bigg[\gamma\{Q^{(1)}(\delta^{[t]}) + \sum_{a\in\{1,\ldots,p\}} 1_{(a)}^{(+)} + 1_{(a)}^{(-)}\} + \sum_{a\in\{1,\ldots,p\}} \gamma_a^{\mathcal{I}}\{Q_{(a)}^{(2)}(\delta^{[t]}) - 1_{(a)}^{(+)} - 1_{(a)}^{(-)}\} + \sum_{a\in\{1,\ldots,p\}} \gamma_a^{+}(-1_{(a)}^{(+)}) + \gamma_i^{-}(-1_{(a)}^{(-)})\Bigg]_{ii}^{\pm} \\
&= \Big[\gamma\{Q_{ii}^{(1)\pm}(\delta^{[t]}) + 1\} + \gamma_i^{\mathcal{I}}\{Q_{(i)ii}^{(2)\pm}(\delta^{[t]}) - 1\} + \gamma_i^{\pm}(-1)\Big] \\
&= \gamma - \gamma_i^{\mathcal{I}} + \gamma_i^{\pm}(-1)
\end{aligned}$$

$$\begin{aligned}
\implies -U_{ii}^{\operatorname{sign}(-U_{ii})} &= \gamma - \gamma_i^{\mathcal{I}} \\
&= \gamma - \gamma + |U_{ii}| = |U_{ii}|, \text{ and,} \\
-U_{ii}^{-\operatorname{sign}(-U_{ii})} &= \gamma - \gamma_i^{\mathcal{I}} - \gamma_i^{-\operatorname{sign}(-U_{ii})} \\
&= \gamma - \gamma + |U_{ii}| - 2|U_{ii}| = -|U_{ii}|
\end{aligned}$$

for all i.

Note, in general $U_{ii}^{+} = U_{ii}$ and so

$$\begin{aligned}
\operatorname{sign}(-U_{ii}) &> 0 \\
\implies -U_{ii} &> 0 \\
\implies -U_{ii} &= |U_{ii}| \\
\implies -U_{ii}^{\operatorname{sign}(-U_{ii})} &= |U_{ii}|, \text{ and,} \\
-U_{ii}^{-\operatorname{sign}(-U_{ii})} &= -|U_{ii}|.
\end{aligned}$$

Similarly, we see that

$$\begin{aligned}
\operatorname{sign}(-U_{ii}) &< 0 \\
\implies -U_{ii}^{\operatorname{sign}(-U_{ii})} &= -(-U_{ii}) = |U_{ii}|, \text{ and} \\
-U_{ii}^{-\operatorname{sign}(-U_{ii})} &= -|U_{ii}|.
\end{aligned}$$

Therefore we see that for all i, the equations of this form hold.

For the equations pertaining to the interaction effects, we see

$$\begin{aligned}
-U_{ij} &= \Bigg[\gamma\{Q^{(1)}(\delta^{[t]}) + \sum_{a\in\{1,\ldots,p\}} 1_{(a)}^{(+)} + 1_{(a)}^{(-)}\} + \sum_{a\in\{1,\ldots,p\}} \gamma_a^{\mathcal{I}}\{Q_{(a)}^{(2)}(\delta^{[t]}) - 1_{(a)}^{(+)} - 1_{(a)}^{(-)}\} \\
&\qquad + \sum_{a\in\{1,\ldots,p\}} \gamma_i^{+}(-1_{(a)}^{(+)}) + \gamma_i^{-}(-1_{(a)}^{(-)})\Bigg]_{ij} \\
&= \left[\gamma\{Q_{ij}^{(1)}(\delta^{[t]})\} + \gamma_i^{\mathcal{I}}\{Q_{(i)ij}^{(2)}(\delta^{[t]})\} + \gamma_j^{\mathcal{I}}\{Q_{(j)ij}^{(2)}(\delta^{[t]})\}\right] \\
&= \left(\gamma + \gamma_i^{\mathcal{I}} + \gamma_j^{\mathcal{I}}\right)(-U_{ij}) / \left(\gamma + \gamma_i^{\mathcal{I}} + \gamma_j^{\mathcal{I}}\right) \\
&= -U_{ij}
\end{aligned}$$

for all i and j where $i \neq j$. Thus all of the equations are satisfied with this update when $|U_{kl}| + |U_{kk}| + |U_{ll}| > 3|U_{zz}|$.

The second update form considered is $\delta_{zz}^{\text{sign}(-U_{zz})[t]} = \epsilon$ where $z = \arg\max |U_{ii}|$ and $3|U_{zz}| \geq |U_{ij}| + |U_{ii}| + |U_{jj}|$ for all i and j such that $i \neq j$. In this case, we set $\gamma = 3|U_{zz}|, \gamma_i^{\mathcal{I}} = \gamma - |U_{ii}|, \gamma_i^{\text{sign}(-U_{ii})} = 0$, and $\gamma_i^{-\text{sign}(-U_{ii})} = 2|U_{ii}|$ for all i. Clearly, $\gamma, \gamma_i^{\mathcal{I}}, \gamma_i^{\pm}$ are all non-negative for all i.

Let $Q_{ij}^{(1)}(\delta^{[t]}) = -U_{ij}/\gamma$ for all $i \neq j$, and $Q_{(i)}^{(2)}(\delta^{[t]})$ is a vector of zeros for all i. Note that $3|U_{zz}| \geq |U_{ij}| + |U_{ii}| + |U_{jj}| \geq |U_{ij}|$, and thus $|-U_{ij}/\gamma| \leq |-U_{ij}/(|U_{ij}|)| \leq 1$. From these values we see that by the construction of $\delta^{[t]}$ and selection of z, Equations (3.3)–(3.5) are satisfied. This leaves only the conditions specified by Equation (3.2).

As with the first update form, we examine the equations pertaining to the main

effects and the interaction effects separately.

For the equations pertaining to the main effects, we see that

$$
\begin{aligned}
-U_{ii}^{\pm} &= \Bigg[\gamma\{Q^{(1)}(\delta^{[t]}) + \sum_{a\in\{1,\ldots,p\}} 1_{(a)}^{(+)} + 1_{(a)}^{(-)}\} + \sum_{a\in\{1,\ldots,p\}} \gamma_a^{\mathcal{I}}\{Q_{(a)}^{(2)}(\delta^{[t]}) - 1_{(a)}^{(+)} - 1_{(a)}^{(-)}\} \\
&\qquad + \sum_{a\in\{1,\ldots,p\}} \gamma_a^{+}(-1_{(a)}^{(+)}) + \gamma_a^{-}(-1_{(a)}^{(-)})\Bigg]_{ii}^{\pm} \\
&= \Big[\gamma\{Q_{ii}^{(1)\pm}(\delta^{[t]}) + 1\} + \gamma_i^{\mathcal{I}}\{Q_{(i)ii}^{(2)\pm}(\delta^{[t]}) - 1\} + \gamma_i^{\pm}(-1)\Big] \\
&= \gamma - \gamma_i^{\mathcal{I}} + \gamma_i^{\pm}(-1) \\
\implies -U_{ii}^{\operatorname{sign}(-U_{ii})} &= \gamma - \gamma_i^{\mathcal{I}} \\
&= \gamma - \gamma + |U_{ii}| \\
&= |U_{ii}| = -U_{ii}^{\operatorname{sign}(-U_{ii})}, \text{ and,} \\
-U_{ii}^{-\operatorname{sign}(-U_{ii})} &= \gamma - \gamma_i^{\mathcal{I}} - \gamma_i^{-\operatorname{sign}(-U_{ii})} \\
&= \gamma - \gamma + |U_{ii}| - 2|U_{ii}| \\
&= -|U_{ii}| = -U_{ii}^{-\operatorname{sign}(-U_{ii})}
\end{aligned}
$$

for all i. Therefore the equations of pertaining to the main effects hold for this update form.

For the equations pertaining to the interaction effects, we see

$$
-U_{ij} = \Bigg[\gamma\{Q^{(1)}(\delta^{[t]}) + \sum_{a\in\{1,\ldots,p\}} 1_{(a)}^{(+)} + 1_{(a)}^{(-)}\} + \sum_{a\in\{1,\ldots,p\}} \gamma_a^{\mathcal{I}}\{Q_{(a)}^{(2)}(\delta^{[t]}) - 1_{(a)}^{(+)} - 1_{(a)}^{(-)}\}
$$

$$
\begin{aligned}
&\quad + \sum_{a\in\{1,\ldots,p\}} \gamma_i^+(-1_{(a)}^{(+)}) + \gamma_i^-(-1_{(a)}^{(-)}) \Bigg]_{ij} \\
&= \left[\gamma\{Q_{ij}^{(1)}(\delta^{[t]})\} + \gamma_i^{\mathcal{I}}\{Q_{(i)ij}^{(2)}(\delta^{[t]})\} + \gamma_j^{\mathcal{I}}\{Q_{(j)ij}^{(2)}(\delta^{[t]})\}\right] \\
&= \left[\gamma\{-U_{ij}/\gamma\} + \gamma_i^{\mathcal{I}}\{0\} + \gamma_j^{\mathcal{I}}\{0\}\right] \\
&= -U_{ij}
\end{aligned}
$$

for all i and j where $i \neq j$. Therefore, for this update form Equation (3.2) is satisfied when $3|U_{zz}| \geq |U_{ij}| + |U_{ii}| + |U_{jj}|$. Therefore, the condition specified by Equation (3.2) is satisfied for both update forms. Thus all of the conditions are satisfied. □

3.3.3 ACTS Stagewise Estimating Equations

Though HiLa has a desirable connection to the hierarchical lasso, it does have some limitations. First, it is making a comparison among all possible interactions, which can result in a great deal of computation when the number of covariates is very large. Secondly, simpler comparison mechanisms may exist to perform optimal interaction selection. Therefore, HiLa has a higher computational cost than is typically desired in a stagewise technique. Zhu et al. (2014) introduced a different approach that address these concerns. Instead of comparing all interaction terms, their approach makes use of an active set that keeps track of which main effects have already been selected. In each iterative step, interactions are considered to be updated only if both (or one for weak hierarchy) of its main effects are in the active set.

We re-characterize the active set approach of Zhu et al. (2014) so that it can be used in the SEE framework (Tibshirani, 2015; Vaughan et al., 2017) to allow for non-Gaussian clustered data. The resulting algorithm is hence referred to as ACTS. Let $\mathcal{A}^{[t]} = \{i : \beta_{ii}^{[t]} \neq 0\}$ and $\mathcal{A}^{2[t]} = \{(i,j) : i,j \in \mathcal{A}^{[t]}\}$, and consider the stagewise regression procedure of the following form:

$$\begin{aligned}
&(t.a)\ \text{Given } \beta^{[t-1]}, \text{ update the nuisance parameters to obtain } \nu^{[t]},\\
&(t.b)\ \delta^{[t]} = \underset{\substack{\delta\in\mathbb{R}^{(p^2+p)/2}\\ \delta_{ij}=0\forall(i,j)\notin\mathcal{A}^{2[t]}}}{\arg\min} \langle U_{[0]}(\beta^{[t-1]}, \nu^{[t]}), \delta\rangle \text{ subject to } \phi(\delta) \leq \epsilon,\\
&(t.c)\ \beta^{[t]} = \beta^{[t-1]} + \delta^{[t]},
\end{aligned} \tag{3.6}$$

where ϕ is a properly chosen penalty function. Using an ℓ_1 norm for ϕ, as a slight extension of the work presented by Wolfson (2011), the appropriate update as a result of the solution to $(t.b)$ will be of the form

$$\begin{aligned}
\delta_{kl}^{[t]} &= \epsilon\,\text{sign}(-U_{kl}),\\
\delta_{ij}^{[t]} &= 0, \quad \forall(i,j) \neq (k,l),
\end{aligned}$$

where

$$(k,l) = \underset{(i,j)\in\{(1,1),\ldots,(p,p)\}\cup\mathcal{A}^{2[t]}}{\arg\max} |U_{ij}|.$$

The updating rules are very simple while preserving the desired strong hierarchy.

