Variable screening and graphical modeling for ultra-high dimensional longitudinal data

Yafei Zhang

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Statistics

Pang Du, Chair
Yili Hong
Inyoung Kim
Xiaowei Wu

July 1, 2019
Blacksburg, Virginia

Keywords: Variable screening, graphical modeling, longitudinal data

Variable screening and graphical modeling for ultra-high dimensional longitudinal data

Yafei Zhang

(ABSTRACT)

Ultrahigh-dimensional variable selection is of great importance in the statistical research. And independence screening is a powerful tool to select important variable when there are massive variables. Some commonly used independence screening procedures are based on single replicate data and are not applicable to longitudinal data. This motivates us to propose a new Sure Independence Screening (SIS) procedure to bring the dimension from ultra-high down to a relatively large scale which is similar to or smaller than the sample size. In chapter 2, we provide two types of SIS, and their iterative extensions (iterative SIS) to enhance the finite sample performance. An upper bound on the number of variables to be included is derived and assumptions are given under which sure screening is applicable. The proposed procedures are assessed by simulations and an application of them to a study on systemic lupus erythematosus illustrates the practical use of these procedures. After the variables screening process, we then explore the relationship among the variables. Graphical models are commonly used to explore the association network for a set of variables, which could be genes or other objects under study. However, graphical modes currently used are only designed for single replicate data, rather than longitudinal data. In chapter 3, we propose a penalized likelihood approach to identify the edges in a conditional independence graph for longitudinal data. We used pairwise coordinate descent combined with second order cone programming to optimize the penalized likelihood and estimate the parameters. Furthermore, we extended the nodewise regression method the for longitudinal data case. Simulation and real data analysis exhibit the competitive performance of the penalized likelihood method.

Variable screening and graphical modeling for ultra-high dimensional longitudinal data

Yafei Zhang

(GENERAL AUDIENCE ABSTRACT)

Longitudinal data have received a considerable amount of attention in the fields of health science studies. The information from this type of data could be helpful with disease detection and control. Besides, a graph of factors related to the disease can also be built up to represent their relationships between each other. In this dissertation, we develop a framework to find out important factor(s) from thousands of factors in longitudinal data that is/are related to the disease. In addition, we develop a graphical method that can show the relationship among the important factors identified from the previous screening. In practice, combining these two methods together can identify important factors for a disease as well as the relationship among the factors, and thus provide us a deeper understanding about the disease.

Acknowledgments

First and foremost, I would fisrt like to express my thankfulness to my supervisor, Dr. Pang Du, whose immense knowledge, encouragement, patience, and enthusiasm are invaluable to my Ph.D. study and research. His expertise helped me with formulation research topic, methodology, and writing of this dissertation. Besides my advisor, I would like to acknowledge the rest of my thesis committee: Dr. Yili Hong, Dr. Inyoung Kim, and Dr. Xiaowei Wu, for their wonderful motivation, guidance, and insightful comments. I would also like to thank all professors and staff members in the Department of Statistics in Virginia Tech for their help. I am grateful to all my fellow students who have offered encouragement, assistance, and friendship during my course study. Last but not the least, I want to thank my family, my wife and my parents, for their love and consistent support.

Contents

List of Figures viii

List of Tables ix

1 Introduction 1

1.1 Motivation example . . . 2

1.2 Variable selection for high dimensional data . . . 3

1.2.1 Classical High Dimensional Variable Selection . . . 3

1.2.2 Ultra-high Dimensional Variable Screening . . . 7

1.2.3 Other Independence Screening . . . 11

1.2.4 Variable selection for linear mixed-effect model . . . 18

1.3 Graphical modeling . . . 19

1.4 Motivation for proposal work . . . 23

2 SIS for ultra-high dimensional Longitudinal Data 25

2.1 Introduction . . . 26

2.2 Method . . . 30

2.2.1 Notation and Model . . . 30

2.2.2 SIS for longitudinal data . . . 31

2.2.3 Iterative SIS for longitudinal data . . . 33
2.3 Theoretical property . . . 34
2.3.1 Equivalence between p-value screening and MMLE screening . . . 34
2.3.2 Sure screening property . . . 35
2.3.3 Uniform convergence of MMLE . . . 36
2.3.4 Selection of threshold parameter in MMLE . . . 38
2.3.5 p-value screening property . . . 39
2.4 Numerical Analysis . . . 41
2.4.1 Simulation . . . 41
2.4.2 SLE Data . . . 43
2.5 Discussion . . . 45

3 Graphical modeling for longitudinal data 49

3.1 Introduction . . . 50
3.2 Method . . . 52
3.2.1 Models and notation . . . 52
3.2.2 Neighborhood selection . . . 53
3.2.3 SOCP estimation . . . 55
3.2.4 Nodewise LASSO . . . 57
3.3 Numerical Analysis . . . 61

3.3.1 Simulation 62
3.3.2 Real data example 63
3.4 Discussion 64

Appendices **67**

Appendix A Appendix for SIS **68**

A.1 Proof of Theorem 1 69
A.2 Proof of Theorem 2 70
A.3 Proof of Theorem 3 70
A.4 Proof of Theorem 4 72
A.5 Proof of Theorem 5 72

Bibliography **79**

List of Figures

3.1 Networks identified for the SLE data by nodewise step method, SOCP method and nodewise LASSO method 66

List of Tables

2.1 MMMS, and its RSD (in parentheses), FP, and FN over 200 replicates for Setting 1, $\epsilon \sim \mathcal{N}(0,1), N = 200$ 43

2.2 MMMS, and its RSD (in parentheses), FP, and FN over 200 replicates for Setting 2, $\epsilon \sim \mathcal{N}(0,1), N = 200$ 44

2.3 MMMS, and its RSD (in parentheses), FP, and FN over 200 replicates for Setting 1, $\epsilon \sim \mathcal{N}(0,5), N = 800$ 45

2.4 MMMS, and its RSD (in parentheses), FP, and FN over 200 replicates for Setting 2, $\epsilon \sim \mathcal{N}(0,5), N = 800$ 46

2.5 The number of selected variable ($|S|$) and MSE by four types of screening methods. 47

3.1 Comparison of average (standard deviation) over 200 replications with $p = 10$ and $n = 200$ 63

3.2 Comparison of average (standard deviation) over 200 replications with $p = 25$ and $n = 500$ 64

3.3 Comparison of average (standard deviation) over 200 replications with $p = 40$ and $n = 800$ 65

Chapter 1

Introduction

1.1 Motivation example

The example we use in this dissertation is from an immunomonitoring study [3] on systemic lupus erythematosus (SLE). SLE is a chronic disease that affects young women with a breakdown of nucleic acid tolerance and highly varied clinical manifestations. In this study, children and adolescents with SLE have been enrolled at the Texas Scottish Rite Hospital for Children and Children's Medical Center Dallas from the Rheumatology clinics. Study procedures followed protocols approved by the Institute Review Boards of the Southwestern Medical Center University of Texas and the Medical Center of Baylor University. During routine morning clinic visits every three months, patients were evaluated by a standardized protocol and more frequently if clinical symptoms justified re-evaluation. Blood has been collected for flowcytometry studies and laboratory measurement in Tempus blood RNA tubes (Life Technologies) for the microarray and in the Citrate Dextrose Acid (ACD) tubes.