Additionally, as in Zhu et al. (2014), ACTS can be modified to preserve weak hierarchy by redefining $\mathcal{A}^{2[t]} = \{(i,j) : i \in \mathcal{A}^{[t]} \text{ or } j \in \mathcal{A}^{[t]}\}$.

3.3.4 Algorithm Details

Specifics of the HiLa and ACTS methods are summarized in Algorithms 3 & 4. Both Algorithms begin with an empty model, represented by $\beta^{[0]} = 0$. The intercept is updated first to be $\beta_0^{[t]}$, the root of $U_0(\beta^{[t-1]}, \beta_0, \psi^{[t-1]}, \alpha^{[t-1]})$ with respect to β_0. Then, ψ and α, nuisance parameters, are updated using method of moments estimators constructed from the Pearson residuals evaluated at $\beta^{[t-1]}$ (Liang and Zeger, 1986). In the next step the optimal update is determined based on the approach being used. Ultimately, the update is applied to yield the next est of estimates $\beta^{[t]}$, and the process is repeated.

Vaughan et al. (2017) discussed termination procedures for stagewise techniques, noting that the algorithms can be run for a set number of iterations, or they can be terminated if the Taylor approximation to the change in the loss function, $|\langle U_{[0]}(\beta^{[t-1]}, \nu^{[t]}), \delta^{[t]} \rangle|$, falls below a pre-specified threshold. In either case, if it is determined that the algorithm was terminated pre-maturely, it is simple to restart the algorithm for its terminal iteration.

As explained in Section 3.3.1, the initial step size choice is important, and it is possible to diagnose a poor step size choice by visual examination of trace plots. We propose to improve this approach by making the step size adaptive to the amount of change in the loss function at the current iteration. In principle, this is executed by

checking if the previous update is undone by the current one; if it is the previous iteration is re-done with a reduced step size. In practice, what it means for a step to "undo" the previous step can be a little unclear, specifically in the cases where multiple effects are being updated at the same time. We propose that ℓ_1-norm of the difference between the two updates be checked against some threshold ζ; if the threshold is not exceeded, then the step size is reduced and the previous iteration is repeated. Eventually, when the step size becomes less than threshold, the threshold is never exceeded. Once this happens, we propose reducing the threshold.

We note that in Algorithm 3 since the update form where an interaction term and both main effects are updated at the same time, because the update direction for each main effect matches the sign of the corresponding estimating equations, we need not actually track the positive and negative components. Instead, we can simply update the original main effect form by a small amount in the desired direction and have the same effect.

3.3.5 An Illustration

To demonstrate the efficacy of the proposed techniques, we consider a clustered Gaussian regression with a simulated data set of 40 clusters of size 4 with pairwise correlation of 0.6. There are 5 covariates resulting in 5 main effects, and 10 interaction effects. Three of the main effects have values of $-0.2, 0.2$, and -0.2, while the remaining two are 0. Three of the interaction effects are non-zero with values of 0.1, -0.2, and one of -0.1;

the rest are zero. Strong hierarchy is preserved in this setting.

The solution paths of β generated by the all-pairs approach, which results from applying lasso directly to all main and interaction effects; hierarchical lasso (hierNet), which has been implemented in the R package `hierNet`; HiLa; and ACTS are presented in Figure 9. The all-pairs approach very quickly broke the hierarchy structure, adding an interaction effect before any main effects were added. The hierarchical lasso approach preserves the strong hierarchy, but doesn't capture all important effects before including

Algorithm 3 Hierarchical Lasso Stagewise Estimating Equations (HiLa)

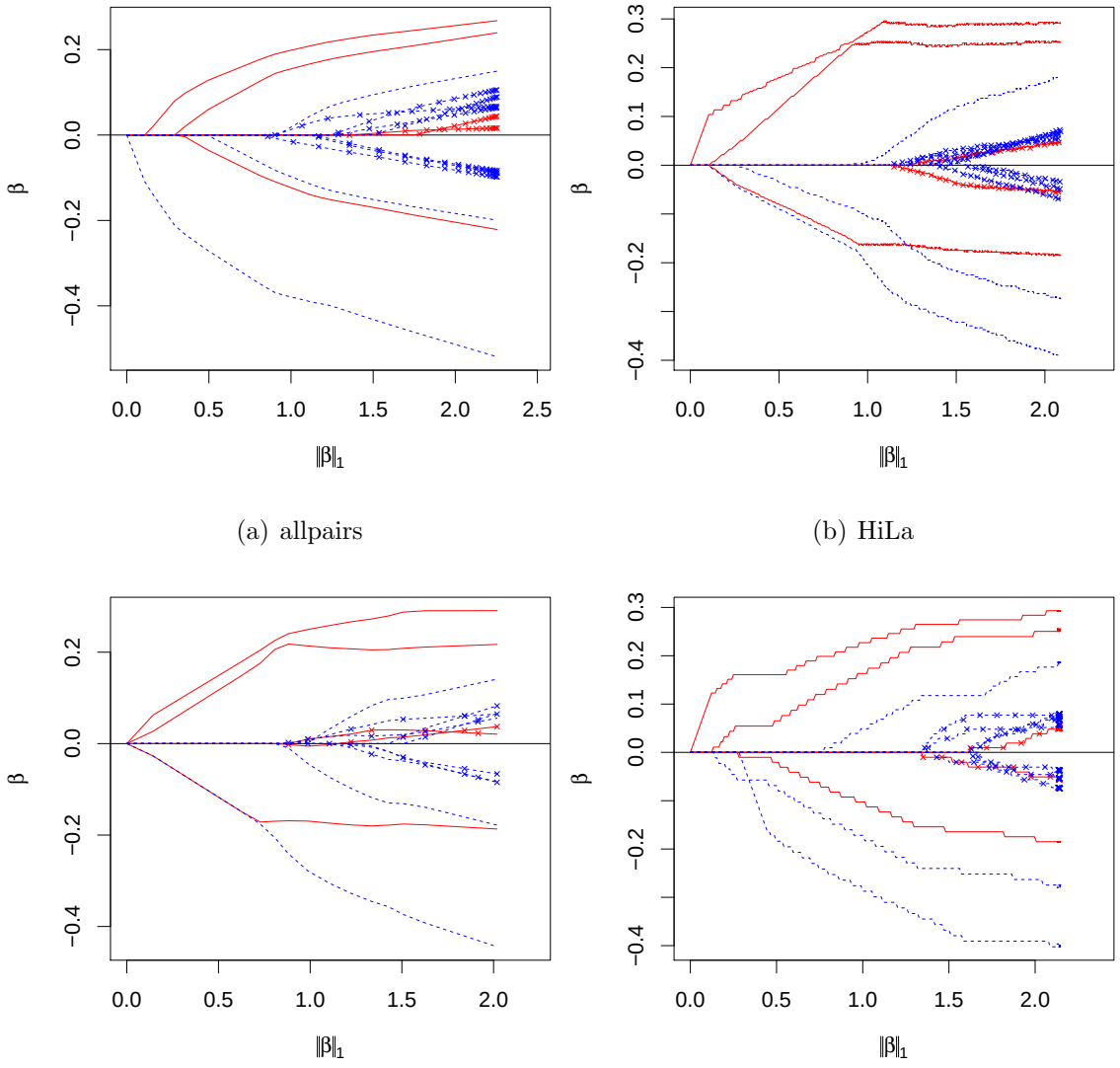
Initialize: $\beta^{[0]} = 0$, $\nu^{[0]}$, $\epsilon > 0$, and $\zeta > 0$.
  for $t = 1, 2, \ldots$ **do**
    $(H.1)$ Given $\beta^{[t-1]}$, update the nuisance parameters to obtain $\nu^{[t]}$.
    $(H.2)$ $(k, l) = \arg\max_{i \neq j} |U_{ij}(\beta^{[t-1]}; \nu^{[t]})| + |U_{ii}(\beta^{[t-1]}; \nu^{[t]})| + |U_{jj}(\beta^{[t-1]}; \nu^{[t]})|$, and,
    $z = \arg\max |U_{ii}(\beta^{[t-1]}; \nu^{[t]})|$.
    if $|U_{kl}(\beta^{[t-1]}; \nu^{[t]})| + |U_{kk}(\beta^{[t-1]}; \nu^{[t]})| + |U_{ll}(\beta^{[t-1]}; \nu^{[t]})| > 3|U_{zz}(\beta^{[t-1]}; \nu^{[t]})|$ **then**
      $(H.3a)$ $\delta_{kl}^{[t]} = \epsilon \operatorname{sign}(-U_{kl}\left(\beta^{[t-1]}; \nu^{[t]}\right)/3$,
      $\delta_{kk}^{[t]} = \epsilon \operatorname{sign}\left(-U_{kk}(\beta^{[t-1]}; \nu^{[t]})\right)/3$,
      $\delta_{ll}^{[t]} = \epsilon \operatorname{sign}\left(-U_{ll}(\beta^{[t-1]}; \nu^{[t]})\right)/3$.
    else
      $(H.3b)$ $\delta_{zz}^{[t]} = \epsilon \operatorname{sign}\left(-U_{zz}(\beta^{[t-1]}; \nu^{[t]})\right)$,
    end if
    if $\|\delta^{[t-1]} - \delta^{[t]}\|_1 < \zeta$ **then**
      $(H.4)$ Set $t = t - 1, \epsilon = \epsilon/2$
      if $\epsilon \leq \zeta$ **then**
        $(H.5)$ Set $\zeta = \zeta/10$
        $(H.6)$ Go To $(H.1)$
      end if
    else
      $(H.7)$ $\beta^{[t]} = \beta^{[t-1]} + \delta^{[t]}$.
    end if
  end for

some unimportant ones; this is likely due the high within-cluster correlation. The proposed techniques, HiLa and ACTS, however both well distinguish the important terms from the irrelevant ones while preserving the strong hierarchy structure.

To demonstrate the computational advantages of ACTS and HiLA over hierarchical lasso, we conduct a small time trial simulation. In this time trial, we use the same setting as the previous illustrative example, but we vary the number of covariates being considered while keeping the non-zero effects the same. We consider when p, the number of covariates, is 10, 50, 100, and 200. Each setting is completed with 100 replicates. HiLa and ACTS both are run for 400 iterations to generate their paths. Two hierarchical

Algorithm 4 Active Set Stagewise Estimating Equations (ACTS)

Initialize: $\beta^{[0]} = 0$, $\nu^{[0]}$, $\epsilon > 0$, $\mathcal{A}^{[0]} = \varnothing$, $\mathcal{A}^{2[0]} = \varnothing$, and $\zeta > 0$.
  for $t = 1, 2, \ldots$ **do**
    $(A.1)$ Given $\beta^{[t-1]}$, update the nuisance parameters to obtain $\nu^{[t]}$.
    $(A.2)$ $(k, l) = \underset{(i,j):(i,j)\in\mathcal{A}^{2[t]}\cup\{(1,1),\ldots(p,p)\}}{\arg\max} \|U_{ij}(\beta^{[t-1]}; \nu^{[t]})\|$.
    $(A.3)$ $\delta_{kl} = \epsilon \operatorname{sign}\left(-U_{ij}(\beta^{[t-1]}; \nu^{[t]})\right)$
    if $\|\delta^{[t-1]} - \delta^{[t]}\|_1 < \zeta$ **then**
      $(A.4)$ Set $t = t - 1, \epsilon = \epsilon/2$
      if $\epsilon \leq \zeta$ **then**
        $(A.5)$ Set $\zeta = \zeta/10$
        $(A.6)$ Go To $(H.1)$
      end if
    else
      if $k = l$ **then**
        $(A.7)$ $\mathcal{A}^{[t]} = \mathcal{A}^{[t-1]} \cup \{k\}$
      end if
      $(A.8)$ $\mathcal{A}^{2[t]} = \{(i, j) : i, j \in \mathcal{A}^{[t]}\}$
      $(A.9)$ $\beta^{[t]} = \beta^{[t-1]} + \delta^{[t]}$.
    end if
  end for



(a) allpairs

(b) HiLa

(c) hierarchical lasso

(d) ACTS

Figure 9: The illustration example: paths of individual coefficient estimates against the ℓ_1 norm of the estimates, generated by the all-pairs approach(lasso) (a), HiLa (b), hierarchical lasso (c), and ACTS (d). All the paths are plotted against the ℓ_1 norm of the solution, e.g., $\|\widehat{\beta}\|_1$, along the path. Main effects are denoted with a solid line while interaction effects are denoted with a dashed line. Paths of irrelevant predictors are marked with "x" and those of important predictors are left unmarked.

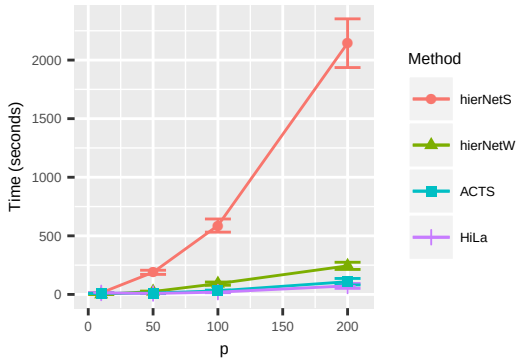


Figure 10: Time trials: the average run times in seconds to generate a full path for HiLa, ACTS, and hierarchical lasso with strong hierarchy (hierNetS) and weak hierarchy (hierNetW). Error bars are constructed using one standard deviation.

lasso solution paths are genereated using the `R` package `hierNet`; one using the strong hierarchy setting (hierNetS), and one using the weak hierarchy setting (hierNetW). The resulting average run times as a function of the number of covariates are presented in Figure 10. We see that as the number of covariates increases, the runtime for the hierarchical lasso's strong setting grows exponentially and is quickly out paced by HiLa and ACTS. Even when compared to hierNet's implementation only enforcing weak hierarchy both HiLa and ACTS have shorter run times.

3.4 Simulation

We conduct two longitudinal simulation studies, one with a Gaussian response, and one with Poisson response. Three simulation factors are considered. First, sample size size is either 45 or 90 clusters of size $k_i = k = 4$. Second, the number of covariates p is either 200 or 400. Finally, we consider two interaction hierarchy settings: (I) strong hierarchy

and (II) weak hierarchy. In (I) the true model adheres to a strong hierarchy. In (II) the true model only adheres to a weak hierarchy structure. Each configuration is replicated 100 times.

In the Gaussian study, the $k \times 1$ response vector for each cluster is generated from the multivariate normal distribution. The mean of the multivariate distribution for cluster i is $X_i\beta$, where the intercept value is $\beta_0 = 0$ and X_i is a $k \times p$ covariate matrix. The covariance matrix Σ_y for the distribution of cluster i has an exchangeable correlation structure, i.e., Σ_y has diagonal elements σ_y^2, and off diagonal elements $\sigma_y^2\rho_y$ with $\rho_y = 0.3$ to indicate a moderate amount of correlation. The variance σ_y^2 is chosen to fix the signal to noise ratio (SNR) to be 2, where SNR $= V[\sum_{i=1}^{p} X_i\beta]/\sigma_y^2$, with $V[\cdot]$ indicating the variance.

All covariates are independently generated on an individual level from a Gaussian distribution with a mean of 0 and a variance of 1. All pairwise interactions are also considered in the modeling process. In all settings, there is a total of 15 non-zero main effects and 5 non-zero interaction effects. All non-zero effects that are randomly generated from a uniform distribution from 0.5 to 1 for each replicate.

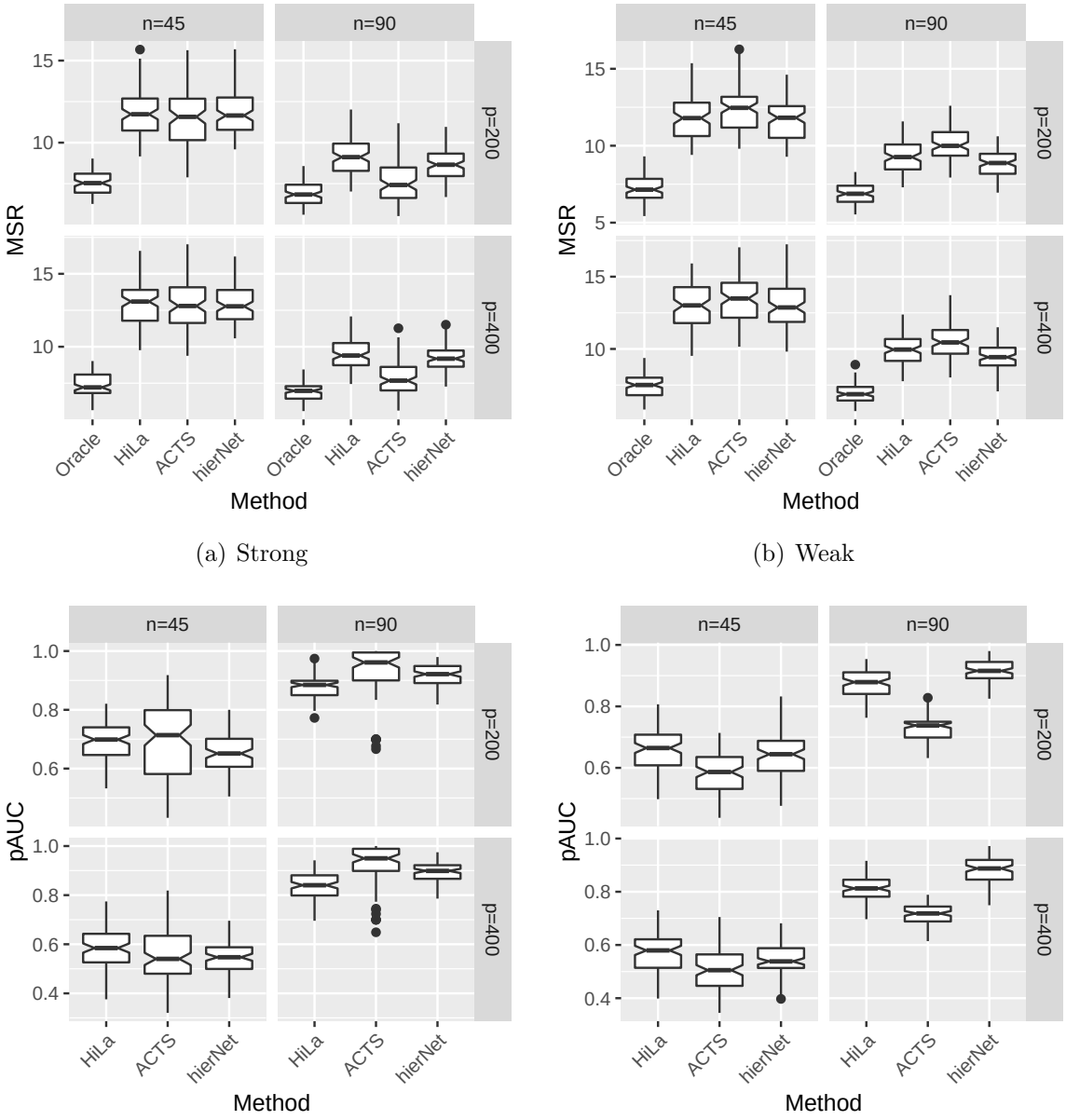
The proposed methods HiLa and ACTS were compared to the hierarchical lasso implementation in `hierNet`. In HiLa and ACTS, we use an exchangeable working correlation, and set the number of iterations as $N = 200$ and the initial step size as $\epsilon = 0.5$. Due to the significant computational cost of enforcing strong hierarchy with hierNet, in order to make our large scale simulation feasible we enforce only a weak hierarchy

when implementing the hierarchical lasso. To compare all methods fairly, they are tuned based on independently generated testing data of large size. Specifically, the prediction error of a given solution on a given solution path is defined as the predictive MSE, ignoring the within-cluster dependence. The solution with the lowest prediction error in each solution path is then selected as the final solution for that path. We also use the lowest prediction error as a predictive measure for comparing different methods. To evaluate the variable selection performance, we use the partial area under the ROC curve (pAUC) from the beginning of he solution path up the the point in the path when the false positive count first breaches 80, as well as the true positive count at the point in the path when the model includes 40 predictors (TP40).

Table 6 reports the simulation results across all simulation settings. Figure 11 shows the boxplots of the predictive measures and partial AUCs. Consistently, the proposed methods outperform hierNet predictive performance and model selection. The one case where hierNet outperforms the proposed methods is when $n = 90$ and the true model structure has a weak hierarchy. This is likely a result of the proposed methods being misspecified as enforcing strong hierarchy whereas hierNet does not suffer from this misspecification in this study. This advantage masks any gains from accounting for the within cluster correlation, which is also reduced due to large sample size. We see that in the weak hierarchy settings, HiLa tends to have an advantage over ACTS. This is due to the fact that when an interaction effect is strong, but the corresponding main effects are weak, ACTS will not consider it until its main effects have been selected. In the

Table 6: Simulation results with from 100 replicates. Reported are the mean and in the parentheses the standard deviations of the predictive measure (Msr), the partial area under the curve (pAUC), and the true positive count at model size 40 (TP40).