The blood transcriptome of 158 patients was collected and profiled for up to 1,412 days, representing 924 visits, in order to assess its molecular heterogeneity. SLE Disease Activity Index (SLEDAI) for each patient, which is a weighted metric combining 24 components, was measured as well as 43,739 gene expressions measured at each visit. The number of visits per patient ranged from 1 to 22 (The dataset described in this dissertation is deposited in the NCBI Gene Expression Omnibus under GEO Series accession number GSE65391). The repeated measures within each subject (patient) warrants the need for a longitudinal data analysis approach. What we shall explore in this dissertation is the relationship between SLEDAI and gene expressions, as well as the interaction among gene regulatory relationships. As reviewed below, despite the rich development on variable selection and feature screening procedures for high dimensional data, there has been little work on any similar development under the longitudinal data setting. Therefore, in this dissertation, I will first propose a feature screening procedure for longitudinal data to reduce the dimension of im-

portant immune response factors and then design a graphical model for longitudinal data to understand the interaction between these factors.

1.2 Variable selection for high dimensional data

Variable selection in high dimensional statistical modeling is of great importance for improving performance of linear regressions, which have been substantially studied and widely used in many areas of scientific discoveries. Some classical variable selection methods, for example Akaike information criterion (AIC) [1], Bayesian information criterion (BIC) [57], Mallow's C_p [48], as well as regularization methods involve NP-hard combinatorial optimization problem. These traditional methods grant a top notch interpretation of best subset selection and admit nice sampling properties. However, with large dimensionality of p, estimation accuracy and computational cost make the statistical procedures infeasible for the analysis of high dimensional data. To solve this problem, Fan and Lv [21] developed the sure independent screening (SIS) and iterative SIS. SIS and iterative SIS take the super high dimension feature space down to a much smaller feature space, which could be then solved by, for example, LASSO [62]. However, these methods mentioned are all for non-longitudinal data. For high dimensional longitudinal data analysis, other variable selection should be used.

1.2.1 Classical High Dimensional Variable Selection

Consider a linear regression

$$y = X\beta + \epsilon \tag{1.1}$$

where y is a n-vector observations of response variable Y, X is an $n \times p$ matrix consisting of p predictors $X_1, X_2, \ldots, X_p$, and $\beta = (\beta_1, \ldots, \beta_p)^T$ is p-dimensional coefficient vector, and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is the random error. Suppose the response variable is centered and the predictor variables are standardized with mean zero and standard deviation one. A generalized form of the penalized least squares is

$$||y - X\beta||^2 + \sum_{j=1}^{p} p_\lambda(\beta_j), \tag{1.2}$$

where p_λ is a penalty function on individual coefficient with penalty parameter(s) λ. A popular generalization of penalty p_λ is penalized L_q regularization, referred as the bridge regression [28], in which $p_\lambda(t) = \lambda|t|^q$ for $0 < q \leq 2$, which refers to the best subset selection (penalized L_0 regularization) and Ridge regression (penalized L_2 regularization). In particular, the well-known LASSO is the penalized L_1 regression.

There are three properties for the penalty function [20] are considered by Fan and Li [19]:

1. Sparsity: The estimator computed from the loss function leads small estimated coefficients to zero.

2. Unbiasedness: The estimator is nearly unbiased.

3. Continuity: The estimator is continuous in the data to lower the variation of the model prediction.

Specifically, the convex L_q penalty does not meet the sparsity condition, when $q > 1$. The concave L_q penalty does not satisfy the continuity condition, when $0 \leq q < 1$. Particularly, the large coefficients are usually biased for LASSO estimator, which makes Fan and Li [19]

to develop the smoothly clipped absolute deviation(SCAD) penalty

$$p'_\lambda(t) = \lambda \left\{ I(t \leq \lambda) + \frac{(a\lambda - t)_+}{(a-1)\lambda} I(t > \lambda) \right\} \quad for\ some\ a > 2, \tag{1.3}$$

where $p_\lambda(0) = 0$ and, $a = 3.7$ is commonly applied. The SCAD penalty meets all the above three properties and the estimator has the oracle property [19]. Fan and Li [19] also developed an efficient and unified local quadratic approximation (LQA) algorithm, since the SCAD penalty is nonconcave, which is challenging to be optimized and the involving computation is difficult. The main thought is to use a quadratic function to locally approximate the objective function. Zou and Li [80] proposed to use a local linear approximation, which borrows the strength from LARS.

Efron et al. [16] proposed a least angle regression (LARS) algorithm to calculate the whole solution path of the LASSO according to the piecewise linear path of the LASSO solution. And the LARS algorithm can also be extended to solve the penalized least squares problems. Fu [31], Daubechies et al. [13], Osborne et al. [53], and Wu and Lang [69] proposed the coordinate descent algorithm, which is very fast and efficient for large LASSO problems, which can also be used to solve the group LASSO [72] problem as shown in Meier et al. [50]. Bach [2] and Nardi and Rinaldo [52] have already explored the asymptotic properties of the group LASSO. Some other approaches at group level were also proposed by Wang et al. [65], Kim et al. [37], and Zhao et al. [75].

Zhang[74] developed a minimax concave penalty (MCP), which is

$$p'_\lambda(t) = (a\lambda - t)_+ / a, \tag{1.4}$$

and proved that the resulting estimator also has the oracle property. Furthermore, a penalized linear unbiased selection (PLUS) procedure was developed to solve for the MCP.

Zou [78] proposed a weighted version of L_1 penalty, which is called adaptive LASSO, to solve the problem of the lack of oracle property of the LASSO estimator.

$$J_\lambda(\beta) = \lambda \sum_{j=1}^{p} \frac{|\beta_j|}{|\hat{\beta_{init,j}}|^\gamma} \tag{1.5}$$

where the $\gamma > 0$ is a pre-selected constant and the $\hat{\beta_{init,j}}$ are root-n-consistent initial estimates of β. Under some regularity conditions, the resulting adaptive LASSO estimator has the oracle property [35], but the penalty at zero is infinite. But for some penalty functions, SCAD and MCP for instance, do not possess this undesired property. The estimates of adaptive LASSO solution can be obtained by using the same procedures for LASSO.