			Strong Hierarchy			Weak Hierarchy		
			Msr	pAUC	TP40	Msr	pAUC	TP40
Gaussian	$n = 45$	$p = 200$	Oracle	7.53 (0.85)			7.28 (0.99)		
			hierNet	11.89 (1.56)	0.65 (0.07)	13.42 (1.76)	11.67 (1.48)	0.64 (0.08)	13.53 (1.99)
			HiLa	11.88 (1.65)	0.69 (0.07)	13.91 (1.65)	11.86 (1.68)	0.65 (0.08)	13.15 (1.87)
			ACTS	11.57 (2.07)	0.69 (0.15)	14.07 (3.06)	12.38 (1.68)	0.58 (0.08)	11.93 (1.59)
		$p = 400$	Oracle	7.37 (0.90)			7.44 (0.93)		
			hierNet	13.00 (1.49)	0.54 (0.08)	11.44 (1.87)	13.04 (1.73)	0.55 (0.07)	11.37 (1.86)
			HiLa	13.04 (1.63)	0.58 (0.09)	11.75 (1.91)	12.98 (1.78)	0.57 (0.09)	11.49 (1.85)
			ACTS	12.89 (1.92)	0.56 (0.13)	11.39 (2.66)	13.43 (1.89)	0.51 (0.09)	10.34 (1.79)
	$n = 90$	$p = 200$	Oracle	6.93 (0.80)			6.90 (0.73)		
			hierNet	8.69 (1.12)	0.92 (0.04)	18.65 (1.21)	8.88 (1.01)	0.92 (0.04)	18.82 (1.13)
			HiLa	9.14 (1.23)	0.88 (0.05)	17.49 (0.99)	9.33 (1.17)	0.87 (0.05)	17.38 (1.25)
			ACTS	7.71 (1.46)	0.94 (0.09)	18.96 (1.84)	10.15 (1.33)	0.73 (0.05)	14.57 (1.00)
		$p = 400$	Oracle	6.95 (0.75)			6.95 (0.75)		
			hierNet	9.24 (1.01)	0.89 (0.05)	18.23 (1.35)	9.43 (1.08)	0.88 (0.06)	18.13 (1.36)
			HiLa	9.50 (1.25)	0.84 (0.06)	16.66 (1.22)	9.97 (1.27)	0.81 (0.06)	16.34 (1.23)
			ACTS	7.90 (1.38)	0.94 (0.09)	18.81 (2.01)	10.48 (1.33)	0.71 (0.05)	14.23 (0.99)
Poisson	$n = 45$	$p = 200$	Oracle	1.43 (0.32)			1.35 (0.12)		
			allPairs	4.06 (1.59)	0.42 (0.14)	5.21 (1.92)	3.10 (0.63)	0.41 (0.11)	5.15 (1.59)
			HiLa	2.59 (0.85)	0.80 (0.09)	9.95 (1.15)	2.31 (0.40)	0.76 (0.09)	9.60 (1.19)
			ACTS	2.24 (0.62)	0.87 (0.13)	10.69 (1.64)	2.50 (0.46)	0.69 (0.09)	8.47 (1.18)
		$p = 400$	Oracle	1.42 (0.21)			1.36 (0.12)		
			allPairs	4.74 (1.69)	0.27 (0.12)	3.60 (1.63)	3.43 (0.56)	0.26 (0.10)	3.35 (1.42)
			HiLa	3.06 (1.35)	0.74 (0.11)	9.42 (1.36)	2.50 (0.44)	0.68 (0.10)	8.54 (1.37)
			ACTS	2.55 (1.08)	0.85 (0.14)	10.50 (1.64)	2.65 (0.49)	0.62 (0.11)	7.64 (1.38)
	$n = 90$	$p = 200$	Oracle	1.38 (0.25)			1.30 (0.10)		
			allPairs	2.20 (0.80)	0.86 (0.07)	11.04 (1.04)	1.93 (0.24)	0.86 (0.08)	10.73 (1.26)
			HiLa	1.83 (0.67)	0.94 (0.05)	11.53 (0.58)	1.72 (0.18)	0.95 (0.05)	11.53 (0.63)
			ACTS	1.30 (0.11)	1.00 (0.01)	12.00 (0.00)	2.02 (0.29)	0.79 (0.07)	9.48 (0.88)
		$p = 400$	Oracle	1.37 (0.15)			1.31 (0.10)		
			allPairs	2.81 (1.25)	0.73 (0.10)	9.25 (1.45)	2.23 (0.34)	0.74 (0.10)	9.21 (1.58)
			HiLa	1.99 (0.60)	0.93 (0.05)	11.41 (0.67)	1.83 (0.24)	0.91 (0.06)	11.12 (0.82)
			ACTS	1.39 (0.27)	0.99 (0.04)	11.95 (0.50)	2.10 (0.31)	0.76 (0.06)	9.23 (0.75)



(a) Strong

(b) Weak

(c) Strong

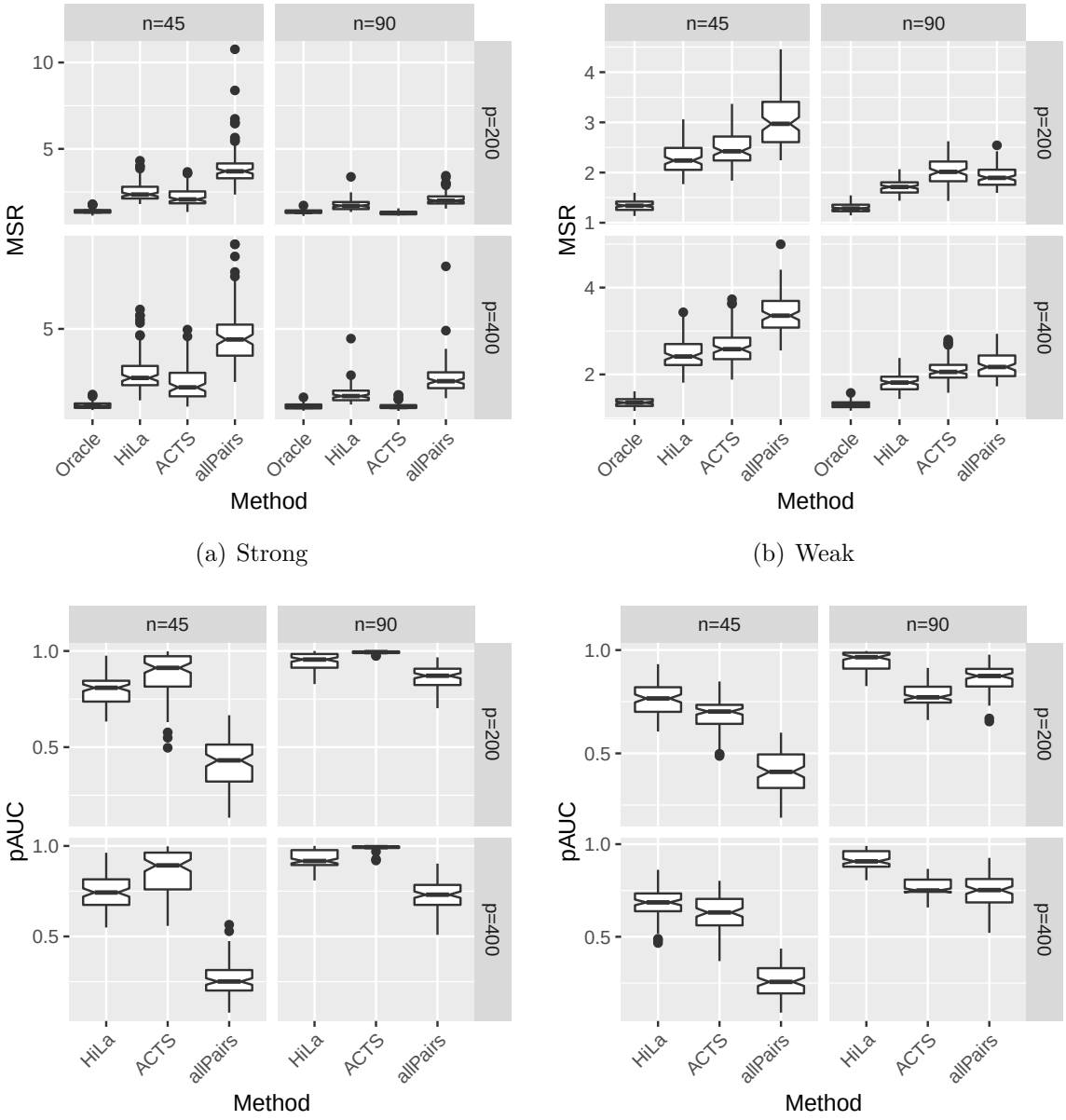
(d) Weak

Figure 11: Gaussian setting: boxplots of the trimmed predictive performance (MSR) and trimmed partial area under the curve paths (pAUC) over 100 replicates, where the top and bottom 2% have been excluded.

strong hierarchy settings however, ACTS tends to be the best, especially as the sample size increases. Finally, both proposed techniques are executed far faster than hierNet, with the largest discrepancy having a factor of 6, on average.

The setting of our study with Poisson response is similar to that with Gaussian response. The within-cluster dependence of Poisson responses is set to be a normal copula with an exchangeable correlation structure whose off-diagonal values are $\rho_y = 0.3$. The marginal Poisson distributions are set to have mean $g^{-1}(X_{ij}\beta)$ for the jth observation in the ith cluster, where g is the log link function. The number of non-zero main effects is reduced to 9 and the number of non-zero interaction effects is reduced to 3, all of which are randomly generated from a uniform distribution from 0.15 to 0.3. The techniques used were HiLa, ACTS, and lasso applied to all possible interactions (allPairs), provided by `glmnet`, as hierNet is not been implemented for Poisson regression. Instead of the predictive MSE, the model paths are evaluated based on the predictive deviance assuming independence. In HiLa and ACTS, we use an exchangeable working correlation, and set the number of iterations as $N = 300$ and the initial step size as $\epsilon = 0.1$.

The simulation results are summarized in the lower block of Table 6. Figure 12 shows the boxplots of the predictive measures and partial AUCs. Most observations in the Gaussian case remain. Both proposed techniques outperform allPairwise in prediction, and have far better variable selection performance, as the allPairwise approach fails to account for the hierarchy. In one exceptional performance, we see that in the Strong setting with $n = 90$ and $p = 200$, ACTS manages to not only perform near perfect



(a) Strong

(b) Weak

(c) Strong

(d) Weak

Figure 12: Poisson setting: boxplots of the trimmed predictive performance (MSR) and trimmed partial area under the curve paths (pAUC) over 100 replicates, where the top and bottom 2% have been excluded.

model selection on average, but also has superior predictive performance compared to the Oracle estimator. This superior predictive performance is attributable to the shrinkage effect of our penalization approximation.

3.5 Connecticut Adolescent Suicide Risk Study

Suicide among youth is among the most serious public health problems in the United States (e.g., Chen and Aseltine, 2017). The Centers for Disease Control (CDC) reported that the suicide rate was 11.1 per $100,000$ among youth aged 15–24 in 2013, making it the third leading cause of death in this age group. Among adolescents aged 15–19 years, the number of suicides tripled between 1950 and 2011. Addressing the issue of teen suicide is difficult and requires a thorough investigation into the potential risk factors in order to optimize resource allocation. It is important that the interaction terms be considered; if interaction effects are found, that would dramatically affect the analysis of the study.