The elastic net (ENet) which was proposed by Zou and Hastie [79] is a combination of the L_1 and L_2 penalties:

$$p_\lambda(t) = \lambda_1|t| + \lambda_2 t^2, \tag{1.6}$$

where the L_1 penalty refers to the sparsity in the coefficients and the L_2 penalty encourages some grouping effects. Similarly, Liu and Wu [45] developed a L_0L_1 penalty, which combines the L_0 and L_1 penalties:

$$p_\lambda(t) = (1 - \lambda_1)\min\{|t|/\lambda_2, 1\} + \lambda_1|t|. \tag{1.7}$$

This L_0L_1 penalty overcomes the disadvantages of the L_0 and L_1 penalties. While, Wu et al. [68] proposed a procedure that is a combination of the L_1 and L_∞ penalties:

$$J_\lambda(\beta) = \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_\infty ||\beta||_\infty, \tag{1.8}$$

where $||\beta||_\infty = \max_{1\le j\le p}|\beta_j|$. The L_∞ penalty leads to group among highly correlated predictors.

1.2.2 Ultra-high Dimensional Variable Screening

Those regularization procedures mentioned above are able to deal with high dimensional data that p is as large as sample size n, but may have problems when p is much larger than n, i.e., p is in an exponential order $\exp\{O(n^{\alpha})\}$, $\alpha > 0$ of n. To deal with the ultra high dimensionality, an attractive idea is that one method that is fast and efficient is used to reduce the huge dimensionality p to a relatively large scale d (e.g., $O(n^{b})$ for some $b > 0$), which is close to or smaller than the sample size, then the regularization procedures can be used to further reduced feature space. This indicates a two-step method: a crude large scale screening and a moderate scale selection [21]. Many screening techniques can be chosen in the first step, once screening property introduced by Fan and Lv[21] is met, such that the variables screened in the first step contains all the important variables.

Sure Independence Screening

The main theme behind independence screening is as follows: each feature is used independently as a predictor for predicting the response and, subsequently, those features which appear highly related to the response are selected. In linear regression, the marginal correlation coefficient could be used as a case of a measurement of association between the predictor variables and the response variable. Fan and Lv [21] proposed to rank the features based on the magnitude of its sample correlation with the response variable. Let $\omega = (\omega_1, \ldots, \omega_p)^T = X^T y$ denote the p-vector calculated by componentwise regression. And each column of the design matrix X, whose dimension is $n \times p$, is standardized with mean zero and variance one. For any predefined d_n, select the predictor variables satisfying

$$\mathcal{M}_{d_n} = \{1 \leq i \leq p : |\omega_i| \text{ is among the first } d_n \text{ largest of all}\}.$$

and d_n is suggested to be $[n/\log n]$ in the paper as a conservative practical choice. Such correlation learning kicks out those predictor variables that are weakly marginally correlated with the response variable. SIS has been proved to have the sure screening property, which means that with high probability (tending to 1) it is able to detect a subset of covariates which contains the important ones and its size is much smaller than p. After the huge scale dimensionality p is reduced to a relatively moderate scale d, other variable selection methods, for example Lasso and elastic net, can be used to further select variables for parameter estimation and predicting. The sure screening methods sampling properties can be achieved by merging the sure independent screening property and theory of penalization methods [22]. However, the screening procedure could also have poor performance with some invalid key conditions.For instance, important variables that are jointly correlated but marginal uncorrelated with the response may not be selected by SIS, which in this condition is more likely to select the unimportant variables which are jointly uncorrelated but marginally correlated with the response variable. To overcome these potential problems, Fan and Lv [21] also proposed an iterative SIS procedure by replacing the response variable with the residual obtained in the previous step by the regressing the response and selected variables; see Section 1.2.2 for more details. Wang [64] proposed to use the extended BIC [10] to define the active predictor set size and explored the forward regression property with the ultrahigh-dimensional predictors. Hall and Miller [32] proposed to use the generalized correlation and rank the predictor variables according to the magnitude of estimated generalized correlation coefficients. Li et al. [40] developed a robust rank correlation screening (RRCS) approach by ranking the Kendall correlation to handle some heavy-tail distributions. And the RRCS approach is robust to influence points and outliers in the samples, which is not the case for the Pearson correlation in SIS procedure.

Since Fan and Lv [21] proposed SIS for ultra high dimensional variable selection, many

authors further extended the SIS procedure to various statistical methodologies.

Iterative Sure Independence Screening

Fan and Lv [21] pointed out SIS has three possible problems: (1) SIS may not select the important predictor variable that is marginally uncorrelated but jointly correlated with the response variable; (2) SIS tends to select some unimportant predictor variables that are highly correlated with the important predictor variables than those important predictor variables relatively less corelated to the response variable; (3) the problem of multi-collinearity among the predictor variables increases the difficulty of variable screening. Fan and Lv [21] proposed to overcome these potential problems by an iterative SIS.

Fan et al. [24] extended and improved the idea of iterative SIS and developed an iterative variable screening method under some more general statistical models. Assume that our goal is to obtain a sparse feature space β to optimize

$$n^{-1}\sum_{i=1}^{n}\mathcal{L}(Y_i, x_i^T\beta) + \sum_{j=1}^{p} p_\lambda(|\beta_j|). \tag{1.9}$$

The proposed iterative procedure consists of the following steps.

1. Pick a set $\mathcal{A}_1$ of size k_1 using a SIS procedure, and then apply a regularization approach like LASSO, SCAD or ENET to select a subset $\mathcal{M}_1$.

2. Instead of calculating the residuals from the regression as in Fan and Lv [21], for $j \notin \mathcal{M}_1$, compute

$$\mathcal{L}_j^{(2)} = \min_{\beta_0,\beta_{\mathcal{M}_1},\beta_j} n^{-1}\sum_{i=1}^{n}\mathcal{L}(Y_i, \beta_0 + x_{i,\mathcal{M}_1}{}^T\beta_{\mathcal{M}_1} + X_{ij}\beta_j) \tag{1.10}$$

   where $x_{i,\mathcal{M}_1}$ are the variables contained in $\mathcal{M}_1$. This equation quantifies the additional

contribution of variable X_j with the variables $x_{\mathcal{M}_1}$. Pick the k_2 smallest $\left\{\mathcal{L}_j^{(2)}, j \notin \mathcal{M}_1\right\}$ variables and denote $\mathcal{A}_2$ the resulting set.

3. Estimate the parameters using

$$\hat{\beta}_2 = \underset{\beta_0, \beta_{\mathcal{M}_1}, \beta_{A_2}}{argmin}\ n^{-1} \sum_{i=1}^{n} \mathcal{L}(Y_i, \beta_0 + x_{i,\mathcal{M}_1}^T \beta_{\mathcal{M}_1} + x_{i,A_2}^T \beta_{A_2}) + \sum_{j \in \mathcal{M}_1 \cup A_2} p_\lambda(|\beta_j|). \tag{1.11}$$

4. Iterate step 2-3 until some predefined number d variables are selected or $\mathcal{M}_l = \mathcal{M}_{l-1}$.

The eventual estimation is the $\hat{\beta}_{\mathcal{M}_l}$. This iterative procedure extends the iterative SIS to a general statistical framework. It can be easily extended to many procedures. It is also able to delete predictors from the previously selected set.

Xu and Chen [70] found that the iterative procedure has heavily computational cost and increased complexity, which motivated them to develope a sparsity restricted MLE (SRMLE) procedure for the generalized linear models (GLMs), which is computationally cheaper. They further proved that the SRMLE procedure has the sure screening property.

SIS for GLMs

Consider a GLM with canonical link. The conditional likelihood is

$$f(y|x) = exp\left\{y\theta(x) - b(\theta(x)) + c(y)\right\}, \tag{1.12}$$

$b(\cdot)$, $c(\cdot)$, and $\theta(x) = x^T\beta$ are some known function once the distribution of response variable is known. Without loss of generality, suppose the dispersion parameter $\phi = 1$. We also assume that each column of the design matrix is standardized with mean zero and variance

one. The penalized likelihood for GLMs is

$$\frac{1}{n}\sum_{i=1}^{n}\ell((x_i)^T\beta, y_i) + \sum_{j=1}^{p} p_\lambda(|\beta_j|), \tag{1.13}$$

where $\ell(\theta, y) = b(\theta) - y\theta$. Denote $\hat{\beta}_j^M$ to be the maximum marginal likelihood estimator (MMLE), which is the minimizer of the marginal regression

$$\hat{\beta}_j{}^M = (\hat{\beta}_{j,0}^M, \hat{\beta}_j^M) = \arg\min_{\beta_0,\beta_j} \sum_{i=1}^{n} \ell(\beta_0 + \beta_j X_{ij}, Y_i), \tag{1.14}$$

where the X_{ij} is the ith observation of the jth variable. It is rational to treat the magnitude of $\hat{\beta}_j^M$ to rank the importance of the predictor variables. Fan and Song [25] picked the predictor variables whose estimates exceeded a predefined threshold value γ_n:

$$\mathcal{M}_{\gamma_n} = \{1 \le j \le p : |\hat{\beta}_j{}^M| \ge \gamma_n\}. \tag{1.15}$$

Fan and Song[25] further proved that MMLE screening procedure has the sure screening property, and under some technical assumptions, the MMLEs are uniformly convergent to the population values. They also discussed the size of the selected model.

1.2.3 Other Independence Screening

Nonparametric Independence Screening

In the real world,Rare previous information indicates that the impact of the predictor variable takes a linear form or belongs to any other parametric family of the finite dimensions. Sometimes significant improvements can be made by using some flexible nonparametric models, like the additive model [60]. For the ultra high dimensional additive model, Fan et al.

[18] developed a nonparametric independence screening (NIS)

$$Y = \sum_{j=1}^{p} m_j(X_j) + \epsilon, \tag{1.16}$$

where $\{m_j(X_j)\}_{j=1}^{p}$ have mean 0. They rank the utility of variables according to $E(f_j^2(X_j))$, where $f_j(X_j) = E(Y|X_j)$ is the projection of Y onto X_j. $f_j(x)$ can be estimated via a normalized B-spline basis $B_j(x) = \{B_{j1}(x), \ldots, B_{jd_n}(x)\}^T$:

$$\hat{f}_{nj}(x) = \hat{\beta}_j^T B_j(x), 1 \leq j \leq p, \tag{1.17}$$

where $\hat{\beta}_j = (\beta_{j1}, \ldots, \beta_{jd_n})^T$ is computed by the marginal least squares estimation:

$$\hat{\beta}_j = \underset{\beta_j \in \mathcal{R}^{d_n}}{\arg\min} \sum_{i=1}^{n} (Y_i - \beta_j^T B_j(X_{ij})). \tag{1.18}$$

Thus the screened model index set is

$$\mathcal{M}_\nu = \{1 \leq j \leq p : ||\hat{f}_{nj}||_n^2 \geq \nu_n\}, \tag{1.19}$$

where ν_n is a predefined threshold value and $||\hat{f}_{nj}||_n^2 = n^{-1} \sum_{i=1}^{n} \hat{f}_{nj}(X_{ij})^2$. This independence screening ranks the importance of variables based on the the marginal nonparametric regression strenth. Fan et al.[18] further introduced INIS-penGAM procedure to decrease the false selection rate and greedy INIS (g-INIS) algorithm to deal with the highly correlated covariates.

Fan et al. [23], and Liu et al. [44] furhter expand the NIS to the area of varying coefficient models. Fan et al. [23] developed a conditional correlation screening method according to the kernel regressions. Liu et al. [44] proposed a conditional correlation learnin (CC-SIS).

Model-free Feature Screening for Continuous Variables

Most of the screening procedures we reviewed above focuses on a class of specific model and its performance is based upon the belief that the imposed working model is close to the true model. However, it may be very challenging to specify the model structure on the regression function in ultra high dimensional modelling.

Zhu et al.[77] proposed a sure independent ranking screening(SIRS) procedure, which is a model-free variable screening procedure. Let Y be the response variable with support Ψ_y, and Y can be both univariate and multivariate. Let $x = (X_1, \ldots, X_p)^T$ be a covariate vector. Zhu et al.[77] first developed the notion of active predictors and inactive predictors without specifying a regression model. Consider the conditional distribution function of Y given x, denoted by $F(y|x) = P(Y < y|x)$. Define the true model

$$\mathcal{M}_* = \{k : F(y|x) \text{ functionally depends on } X_k \text{ for some } y \in \Psi_y\}, \tag{1.20}$$

if $k \in \mathcal{M}_*$, X_k is referred to as an active predictor, otherwise it is referred to as an inactive predictor. Zhu et al.[77] considered a general model framework under which a unified screening approach was developed. Assume that

$$F(y|x) = F_0(y|B^T x_{\mathcal{M}_*}), \tag{1.21}$$

where $F_0(\cdot|B^T x_{\mathcal{M}_*})$ is an unknown distribution function for a given $B^T x_{\mathcal{M}_*}$. Assume that $E(X_k) = 0$ and $Var(X_k) = 1$ for $k = 1, \ldots, p$. Define $\Omega(y) = E[xF(y|x)]$. It then follows by the law of iterated expectations that $\Omega(y) = E[xE\{1(Y < y)|x\}] = cov\{x, 1(Y < y)\}$. Let $\Omega_k(y)$ be the kth element of $\Omega(y)$, and define

$$\omega_k = E\{\Omega_k^2(Y)\}, k = 1, \ldots, p. \tag{1.22}$$

Then ω_k is to serve as the marginal utility measure for predictor ranking. A natural estimator for ω_k is

$$\hat{\omega}_k = \frac{1}{n}\sum_{j=1}^{n}\widehat{\Omega}_k^2(Y_j) = \frac{1}{n}\sum_{j=1}^{n}\left\{\frac{1}{n}\sum_{j=1}^{n}X_{ik}1(Y_i < Y_j)\right\}^2, k = 1, \ldots, p, \tag{1.23}$$

where X_{ik} denotes the kth element of x_i. Zhu et al.[77] proposed ranking all the candidate predictor X_k according to $\hat{\omega}_k$ from the largest to smallest, and then selecting the top ones as the active predictors. And they empirically demonstrated that the combination of the soft cutoff and hard cutoff by setting $d = [n/\log(n)]$ works quite well in their simulation studies.