Hospitalization data for teens aged 15–19 from 2010–2014 was collected from the Connecticut Hospital Inpatient Discharge Database (Vaughan et al., 2017). The hospitalization counts for each of 119 school districts that contain a high school in the state of Connecticut were recorded and compared with various socio-economic, academic, and demographic measure for each school district. Two different hospitalization types were examined. First, as a proxy for suicide risk in a given school district, the ratios of annual

counts of inpatient hospitalizations due to intentional self-injury and the population sizes of the 15–19 age group for each school district were considered. Second, the ratios of the annual overall inpatient hospitalizations to the same population size for the same age group were used as a relative measure of the general well-being of the teens aged 15–19 in each school district in order to compare with the analysis of the relative suicide risk level. One school district (Thomaston School District) had missing values and was thus removed, leaving a final sample size of 118 clusters of size 5.

Many characteristics of the school districts were collected as covariates: average household size, proportion of population under the age of 18, the proportion of population that is white, the average score on the Connecticut Academic Performance Test (CAPT) for the school district, the average attendance rate of the high schools in the district, the incidence rate (the ratio between the number of incidences and the total enrollment), the median income of the district, and grant status (indicating whether a school district has ever received a state grant related to suicide prevention). We also use the linear term of time to capture the trend over time. All pairwise interaction terms were also considered.

The proposed techniques HiLa and ACTS were used to conduct a Poisson regression analysis, in which the annual hospitalization counts served as the clustered response and the logarithm of the district population of the 15–19 age group as the offset. The aforementioned variables are standardized and used as predictors. The step size was set

to $\epsilon = 0.005$. The model selection was done using 10-fold cross-validation on models selected by HiLa and ACTS but re-fit with traditional GEE, using the deviance to measure the predictive performance of the models. While HiLa and ACTS resulted in similar solution paths, ACTS produced the best model for modeling the overall hospitalization counts, and HiLa produced the best model for modeling the suicide-related hospitalization counts. For each hospitalization type, the data was refit using the entire dataset and regular GEEs with the optimal model structure selected in the cross-validation step.

Table 7 reports the refitting estimates from GEE. The interaction between the proportion of the population that is white and the average CAPT is negatively associated with the overall hospitalization rate. This means that the negative association between the average CAPT and the overall hospitalization is weaker when the proportion of whites in the district's population is low than when the proportion is high. The interactions between the proportion of the population that is white and time and between the average CAPT and time are both positively associated with the overall hospitalization rate, indicating while both the proportion of the population that is white or the average CAPT in a district are negatively associated with the hospitalization rate, the association becomes weaker over the course of the study. Additionally, the median income is negatively related to the overall hospitalization rate, indicating that the general well-being of teens aged 15–19 tends to be better in wealthier communities. Also, there is a negative association between the overall hospitalization rate and time, indicating that in general, the well-being of teens is improving over this time period. The main effects

Table 7: Suicide study: the fitted Poisson regression models for the overall hospitalization counts and the suicide-related hospitalization counts. HiLa and ACTS were used for model selection, and the estimation results were from refitted models using GEE. Between HiLa and ACTS, ACTS produced the best model for the overall hospitalizations, whereas HiLa produced the best model for the suicide-related hospitalizations.

	Overall		Suicidal	
	estimate	ratio	estimate	ratio
(Intercept)	−3.4920	0.0304	−6.5570	0.0014
Main Effects				
Prop. White	−0.0060	0.9944	0.0055	1.0055
Average household size			−0.2782	0.7571
Prop. U. 18			−0.0235	0.9768
Med. Income	−0.0614	0.9404	−0.0062	0.9939
Avg. CAPT	−0.0161	0.9840		
Attendance rate			−0.0622	0.9397
Incidence rate			0.0096	1.0097
Grant			−0.1798	0.8354
Time	−0.0471	0.9540	0.0286	1.0290
Interaction Effects				
Prop. White & Avg. CAPT	−0.0002	0.9998		
Prop. White & Time	0.0002	1.0002		
Avg. CAPT & Time	0.0065	1.0065		
Prop. U. 18 & Med. Income			−0.0037	0.9964
Prop. White & Attendance rate			0.0017	1.0017
Attendance rate & Incidence rate			0.0118	1.0118
Grant & Time			0.0460	1.0471

selected here echoes those selected in previous studies (Vaughan et al., 2017).

In the analysis of the self-inflicted hospitalizations, the interaction between the proportion of the population under 18 and the median income was found to be negatively associated with the suicide related hospitalization rate. This indicates that while there was a negative association between the proportion of the population that was under 18 and the suicide hospitalization rate, the effect was stronger in higher income school districts than in lower income school districts. This may suggest that the benefits from having peers may be outweighed by a lack of available resources in the schools. The interaction between the attendance rate and the incidence rate was found to be positive with a large effect. This positive effect indicates that when the incidence rate is low, the negative association between the attendance rate and the suicide related hospitalization rate is much stronger, but when the incidence rate is high, the association actually becomes a positive one, meaning that among schools with a high incidence rate, those with a higher attendance rate were also more likely to have a higher suicide hospitalization rate, on average. The interaction between the proportion of the population that is white and the attendance rate was also found to be positive. Furthermore, the interaction effect between the grant status of a school district and time was found to be positive, which indicates that while overall the grant status of a school district was negatively associated with the suicide hospitalization rate, the effect lessened over time. Finally, the average household size and the proportion of the population under 18, for a given median income, were both found to be negatively associated with the suicide related

hospitalization rate, when holding other variables fixed, which may suggest that the more communal options available to teens result in better overall conditions for their mental health.

Chapter 4

Discussion and Future Work

4.1 Discussion

Several directions merit further research. The HiSEE technique can be extended to address more complicated grouping structures, such as hierarchical structure and group overlapping. It is seen that certain types of constraints on the model structures can be more easily handled by stagewise estimation approaches. The proposed stagewise approaches can be applied in the integrative analysis of multiple data sets, especially in high-throughput genomic studies (Ma et al., 2011; Liu et al., 2014).

Integrating the work of Vaughan et al. (2017), both ACTS and HiLa can be extended to allow for groups of covariates. In such an extension, questions about how to characterize the interaction between groups of covariates would need to be addressed. Such techniques could be applied to genomic studies in which a priori grouping structures and interactions would both be of great interest. Alternatively, ACTS could potentially be extended to allow for higher order levels of hierarchy such as three-way interaction models.

4.2 Grouped Interaction Selection

The nature of grouped covariates is such that there is some assumed dependence within the group. This being the case, it is unreasonable to include interaction terms for covariates within the same group when accounting for group structures as it is effectively accounting for the dependence within the group twice. Furthermore, in the case where the group is defined based on a dummy coding for factor levels, including interaction terms is purely nonsensical.

So, it is proposed that only interactions of covariates between groups be considered when a grouping structure is being incorporated into the regression. This leads to non-overlapping groups of covariate interactions where each group of interactions contains the interaction terms corresponding to the main effects in two different groups.

Consider the following expansion of the original problem description. Suppose that there are J groups of main effects with p_j elements in group j, where $\sum_{j=1}^{J} p_j = p$. The corresponding coefficients are represented in the $p_j \times 1$ vector $\beta_{\mathcal{I}_{jj}}$, where $\mathcal{I}_{jj}$ represents the set of all indecies of the form (s, s) that correspond to the main effects in group j. For notational simplicity, we may refer to $\mathcal{I}_{jj}$ as $\mathcal{J}$. There are an additional $\sum_{j>k=}^{J} p_j p_k$ interaction terms, for each pair of groups indexed by j and k, there is an additional group of $p_j p_k$ interaction terms. The corresponding coefficients are represented by the $p_j p_k \times 1$ vector $\beta_{\mathcal{I}_{jk}}$. similar to $\mathcal{I}_{jj}$, $\mathcal{I}_{jk} = \{(s, r) : (s, s) \in \mathcal{I}_{jj}, (r, r) \in \mathcal{I}_{kk}\}$, the set of indecies corresponding the individual interaction terms contained in the interaction

group indexed by j and k. For notational ease, we again refer to interaction groups using subscripts out of order; that is we may use either $\beta_{\mathcal{I}_{kj}}$ or $\beta_{\mathcal{I}_{jk}}$. To further ease notational simplicity, we may refer to $\mathcal{I}_{jk}$ as $\mathcal{J}\mathcal{K}$ and $\mathcal{I}_{kj}$ as $\mathcal{K}\mathcal{J}$. In cases where $\mathcal{J} = \mathcal{K}$ we note that $\mathcal{K}\mathcal{J} = \mathcal{J}\mathcal{K} = \mathcal{J}$. This yields the full $p + \sum_{j>k=}^{J} p_j p_k \times 1$ coefficent vector

$$\beta^\top = (\beta^\top_{\mathcal{I}_{11}}, \ldots, \beta^\top_{\mathcal{I}_{JJ}}, \beta^\top_{\mathcal{I}_{12}}, \ldots \beta^\top_{\mathcal{I}_{1J}}, \ldots, \beta^\top_{\mathcal{I}_{(J-1)J}})^\top$$

Similar notation can be applied to U as the two vectors share a correspondence and are of the same dimension. This gives the linear predictor for the rth entry in the sth cluster the form

$$\begin{aligned}\eta_{sr} &= \beta_0 + \sum_{j\in\{1,\ldots,J\}} X^\top_{sr\mathcal{I}_{jj}}\beta_{\mathcal{I}_{jj}} + \sum_{k\in\{1,\ldots,p\}}\sum_{j>k} X^\top_{sr\mathcal{I}_{kj}}\beta_{\mathcal{I}_{kj}} \\ &= \beta_0 + \sum_{j\in\{1,\ldots,J\}} X^\top_{sr\mathcal{J}}\beta_{\mathcal{J}} + \sum_{k\in\{1,\ldots,p\}}\sum_{j>k} X^\top_{sr\mathcal{J}\mathcal{K}}\beta_{\mathcal{J}\mathcal{K}},\end{aligned}$$

where X_{sr} is a $p + \sum_{j>k=}^{J} p_j p_k \times 1$ covariate vector for the rth entry in the sth cluster the includes both the main effects, denoted in groups as $X^\top_{sr\mathcal{I}_{jj}} = X^\top_{sr\mathcal{J}}$ for main effects and $X^\top_{sr\mathcal{I}_{kj}} = X^\top_{sr\mathcal{J}\mathcal{K}}$ for interaction terms.

4.2.1 Algorithm

Because of the non-overlapping structure of the main effect groups and the interaction effect groups regular penalty forms can be used. For example the group lasso and the

sparse group lasso penalties could be applied; however doing so would not preserve the interaction structure. Instead we propose modifications of the previous stagewise approaches such that the interaction structure is preserved.