Several other model-free screening procedures have been proposed. Li et al.[40] proposed a robust rank correlation screening (RRCS) procedure based on the Kendall rank correlation. RRCS can be used against outliers and influence points in the observations and the sure independence screening property can hold only under the existence of a second order moment of predictor variable. Li et al.[41] developed the sure independence screening procedure based on the distance correlation (DC-SIS) under general parametric models. The DC-SIS can be used directly to screen grouped predictor variables and for multivariate response variables.

Model-free Feature Screening for Categorical Data

The aforementioned methods implicitly assume that predictor variables are continuous. Ultra high dimensional data with categorical predictors and categorical responses are frequently encountered in practice.

To deal with the cases when the predictors and the responses are all categorical, Huang et al.[34] employed the Pearson χ^2 test statistic as a marginal utility for feature screening. Let $Y_i \in \{1, \ldots, K\}$ be the corresponding class label, and $X_i = (X_{i1}, \ldots, X_{ip})^T \in \mathcal{R}^p$ be the associated categorical predictor. Define $P(Y_i = k) = \pi_{yk}$, $P(X_{ij} = k) = \pi_{jk}$, and $P(Y_i = k_1, X_{ij} = k_2) = \pi_{yj,k_1k_2}$. Those quantities can be estimated by $\hat{\pi}_{yk} = n^{-1}\sum I(Y_i = k)$,

$\hat{\pi}_{jk} = n^{-1}\sum I(X_{ij} = k)$, and $\hat{\pi}_{yj,k_1k_2} = n^{-1}\sum I(Y_i = k_1)I(X_{ij} = k_2)$. Subsequently, a chi-square type statistic can be defined as

$$\hat{\triangle}_j = \sum_{k_1=1}^{K}\sum_{k_2=1}^{2}\frac{(\hat{\pi}_{yk_1}\hat{\pi}_{jk_2} - \hat{\pi}_{yj,k_1k_2})^2}{\hat{\pi}_{yk_1}\hat{\pi}_{jk_2}}, \tag{1.24}$$

which is a nature estimator of

$$\triangle_j = \sum_{k_1=1}^{K}\sum_{k_2=1}^{2}\frac{(\pi_{yk_1}\pi_{jk_2} - \pi_{yj,k_1k_2})^2}{\pi_{yk_1}\pi_{jk_2}}. \tag{1.25}$$

Huang et al.[34] proposed estimating the true model by $\hat{S} = \{j : \hat{\triangle}_j > c\}$, where $c > 0$ is some prespecified constant. They further established the sure screening property under mild conditions.

SIS for Classification

Classification and discriminant analysis are useful for analysis of categorical response data. Traditional methods of classification and discriminant analysis may break down when the dimensionality is extremely large.

Let Y be a categorical response with K classes $\{y_1, \ldots, y_K\}$. If an individual covariate X_j is associated with the response Y, then $\mu_{jk} = E(X_j|Y = y_k)$ are likely different from the population mean $\mu_j = E(X_j)$. It is intuitive to use the test statistic for multi-sample mean problem as a marginal utility for feature screening. Fan and Fan[17] proposed using the two sample t-statistic as marginal utility for feature screening in high dimensional binary classification. They further showed that the t-statistic does not miss any important features with probability 1 under some technical conditions.

Although the variable screening based on two-sample t-statistic performs generally well in the

high-dimensional classification problems, it may break down for heavy-tailed distributions or data with outliers. To overcome this drawback, Mai and Zou[47] proposed a feature screening method for binary classification based on the Kolmogorov-Smirnov statistic. Let $F_{+j}(x)$ and $F_{-j}(x)$ denote the conditional cumulative probability functions of K_j given $Y = 1, -1$, respectively. Define $K_j = sup_{-\infty<x<\infty}|F_{+j}(x) - F_{-j}(x)|$. The sample version of K_j is defined as $K_{nj} = sup_{-\infty<x<\infty}|\hat{F}_{+j}(x) - \hat{F}_{-j}(x)|$. Mai and Zou[47] proposed ranking all the variables by the K_{nj} statistics, which is the Kolmogorov-Smirnov test statistic for testing the equivalence of two distributions. Mai and Zou[47] further established the sure screening property and showed that this method is almost as fast as t-test screening[17] and is ten times faster than nonparametric maximum marginal likelihood screening[18]. However, it is limited to the binary classification. Cui et al.[12] proposed a model-free feature screening procedure using mean variance index for ultra high dimensional discriminant analysis. It is not only robust to heavy-tailed distributions of predictors and the presence of potential outliers, but also allows the categorical response having a diverging number of classes in the order of $O(n^k)$ with some $k \geq 0$.

Conditional SIS

Where certain important variable sets are previously known, whose conditional contribution to the response, given the known set of variables, leads to conditional independence screening (CSIS), which is a natural assessment of the other predictors are relative importance. Barut et al. [5] developed this CSIS for ultra high dimensional data in GLMs. Consider model 1.12, we wish to recruit additional variables, $X_{\mathcal{T}}$, from rest of the variables given a set of variables $X_{\mathcal{C}}$. Without loss of generality, suppose $\mathcal{C}$ to be the set of first q variables, which are known to be important to the response variable. And the $\mathcal{T}$ is the remaining set of size

$d = p - q$. We use the notation

$$\beta_{\mathcal{C}} = (\beta_1, \ldots, \beta_q)^T \in \mathcal{R}^q \tag{1.26}$$

$$\beta_{\mathcal{T}} = (\beta_{q+1}, \ldots, \beta_p)^T \in \mathcal{R}^d \tag{1.27}$$

and similar notation for $X_{\mathcal{C}}$ and $X_{\mathcal{T}}$. For a data sample $\{(X_i, Y_i)\}_{i=1}^n$ in (1.12), the minimizer of the negative marginal log-likelihood is maximum marginal likelihood estimator (MMLE) $\hat{\beta}^M_{\mathcal{C},j}$ for $j = q+1, \ldots, p$:

$$\hat{\beta}^M_{\mathcal{C},j} = \underset{\beta_{\mathcal{C}}, \beta_j}{\arg\min} \mathbb{P}\left\{\ell(X_{\mathcal{C}}\beta_{\mathcal{C}} + X_j\beta_j, Y)\right\} \tag{1.28}$$

where the empirical measure is $\mathbb{P}_n f(X, Y) = \frac{1}{n}\sum_{i=1}^n f(X, y)$. The strength of the conditional contribution of X_j given $X_{\mathcal{C}}$ is measured by $\hat{\beta}^M_j$. The variables to be kept are based on a theoretically predefined threshold γ:

$$\hat{\mathcal{M}}_\gamma = \left\{1 \le k \le p : |\hat{\beta}^M_k| \ge \gamma\right\}, \tag{1.29}$$

In another words, we keep the variables that have large additional contribution in the presence of $X_{\mathcal{C}}$.

The threshold parameter γ could be selected by two criteria: controlling false discovery rate (FDR) and random decoupling. For controlling FDR, we know each coefficient asymptotically follows

$$\left[I_j(\hat{\beta}^M_j)\right]^{\frac{1}{2}} \hat{\beta}^M_j \sim \mathcal{N}(0, 1) \tag{1.30}$$

where $I_j(\hat{\beta}^M_j)$ is the component corresponds to β_j in the Fisher information matrix. We can keep the variables satisfying $\hat{\mathcal{M}}_\gamma = \left\{j : I_j(\hat{\beta}^M_j)|\hat{\beta}^M_j| \ge \delta\right\}$, where $\delta = \phi^{-1}(1 - f/2d)$ and f is the number of allowable false positives. The other one is to use random decoupling. To

conduct random decoupling, the last d columns of the design matrix are randomly permuted at first, while the first q columns are kept intact. Then for each permuted columns X_j^{M*}, which corresponds to be X_j in the design matrix, the regression coefficient $\hat{\beta}_j^{M*}$ should be statistically estimated zero since X_j^{M*} and Y are not correlated anymore due to the random permutation. Then let $\hat{\gamma}^* = \max_{q+1 \leq j \leq p} |\hat{\beta}_j^{M*}|$, repeat the procedure B times, resulting in

$$\left\{ |\hat{\beta}_b j^{M*}|, j = q+1, \ldots, p \right\}_b^B, \tag{1.31}$$

it is recommended to take the maximum value of $\hat{\gamma}_b^*$, where $\hat{\gamma}_b^* = \max_{q+1 \leq j \leq p} |\hat{\beta}_b j^{M*}|$.

1.2.4 Variable selection for linear mixed-effect model

The variable selection for linear model has been extensively studied for decades, but relatively few articles examine high-dimensional regression problems involving random effects, which usually appear in cluster or repeated measure data. To deal with repeated measures, linear mixed effect model is frequently used by adding random effects to the predictor variables in ordinary linear regression. This model provides a convenient tool for the analysis of longitudinal data and has got a lot of consideration in many applications such as health study, biology, econometrics, etc. The increasing dimensionality of fixed and/or random components results the interpretation as well as the inferences of the linear mixed effect models to be more problematic due to complexity of the mixed effects models. Therefore, selecting important fixed or random effects has been more and more important in the analysis of longitudinal or repeated measurements data using mixed effects models.

Variable selection in longitudinal data has been studied in some literature in the past few decades. For example, Vaida and Blanchard [63] proposed the conditional AIC (cAIC), which is an extension of the commonly-used AIC [1], to determine the degrees of freedom

when incorporating random effects in the mixed effec models. The cAIC then was explored by Liang et al [43]. Chen and Dunson [11] developed a Bayesian variable selection procedure for choosing the important random effects in the mixed effect models by Cholesky decomposition of unknown covariance matrix of random effects, and specified a prior distribution on the standard deviation of random effects with a positive mass at zero to achieve the sparsity of random components. Pu and Niu [55] proposed an extended version of generalized information criterion (GIC) to select linear mixed models and explored the asymptotic property of the selected fixed effects for the proposed method. Bondell et al [6] developed a joint variable selection method to select fixed and random effects in the mixed effects models by a modified Cholesky decomposition in the setting of fixed dimensionality for both fixed effects and random effects. Ibrahim et al [36] proposed a general class of linear mixed effects models of both fixed and random effects in fixed number of dimension by maximizing penalized likelihood method with the SCAD and adaptive LASSO penalty to select fixed and random effects. Fan and Li [27] proposed a proxy matrix in the penalized profile likelihood to deal with the problem of unknown covariance matrix of random effects to select and estimate significant fixed effects. They also proposed to select and estimate important random effects by a group variable selection strategy. Jurg Schelldorfer et al [56] came up with generalized linear mixed models (GLMMs) with an ℓ_1 penalty to fit the high dimensional longitudinal data. This Lasso-type procedure for GLMMs could be applied for variable selection to decrease the dimensionality of feature space.

1.3 Graphical modeling

Graphical model is a frequently used to explore association networks for a set of variables, where the variables can be genes, proteins, or any other objects under study based on the

research. The idea behind the graphic model is to use the partial correlation coefficient for any two variables as a measure of dependence. The partial correlation coefficient equaling to zero suggests a conditional independent relationship between the two variables. Another way of measuring the dependency is to calculate the correlation coefficient of two variables. However, the correlation coefficient is much weaker than the partial correlation coefficient, since all the variables in a real system are more or less correlated. Specifically, let $X = (X_1, ..., X_p)$ be a p-dimensional random vector drawn from a p-dimensional multivariate Normal distribution $\mathcal{N}_p(\mu, \Sigma)$, where μ is the mean vector and Σ is the variance-covariance matrix. A common way of studying graphical model is based on covariance selection [14] by recognizing the nonzero off-diagonal elements in the precision matrix C, which is the inverse of the covariance matrix ($C = \Sigma^{-1}$), since those entries correspond to variables that are conditionally dependent. Moreover, the partial correlation coefficient between any of the two variables X_i and X_j in the data given all other variables is showed by Lauritzen [39]

$$\rho_{i,j|\mathcal{D}\setminus\{i,j\}} = -\frac{C_{i,j}}{\sqrt{C_{i,i}C_{j,j}}} \tag{1.32}$$

where $C_{i,j}$ is the (i,j)-entry of the precision matrix, and $\mathcal{D} = \{1, 2, ..., p\}$ is the set of nodes or variables. Therefore, the estimation of the partial correlation coefficients or the precision matrix is the key to build up a graphical model. But this method could not be applied when $p \ll n$, since the sample precision matrix is singular and thus cannot be directly estimated from the data.