If we are only considering a group level selection, then ACTS can easily be re-structured using the ℓ_2 norm in place of the ℓ_1 norm. The resulting algorithm would operate as follows: Let $\mathcal{A}^{[t]} = \{j : \beta^{[t]}_{\mathcal{J}} \neq 0\}$ and $\mathcal{A}^{2[t]} = \{(j,k) : j,k \in \mathcal{A}^{[t]}\}$, and consider the stagewise regression problem can be formulated in the following way:

$$\begin{aligned}
&(t.a)\ \text{Given } \beta^{[t-1]}, \text{ update the nuisance parameters to obtain } \nu^{[t]},\\
&(t.b)\ \delta^{[t]} = \underset{\substack{\delta \in \mathbb{R}^{p+\sum_{j>k}^{J} p_j p_k} \\ \delta_{\mathcal{J}\mathcal{K}}=0 \forall (j,k) \notin \mathcal{A}^{2[t]}}}{\arg\min} \langle U_{[0]}(\beta^{[t-1]}, \nu^{[t]}), \delta \rangle \text{ subject to } \phi(\delta) \leq \epsilon,\\
&(t.c)\ \beta^{[t]} = \beta^{[t-1]} + \delta^{[t]},
\end{aligned} \tag{4.1}$$

Where $\phi(\delta) = \sum_{k=j}^{J} \sum_{j=1}^{J} \|\beta_{\mathcal{J}\mathcal{K}}\|_2$, the ℓ_2 norm. The appropriate update as a result of the solution to $(t.b)$ will be of the form

$$\begin{aligned}
\delta^{[t]}_{\mathcal{I}_{i\ell}} &= -\epsilon / \|U_{\mathcal{I}_{i\ell}}\|_2,\\
\delta^{[t]}_{\mathcal{J}\mathcal{K}} &= 0, \quad \forall (k,j) \neq (i,\ell),
\end{aligned}$$

where

$$(i,\ell) = \underset{(j,k) \in \{(1,1),\ldots,(p,p)\} \cup \mathcal{A}^{2[t]}}{\arg\max} \|U_{\mathcal{J}\mathcal{K}}\|_2.$$

Such an approach does manage to harness the group structure while preserving the interaction hierarchy structure, but this approach suffers the same short comings of gLasso and group Stagewise estimating equations: the failure to address the bi-level selection problem. Additionally, this technique can quickly produce bloated models as there are many interaction terms in the interaction groups.

To address these issues, we modify the procedure in Equation (3.6) in the spirit of HiSEE by first using the grouping information to make a group selection as above, but then we only select an individual component of the group selected. Let $\beta_{\mathcal{JK}_{(sr)}}$ be the β_{sr} element of the coeffient vector of β if $(s,r) \in \mathcal{JK}$. Let $\mathcal{B}^{[t]} = \{s : \beta_{ss}^{[t]} \neq 0\}$ and $\mathcal{B}^{2[t]} = \{(s,t) : s,r \in \mathcal{A}^{[t]}\}$, and consider the stagewise regression problem can be formulated in the following way:

$$
\begin{aligned}
&(t.a)\ \text{Given } \beta^{[t-1]}, \text{ update the nuisance parameters to obtain } \nu^{[t]},\\
&(t.b)\ \delta^{(g)} = \underset{\substack{\delta \in \mathbb{R}^{p+\sum_{j>k}^{J} p_j p_k} \\ \delta_{\mathcal{JK}}=0 \forall (j,k) \notin \mathcal{A}^{2[t]}}}{\arg\min} \langle U_{[0]}(\beta^{[t-1]}, \nu^{[t]}), \delta \rangle \text{ subject to } \phi_1(\delta) \leq \epsilon, \text{ and let } \mathcal{G}^{[t]} = \{(j,k) : \delta_{\mathcal{JK}}^{(g)} \neq 0\}.\\
&(t.c)\ \delta^{[t]} = \underset{\substack{\delta \in \mathbb{R}^{p+\sum_{j>k}^{J} p_j p_k} \\ \delta_{\mathcal{JK}}=0 \forall (j,k) \notin \mathcal{G}}}{\arg\min} \langle U_{[0]}(\beta^{[t-1]}, \nu^{[t]}), \delta \rangle \text{ subject to } \phi_2(\delta) \leq \epsilon,\\
&(t.d)\ \beta^{[t]} = \beta^{[t-1]} + \delta^{[t]},
\end{aligned}
\tag{4.2}
$$

where $\phi_1(\delta) = \sum_{k=j}^{J} \sum_{j=1}^{J} w_{jk} \|\delta_{\mathcal{JK}}\|_2$, and $\phi_2(\delta) = \|\delta\|_1$

4.3 Future Work with Stagewise Techniques

The stagewise estimation and penalized estimation have strong connections and are both commonly used for sparse modeling. It would be interesting to explore general connections between penalized regression and stagewise methods in hierarchical variable selection settings. In particular, following Zhao and Yu (2007) and Wolfson (2011), it is promising to investigate the equivalence between the solution path generated by BiSEE and HiLa when $\epsilon \to \infty$ to that generated by sparse group lasso. Other theoretical concerns include estimation consistency (Hall and Severini, 1998; Wang et al., 2012) and variable selection consistency (Wang, 2009; Ing and Lai, 2011) in high-dimensional settings. Moreover, it is known that non-convex penalization methods can achieve better selection performance under weaker conditions compared to the convex method; it is thus worth studying the stagewise procedures that mimic non-convex methods (Huang et al., 2012). Last but not least, it is well known that for regularized estimation, inference based on simply refitting selected model without regularization could be invalid and misleading (Berk et al., 2013; Tibshirani et al., 2016; Lee et al., 2016). Statistical inferences based on stagewise estimators merits further investigation.

Appendix A

R package sgee

As part of the development of the stagewise techniques presented in this thesis, accompanying software to implement these techniques was also developed. This software can be found in the `R` package `sgee`, which is available on the Comprehensive R Archive Network (CRAN). This chapter will demonstrate the functionality and features of that `R` package.

A.1 Stagewise Implementation

All of the stagewise methods described in this thesis have been implemented as individual methods in `sgee`. Because the BiSEE approach described in Chapter 2 is a generalization of stagewise techniques involving lasso (SEE) and group lasso (gSEE), implementation of those methods are also included in the package. There are three main functions in `sgee`, `bisee`, `hisee`, and `isee`.

All of these functions are designed to either take a response vector and covariate matrix, or a data set and a formula, as in `glm` or `lm`, describing the desired relationship. Each of these methods is equipped to handle non-Gaussian clustered data with clusters of

differing sizes through the `family`, `clusterID`, and `waves` parameters. These functions may also take in a value for `offset` to add an offset value to the linear predictor. Additionally, a working correlation structure must be specified; currently "`independence`", "`exchangeable`", and "`ar1`" are implemented. Finally, all of these methods handle standardization so that the steps take in each iteration are fairly applied to effects of different scales. All of these approaches also implement the adaptive step size described in Chapter 3.

All of these functions take in a set of stagewise control values that can be set using the `sgee.control` function found in the package. Parameters set by this function include `maxIt`, the maximum number of iterations the stagewise procedure is to take; `epsilon`, the initial step size; `stoppingThreshold`, the predetermined maximum model size; and `undoThreshold` which controls how small the difference in subsequent steps is allowed to be before adjusting the step size.

For the grouped covariate techniques, `bise` and `hisee`, the user must supply an additional vector called `groupID` that identifies what group each covariate belongs to. When calling `bisee`, the function takes in two different lambda values that dictate the emphasis of the group versus the individuals. The function call is below.

```
bisee(formula, data=list(), family,
         clusterID, waves = NULL, groupID,
         corstr="independence", alpha = NULL,
         lambda1 = .5, lambda2 = 1-lambda1,
         intercept = TRUE,
         offset = 0,
```

```
        control = sgee.control(maxIt = 200, epsilon = 0.05,
                  stoppingThreshold =  min(length(y),
                  ncol(x))-intercept,undoThreshold = 0.005),
        standardize = TRUE,
        verbose = FALSE,
        ...)
```

The value for `lambda1` can be anywhere between 0 and 1, and while the same is true for `lambda2` it is recommended that it remain at its default value. When `lambda1` is exactly 0, the function reduces down to an implementation of Forward stagewise estimating equations, or SEE. When `lambda1` is exactly 1, the function reduces to group stagewise estimating equations, gSEE.

The second primary function in `sgee` is `hisee`, which also can be used to perform the bi-level selection task. Unlike `bisee`, `hisee` does not require any weighting between the groups and the individuals. The function call for `hisee` is very similar to that of `bisee`.

The third primary function of `sgee` is the `isee` function, which performs interaction selection. There are two main methods of interaction selection that `isee` can specify in the `method` parameter, either "`ACTS`" (default), or "`HiLa`". In order to execute the algorithm, both methods require a $2 \times p + \binom{p}{2}$ matrix that identifies the main effects that correspond to each term called the `interactionID`. When using a formula, the function automatically creates this parameter, but if the user instead supplies a response vector and matrix, the user must also supply the `interactionID`. Again, the function call is very similar to that of `bisee` and `hisee`.

A.2 Additional Features

In addition to the main functions described in Section A.1, `sgee` also contains additional functions that make the main functions more user-friendly. These functions make it easier for a user to 1) generate data to test the functions, 2) analyze a given solution path, and 3) visualize a solution path.

The first supporting function is the `genData` function, which has the following function call:

```
genData(numClusters,
        clusterSize = 1,
        clusterRho = 0,
        clusterCorstr = "exchangeable",
        yVariance = NULL,
        xVariance = 1,
        numGroups = 1,
        groupSize = 1,
        groupRho = 0,
        beta = 0,
        numMainEffects = NULL,
        family = gaussian(),
        SNR = NULL,
        intercept = 0)
```

The `genData` function allows the user to easily generate the kind of data that the `sgee` functions were designed for: non-Gaussian clustered data. While able to also generate Gaussian clustered data through the `mvtnorm` package, `sgee` focuses on generating non-Gaussian clustered data using copula from the `copula` package. The user can specify the

marginal distributional family (`family`), the number of clusters (`numClusters`), the cluster size (`clusterSize`), the kind of correlation structure (`clusterCorr`), the marginal variance of the response (`yVariance`) and the individual covariates (`xVariance`), and the true coefficient values (`beta` and `intercept`). Additionally, the function can generate covariate with an overlying grouping structure as described in Chapter 2 using the `groupSize` parameter, and the `groupRho` parameter to determine the amount of within group correlation. Finally, an interaction model can be generated through the use of the `numMainEffects` parameter.

When any of the main methods is called to generate a solution path, what is returned is an object of class `sgee`. The `summary` call on an object of this class produces an analysis of that solution path based on given user input. The function is called in the following way:

```
summary(object,
        newX = NULL,
        newY = NULL,
        newOffset = NULL,
        trueBeta = NULL,
        trueIntercept = NULL,
        scale = NULL,
        averaged = TRUE,
        ...)
```

The primary purpose of the `summary` function is to identify the optimal point in the solution path based on predictive performance. When the summary function is called on an `sgee` object without additional parameters, the original data set is used to evaluate

the predictive performance of each point along the solution path. If the parameters `newX`, `newY`, and `newOffset` are specified, out-of-set prediction can be done. Additionally, if the true model can be supplied via `trueBeta` and `trueIntercept`, as in the case of simulation, model selection criteria such as the false positive and false negative rates can also be calculated.

Finally, `sgee` provides a `plot` function that produces a coefficient trace plot when called on an object of the `sgee` class. The function call is:

```
plot(x,
     y,
     penaltyFun = NULL,
     main = NULL,
     xlab = "Iterations",
     ylab = expression(beta),
     dropIntercept = FALSE,
     trueBeta = NULL,
     color = TRUE,
     manualLineColors = NULL,
     pointSpacing = 3,
     cutOff = NULL,
     ...)
```

The default behavior of the `plot` function call is to simply plot the individual coefficient values against the iteration number. If however a penalty function is provided for the parameter `penaltyFun`, the coefficients can be plotted against penalty function values such as the ℓ_1-norm. Additionally, if the true model structure can be provided for `trueBeta`, then the individual coefficient paths that are erroneously non-zero can be marked. Finally, a logical value `color` can change the settings of the plot to either make

a color plot or a plot fit for black and white printing.

A.3 Demonstration

This section will highlight some of the features of `sgee` by doing two walkthroughs of a single simulation run where data is first generated, then processed by a stagewise technique, and then finally analyzed.

A.3.1 Grouped Covariates

The first demonstration will walk through an example with grouped covariates. We have twenty covariates, forming five groups each of size 4. We have an intercept value of 1, with four coefficients with a value of 2, one coefficients with a size of 1, and five coefficients with a size of 0.5; the rest are zero.

```
## Initialize covariate values
p <- 20
beta <- c(rep(2,4),
          c(1, 0, 0, .5),
          rep(0.5,4),
          rep(0,p-12))
interceptValue <- 1
groupSize <- 4
numGroups <- length(beta)/groupSize
```

After defining our parameters we use the `genData` function to generate a data set of the desired form where we have an exchangeable correlation structure with a correlation

parameter of 0.6 and a marginal response variance of 1. Additionally, the variance of the covariates will be zero (mean of 0 is assumed) with a within group exchangeable correlation (fixed structure) with a correlation parameter of 0.3.

```
trainingData <- genData(numClusters = 50, clusterSize = 4,
                        clusterRho = 0.6,
                        clusterCorstr = "exchangeable",
                        yVariance = 1, xVariance = 1,
                        numGroups = numGroups,
                        groupSize = groupSize,
                        groupRho = 0.3,
                        beta = beta, intercept = interceptValue,
                        family = gaussian())
```

After generating our data we use `hisee` to construct a path of model estimates. In this example, we demonstrate the ability to subset the data by only selecting response values that are less than 1. when we do this, we must specify the `waves` parameter value so that the function knows where each observation falls in its respective cluster have removing some observations.

```
## Perform Fitting by providing formula and data
genDF <- data.frame(trainingData$y, trainingData$x)
names(genDF) <- c("Y", paste0("Cov", 1:p))
coefMat <- hisee(formula(genDF), data = genDF,
                 family = gaussian(),
                 subset = Y<1, waves = rep(1:4, 50),
                 clusterID = trainingData$clusterID,
                 groupID = trainingData$groupID,
                 corstr="exchangeable",
                 control = sgee.control(maxIt = 200,
                                        epsilon = 0.1))

##
```

```
## [1] "stopped on"
## [1] 114
## [1] 5.044267
```

After performing a model fitting, we can call `summary` to identify an optimal point in the path based on predictive performance and the corresponding coefficient estimates.

```
summary(coefMat)
```

```
##
## newX and/or newY missing; original data used
## Call:
## NULL
##
## Lowest Predictive Error: 0.6632039
## Achieved at index: 105
## With corresponding Coefficients:
##                Estimate
## (intercept)  0.83878477
## Cov1         1.87958776
## Cov2         1.96822433
## Cov3         1.95990709
## Cov4         1.93505440
## Cov5         1.02815395
## Cov6         0.00000000
## Cov7        -0.10472760
## Cov8         0.33641621
## Cov9         0.53291541
## Cov10        0.49160014
## Cov11        0.42443227
## Cov12        0.45332062
## Cov13        0.00000000
## Cov14        0.00000000
## Cov15       -0.09193923
## Cov16        0.00000000
## Cov17        0.00000000
## Cov18       -0.20457105
## Cov19        0.10389118
## Cov20        0.00000000
```

Finally, we use `plot` to generate a coefficient trace plot of our solution path.In this example we provide the ℓ_1 norm for the penalty function parameter `penaltyFun`. The allows us to plot the coefficient paths not as a function of the iteration, but rather as a function of the overall model size, as measured by the ℓ_1 norm. We use the `dropIntercept` parameter so that only the non-intercept values are plotted. The plot can be found in Figure 13.

```
plot(coefMat,
     penaltyFun = function(x){sum(abs(x))},
     xlab = "L1 Norm",
     dropIntercept = TRUE)
```

A.3.2 Interaction Selection

The second demonstration is an example of interaction selection. We have only five covariates, which gives us five main effects and $\binom{5}{2} = 10$ interaction effects. We again have an intercept value of 1, with four coefficients with main effect values of 1, 1.5, and 0.5, and six interaction effects of size 0.5 and one interaction effect of size 1. The coefficients are listed in the following order: main effects first, interaction effect with the first covariate, interaction effects with the second covariate, and so on. Note, because of which interaction effects are non zero, the true model only adheres to a weak hierarchy.

```
## Initialize covariate values
p <- 5
beta <- c(1, 0, 1.5, 0, 0.5, ## Main effects
```

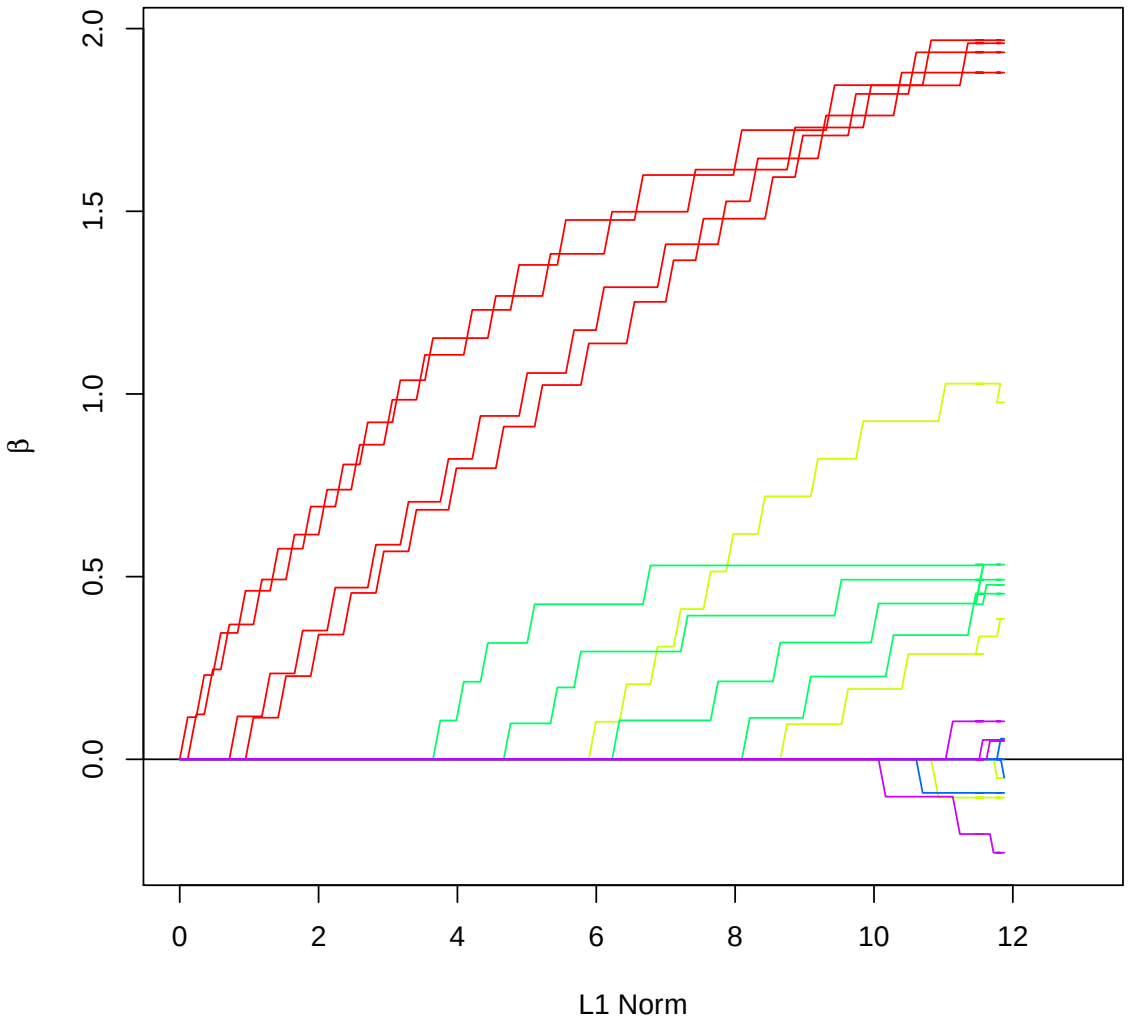


Figure 13: Grouped covariate selection coefficient trace plot. Each line represents the value of a particular coefficient for a particular ℓ_1 norm of the whole model.

```
          rep(0.5,4), ## Interaction terms
          0.5, 0, 0.5,
          0,1,
          0)
interceptValue <- 1
```

Again we use the `genData` function to generate a data set to train our model on. The basic settings remain the same as in the first demonstration. When generating data with interaction terms, `genData` assumes that `beta` includes all interaction effects; therefore, we must use the parameter `numMainEffects` to indicate how many of the effects in `beta` are main effects, and how many are interaction effects.

```
trainingData <- genData(numClusters = 50,
                        clusterSize = 4,
                        clusterRho = 0.6,
                        clusterCorstr = "exchangeable",
                        yVariance = 1,
                        xVariance = 1,
                        beta = beta,
                        numMainEffects = p,
                        family = gaussian(),
                        intercept = interceptValue)
```

This time, we will produce a second testing data set to use with the summary function.

```
testingData <- genData(numClusters = 5*50,
                        clusterSize = 4,
                        clusterRho = 0.6,
                        clusterCorstr = "exchangeable",
                        yVariance = 1,
                        xVariance = 1,
```

```
                          beta = beta,
                          numMainEffects = p,
                          family = gaussian(),
                          intercept = interceptValue)
```

For this example we use `isee` to construct a path of model estimates with the `method` parameter set to "`ACTS`".

```
## Perform Fitting by providing formula and data
genDF <- data.frame(Y = trainingData$y, X = trainingData$xMainEff)

## Using "ACTS" method
coefMat <- isee(formula(paste0("Y~(",
                               paste0("X.", 1:p, collapse = "+"),
                                 ")^2")),
                data = genDF,
                family = gaussian(),
                clusterID = trainingData$clusterID,
                corstr = "exchangeable",
                method = "ACTS",
                control = sgee.control(maxIt = 200,
                                       epsilon = 0.05))
```