Various procedures have been developed in the literature to overcome this problem. And the existing methodologies can be roughly grouped into three classes according to the their schemes: (1) limited order partial correlations; (2) nodewise regression; (3) regularized graphical models.

The first category of work includes but not limited to Magwene and Kim [46], Wille and B uhlmann [67], and Castelo and Roverato ([8], [9]), etc. Magwene and Kim [46], Wille and B uhlmann [67] suggested to use the first-order partial correlation coefficient rather than the full-order partial correlation coefficient. Castelo and Roverato [8] proposed to study the graphical models by a *qp*-procedure according to a the non-rejection rate developed by themselves. This non-rejection rate is defined to be the probability of not rejecting the null hypothesis, for X_i and X_j, $\rho_{ij}|Q = 0$, where Q is a subset of all variables randomly selected from $\mathcal{D}\backslash\{i, j\}$ with size q, which is predetermined by the user. A higher non-rejection rate suggests a higher probability of conditional independence of two variables. The *qp*-procedure first estimates the non-rejection rate for each pair of variables according to the random Q samples, and then excludes those edges when the estimated non-rejection rates going beyond a predefined threshold. The performance of the *qp*-procedure is very obvious to rely on the choice of $q = |Q|$. A higher value of q may result in a better performance of approximating the non-rejection rate to the coefficient of empirical partial correlation, but it can also compromise the power of the statistical tests depending on $n - q$. Castelo and Roverato [9] further developed an averaging *qp*-procedure to enhance the performance of *qp*-procedure by averaging the non-rejection rate estimated from a sequence of q values. It is widely recognized that these procedures will lead to something in the middle of the full graphical model (with correlations conditioned on all $p - 2$ variables) and the correlation graph (with unconditioned correlations) because of their approximation nature. Spirtes et al [59] proposed a remarkable algorithm: the PC algorithm, which operates in an iterative way. It proceeds with a full graph with edges between all variables, and then searches for a subset Q, and $|Q| \leq m$ (which is called the depth of the search and predetermined), for each edge of the current graph. So that the two nodes connected by the edge are independent conditional on Q. The corresponding edge is removed once such a subset Q is discovered. One disadvantage of this PC algorithm is that it is less powerful when p is large, because the

it substantially applies a maxP statistic (i.e., maximum of a set of p-values) to test on the conditional independence. And this algorithm has recently been extended to the variable selection problem for high-dimensional data [7].

The second category is the nodewise regression method, which was developed by Meinshausen and B uhlmann [51]. This nodewise regression identifies the edge between nodes based on the relationship between the regression coefficients and partial correlation coefficients. For a linear model between a marginal variable and the remaining

$$X_j = \beta_{ji} X_i + \sum_{l \in \mathcal{D} \setminus (i,j)} X_l \beta_{jl} + \epsilon_j \tag{1.33}$$

where ϵ_j is the random error when consider X_j as the response variable. Then $\beta_{ji} = -C_{j,i}/C_{j,j}$ and $\beta_{ij} = -C_{i,j}/C_{i,i}$; that is,

$$C_{i,j} \neq 0 \Leftrightarrow \rho_{i,j|V\setminus(i,j)} \neq 0 \Leftrightarrow \beta_{ij} \neq 0 \; and \; \beta_{ji} \neq 0 \tag{1.34}$$

Meinshausen and B uhlmann [51] recommended to use the LASSO regression [62] to collect the nonzero regression coefficients Since p can be greater than n. Let $S^{(j)} = \left\{ l : \beta_l^{(j)} \neq 0 \right\}$ be the set of predictor variables selected by the LASSO for variable X_j. Then the graphical model could be built up by the "or" rule: draw an edge between nodes i and $j \iff i \in S^{(j)}$ or $j \in S^{(i)}$, or the "and" rule: draw an edge between nodes i and $i \in S^{(j)}$ and $j \in S^{(i)}$. This procedure usually leads to a dense network regardless of the consistency of both "or" and "and" rule, which has been theoretically proved by Meinshausen and B uhlmann [51]. The unexplained signal variables will pull in other edges that would otherwise no be included because of the shrinkage of the regression coefficients on the true edge towards 0 by ℓ_1 penalty.

Yuan and Lin [73] proposed to estimate the precision matrix C by minimizing the penalized negative log-likelihood function rather than working on nodewise regressions

$$-\log(\det(C)) + trace(\hat{\Sigma}_{MLE}C) + \lambda\|C\| \tag{1.35}$$

where $\hat{\Sigma}_{MLE}$ is the maximum likelihood estimator of Σ, $\|C\|$ is the ℓ_1-norm of all off-diagonal elements in C, and λ is the regularization parameter. Specifically,when setting $\|C\| = \Sigma_{i<j}|C_{i,j}|$, the minimization problem in (1.35) is convex, and based on which fast algorithms ([30], [4]) have been developed, which is also referred to be graphical LASSO (GLASSO) due to the employment of ℓ_1 penalty. The regularization parameter λ can be selected by cross-validation criterion. Similar to the nodewise regression method, GLASSO also tends to produce dense network. Besides, Mazumeri and Hastie [49] pointed out that the converged precision matrix might not be the inverse of the estimated covariance matrix.

1.4 Motivation for proposal work

Variable screening is of great importance for longitudinal data analysis, but is also very problematic because of within cluster correlation as well as model complexity. In the previous section, we have provided some literature review on feature screening for ultra-high dimensional data and graphical modeling. We summarized a variety of variable screening approaches for linear models, generalized linear models, nonparametric regression, model-free feature screening procedures and several linear mixed effect models. But the methods in Section 1.2.4 can only handle data when the number of predictor variable similar to or smaller than the sample size. They will have difficulty when the number of predictor variable is much larger than the sample size, like the $p = 40,000+$ to $n = 924$ in the SLE study. So a feature screening procedure extending those in Section 1.2.2 is needed to reduce the

ultra-high dimension in longitudinal data to a much smaller dimension manageable by the regularization variable selection methods in Section 1.2.4. In this dissertation, we first extend the SIS and iterative SIS approach for longitudinal data to bring down the dimension to a much smaller scale, so that the methods mentioned in section 1.2.4 could then be applied to further select important fixed or random effects and estimate the coefficients and variance components.

The conditional independence graph is a common tool in describing the relationship between a set of variables. The existing methods such as the dynamic Bayesian network (DBN) treat longitudinal measurements as time series, which often requires a much higher sampling frequency as well as restricts the correlation to be serial. In this dissertation, we proposed a penalize likelihood method as well as a nodewise regression method, which is an extension of Meinshausen and B uhlmann's work [51], for longitudinal data. We use a pairwise coordinate descent combined with second order cone programming (SOCP) to optimize the penalized likelihood and estimate the parameters, and so that the edges in the conditional independence graph can be identified.