Because in this example we have access to an out of training data set, we can identify the best point along the path based on out of sample prediction, rather than in sample prediction, which can lead to over fitting the data.

```
summary(coefMat,
        newX = testingData$x,
        newY = testingData$y)

## Call:
## isee.formula(formula = formula(paste0("Y~(", paste0("X.", 1:p,
##     collapse = "+"), ")^2")), data = genDF, clusterID = trainingData$clusterID,
```

```
##     method = "ACTS", family = gaussian(), corstr = "exchangeable",
##     control = sgee.control(maxIt = 200, epsilon = 0.05))
##
## Lowest Predictive Error: 1.154589
## Achieved at index: 200
## With corresponding Coefficients:
##                 Estimate
## (intercept)  1.082212062
## X.1          0.953119148
## X.2          0.059811060
## X.3          1.560262091
## X.4         -0.005978564
## X.5          0.425598724
## X.1:X.2      0.555574007
## X.1:X.3      0.369532651
## X.1:X.4      0.185057234
## X.1:X.5      0.436368288
## X.2:X.3      0.546038786
## X.2:X.4      0.000000000
## X.2:X.5      0.446051008
## X.3:X.4      0.000000000
## X.3:X.5      0.980397952
## X.4:X.5      0.000000000
```

We conclude with a coefficient trace plot of our solution path. In this example we provide the true model, so that we can see which coefficients `isee` erroneously picked up, and when. The plot can be found in Figure 14.

```
plot(coefMat,
     trueBeta = beta,
     trueIntercept = interceptValue)
```

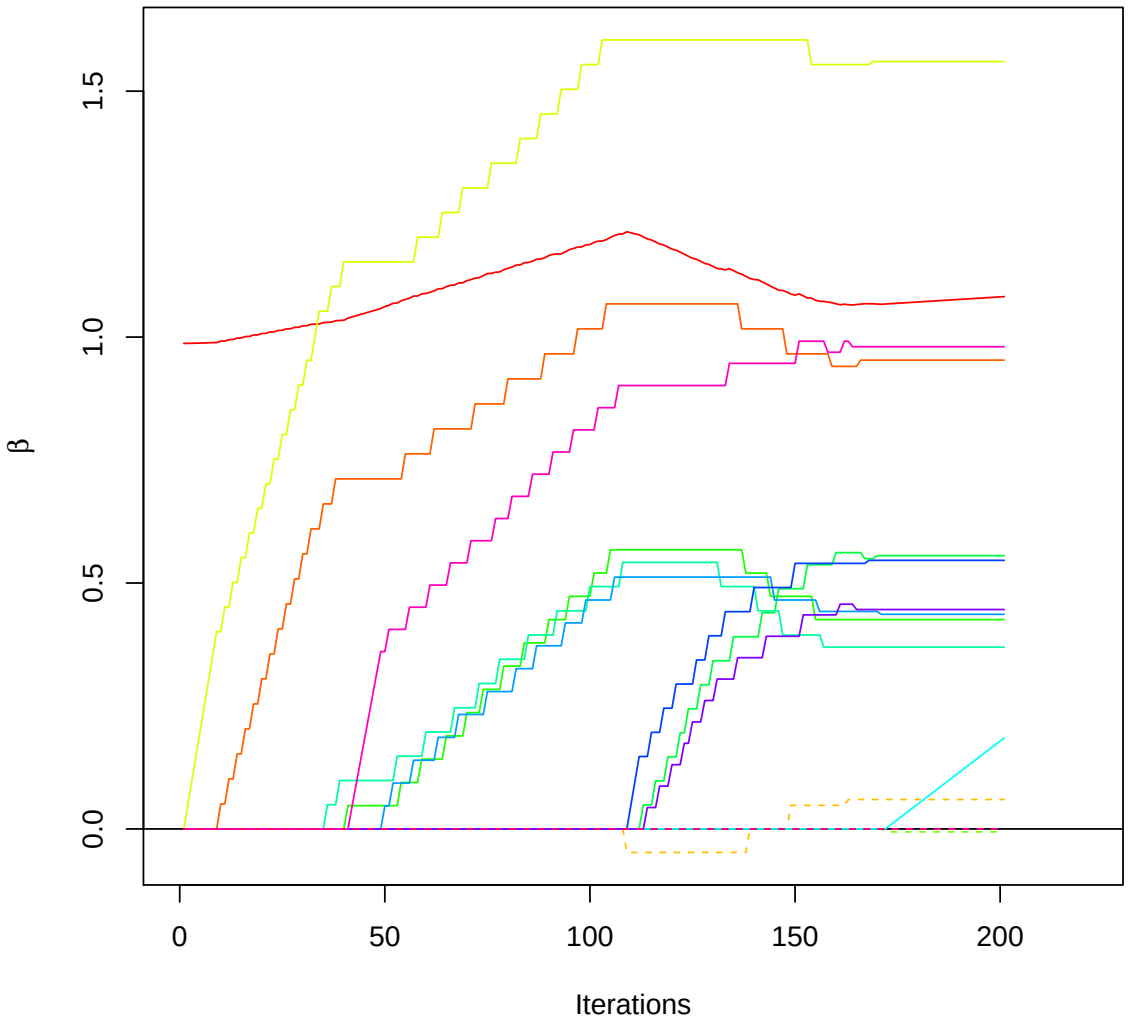


Figure 14: Interaction selection coefficient trace plot. Each line represents the value of a particular coefficient for a particular iteration in the algorithm. Groups Share line colors. Truly non-zero coefficients have solid lines; truly non-zero coefficients have dashed lines.

Bibliography

Bach, F., Jenatton, R., Mairal, J., and Obozinski, G. (2012). Structured sparsity through convex optimization. *Statistical Science* **27,** 450–468.

Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013). Valid post-selection inference. *The Annals of Statistics* **41,** 802–837.

Bien, J., Taylor, J., and Tibshirani, R. (2013). A lasso for hierarchical interactions. *The Annals of Statistics* **41,** 1111–1141.

Breheny, P. and Huang, J. (2009). Penalized methods for bi-level variable selection. *Statistics and Its Interface* **2,** 369–380.

Breiman, L. (1998). Arcing classifier. *The Annals of Statistics* **26,** 801–849.

Bühlmann, P. and Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science* **22,** 477–505.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Science & Business Media.

Chen, K. and Aseltine, R. H. (2017). Using hospitalization and mortality data to identify areas at risk for adolescent suicide. *Journal of Adolescent Health* In Press.

Chen, K., Hoffman, E. A., Seetharaman, I., Lin, C.-L., and Chan, K.-S. (2016). Linking lung airway structure to pulmonary function via composite bridge regression. *Annals of Applied Statistics* **10,** 1880–1906.

Deshpande, V., Dey, D. K., and Schifano, E. D. (2016). Variable selection for correlated bivariate mixed outcomes using penalized generalized estimating equations. *Department of Statistics, University of Connecticut* In Press.

Draper, N. R. and Smith, H. (1998). *Applied Regression Analysis*. John Wiley & Sons.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. (2004). Least angle regression. *The Annals of Statistics* **32,** 407–499.

Fan, J., Han, F., and Liu, H. (2014). Challenges of big data analysis. *National Science Review* **1,** 293–314.

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

Friedman, J., Hastie, T., and Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting. *The Annals of Statistics* **28,** 337–407.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). A note on the group lasso and a sparse group lasso. arXiv preprint arXiv:1001.0736.

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

Halekoh, U., Højsgaard, S., and Yan, J. (2006). The R package geepack for generalized estimating equations. *Journal of Statistical Software* **15/2,** 1–11.

Hall, D. B. and Severini, T. A. (1998). Extended generalized estimating equations for clustered data. *Journal of the American Statistical Association* **93,** 1365–1375.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York.

Huang, J., Breheny, P., and Ma, S. (2012). A selective review of group selection in high dimensional models. *Statistical Science* **27,** 481–499.

Huang, J., Ma, S., Xie, H., and Zhang, C.-H. (2009). A group bridge approach for variable selection. *Biometrika* **96,** 339–355.

Ing, C.-K. and Lai, T. L. (2011). A stepwise regression method and consistent model selection for high-dimensional sparse linear models. *Statistica Sinica* **21,** 1473–1513.

Jenatton, R., Audibert, J.-Y., and Bach, F. (2011). Structured variable selection with sparsity-inducing norms. *Journal of Machine Learning Research* **12,** 2777–2824.

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics* **44,** 907–927.

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

Liang, K.-Y., Zeger, S. L., and Qaqish, B. (1992). Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society: Series B* **54,** 3–40.

Lim, M. and Hastie, T. (2015). Learning interactions via hierarchical group-lasso regularization. *Journal of Computational and Graphical Statistics* **24,** 627–654. PMID: 26759522.

Liu, J., Huang, J., Zhang, Y., Lan, Q., Rothman, N., Zheng, T., and Ma, S. (2013). Identification of geneenvironment interactions in cancer studies using penalization. *Genomics* **102,** 189 – 194.

Liu, J., Ma, S., and Huang, J. (2014). Integrative analysis of cancer diagnosis studies with composite penalization. *Scandinavian Journal of Statistics* **41,** 87–103.

Ma, S., Huang, J., Wei, F., Y., X., and Fang, K. (2011). Integrative analysis of multiple cancer prognosis studies with gene expression measurements. *Statistics in Medicine* **30,** 3361–3371.

Meier, L., van de Geer, S., and Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B* **70,** 53–71.

Miller, A. (2002). *Subset Selection in Regression*. CRC Press.

Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics* **57,** 120–125.

Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics* **44,** 1033–1048.

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

Schapire, R. E. and Freund, Y. (2012). *Boosting: Foundations and Algorithms*. MIT press.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013a). *SGL: Fit a GLM (or Cox Model) with a Combination of Lasso and Group Lasso Regularization*. R package version 1.1.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013b). A sparse-group lasso. *Journal of Computational and Graphical Statistics* **22,** 231–245.

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

Tibshirani, R. J. (2015). A general framework for fast stagewise algorithms. *Journal of Machine Learning Research* **16,** 2543–2588.

Tibshirani, R. J., Taylor, J., Lockhart, R., and Tibshirani, R. (2016). Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association* **111,** 600–620.

Vaughan, G., Aseltine, R., Chen, K., and Yan, J. (2017). Stagewise generalized estimating equations with grouped variables. *Biometrics* In Press.

Wang, H. (2009). Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association* **104,** 1512–1524.

Wang, L., Chen, G., and Li, H. (2007). Group SCAD regression analysis for microarray time course gene expression data. *Bioinformatics* **23,** 1486–1494.

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

Wolfson, J. (2011). EEBoost: A general method for prediction and variable selection based on estimating equations. *Journal of the American Statistical Association* **106,** 296–305.

Yan, J. and Fine, J. (2004). Estimating equations for association structures. *Statistics in Medicine* **23,** 859–874.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B* **68,** 49–67.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38,** 894–942.

Zhao, P., Rocha, G., and Yu, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *Ann. Statist.* **37,** 3468–3497.

Zhao, P. and Yu, B. (2007). Stagewise lasso. *Journal of Machine Learning Research* **8,** 2701–2726.

Zhu, R., Zhao, H., and Ma, S. (2014). Identifying gene-environment and gene-gene interactions using a progressive penalization approach. *Genetic Epidemiology* **38,** 353–368.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101,** 1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67,** 301–320.