The remainder of this dissertation is organized as follows. In Chapter 2, we focus on feature screening procedures of SIS and iterative SIS for longitudinal data. We propose a new robust procedure and conduct simulation studies to assess the performance of the proposed procedure. The goal of this chapter is to identify a robust screening procedure for longitudinal data, and the relationship among selected important variables will be further explored in the next chapter. In Chapter 3, we propose a robust graphical modeling for longitudinal data. We conduct simulation studies to evaluate the performance of the proposed procedure and we further illustrate the procedure using SLE data.

Chapter 2

SIS for ultra-high dimensional Longitudinal Data

In this chapter, we focus on feature screening procedures for longitudinal data and aim to develop feature screening procedure that can dramatically reduce the dimension of the longitudinal predictor to a much smaller scale that is comparable to the sample size. The goal is make the output from the screening procedure manageable by a regularized variable selection procedure for linear mixed-effects model. The response variable and predictors are all continuous variables.

2.1 Introduction

In longitudinal studies, data are commonly correlated, because the measurements are often taken repeatedly over time on the same subject. For example, the immunological bioinformatics for disease monitoring is a common topic and data from such studies often consist of repeated measurements on the same cohort of subjects over time. In response to such applications, statistical analysis tools for longitudinal data have been developed and made available to practitioners linear mixed effect models provide a very convenient tool for the analysis of longitudinal data, where within-subject correlation is modeled by random effects. The monograph by Diggle et al. [15] provides an excellent overview of statistical methods for longitudinal data analysis.

With the rapid development of biotechnology, high-throughput genomic data have become commonly available, where a large number of gene expressions are collected. As the technology becomes more mature , its cost also gets dramatically reduced over the decades. This allows researchers to carry out experiments where gene gene expression data are collected repeatedly over time. Therefore such experiments often produce longitudinal gene expression data of ultra-high dimensions. The common goal of these experiments, similar to that of the older-generation of single-replicate genetic experiments, is to identify important genes

associated with the disease of interest and figure out how they interact with each other in their genetic functions. In this chapter, we focus on feature screening in linear mixed models for ultra-high dimensional longitudinal data with continuous outcomes.

The motivating example of this dissertation comes from a study of systemic lupus erythematosus (SLE). It is an autoimmune disease characterized by breakdown of tolerance to nucleic acids and highly diverse clinical manifestations. The study was to compare transcriptomes of SLE patients at various stages of the disease, as well as to evaluate the transcriptome for individual patients over time while taking a large number of gene expression values into account [3].The blood transcriptome of 158 pediatric patients was longitudinally profiled for up to 1,412 days, representing 924 visits, in order to assess its molecular heterogeneity. SLE Disease Activity Index (SLEDAI) for each patient, which is a weighted metric combining 24 components, as well as 43,739 gene expressions were measured at each visit. The number of visits per patient ranged from 1 to 22. The repeated measures within each subject (patient) warrants the need for a longitudinal data analysis approach.

The first issue to be addressed in our project is the ultra-high dimension of the longitudinal data studied in immunological bioinformatics, of which the dimension is often on the scale of tens of thousands. Performing any kind of analysis on such data is a daunting task and associated with extremely high computational cost. A similar issue in the scenario of single-replicate data has motivated the work of Fan and Lv [21], who introduced the concept of feature screening (FS) and used a screening criterion to reduce the ultrahigh dimension to a much more manageable scale, though possibly still a high dimension like hundreds or thousands. The criterion they used was the marginal correlation between the response and each predictor. Note that their correlation screening procedure was among the first ones supported by a rigorous statistical theory, although testing correlations has been a popular approach in bioinformatics for a while. Various screening procedures have been

proposed since then. For example, Wang [64] proposed a forward regression method for FS in ultrahigh dimensional linear models. Fan et al. [24] and Fan and Song [25] developed FS procedures for generalized linear models and robust linear models. Fan et al. [18] developed a nonparametric FS procedure for additive models. Li et al. [40] proposed a model-free FS procedure based on certain distance correlation [61]. Mai and Zou [47] proposed a FS procedure for binary classification with ultrahigh dimensional predictors based on the Kolmogorov's statistic. Barut et al. [5] present a conditional sure independence screening for the case that some prior knowledge on a certain important set of variables is available. However, all these methods were for the scenario of single-replicate data and have not studied the data with ultra-high dimensional longitudinal measurements.

In this dissertation, we develop an independence screening procedure based on ranking p-value or maximum marginal likelihood estimator (MMLE) for linear mixed effect models. For each predictor, we fit a marginal linear mixed effects model with the response against this predictor. The p-values of the slope coefficient or the magnitude of the absolute values of the MMLEs from these marginal linear mixed effects models can preserve the nonsparsity information of the joint models, provided that the true values of the marginal likelihood preserve the nonsparsity of the joint regression models and that the MMLE estimates the true values of the marginal likelihood uniformly well. We rank the p-values and MMLEs, and select the important variables by theoretically determined parameters. As a practical screening method, independent screening can miss variables that are weakly correlated with the response variables, but jointly highly important to the response variables. It can rank some jointly unimportant variables too high by using marginal methods. Therefore, we extend SIS to iterative SIS to make the procedures robust and practical. This could be viewed as a generalization of the work from Fan and Lv[21] and Fan et al.[24]. The former focuses on ordinary linear models and the latter extends the idea to generalized linear models.

The SIS property can be achieved as long as the surrogate, in this case, the marginal utility, can preserve the nonsparsity of the true parameter values. In this dissertation, we screen variables with p-value or MMLE, which can be viewed as a Wald type of statistic. We show that these three methods are equivalent in terms of sure screening properties in our proposed framework. This generalizes the marginal likelihood ratio screening results for single replicate data in Fan et al[24].

In our simulations, we compare the performance of SIS and iterative SIS based on p-value and MMLE and check the median minimum model size, number of false negative and number of false positive. The application of our methods to SLE data provide critical information about this disease. In summary, our method has the following distinguishing features: (1) it is the first screening method proposed for ultra-high dimensional longitudinal data; (2) we derive the equivalence between p-value screening and MMLE screening; (3) we rigorously establish the sure independence screening property for the proposed screening procedures.

The rest of this chapter is organized as follows. In Section 2, we introduce the sure independent screening with p-value and MMLE. The sure independent screening property and uniform convergence of marginal maximum likelihood estimator are presented in Section 3. In Section 4, we conducted numerical analysis of two types of screening on simulated and real data. A summary of our findings and discussions is in Section 5. The details of the proof are deferred to the Appendix A.

2.2 Method

2.2.1 Notation and Model

Consider longitudinal data (X_{ij}, Y_{ij}, Z_{ij}), $i = 1, \ldots, m$, $j = 1, \ldots, n_i$ collected from m independent subjects. Here Y_{ij} is the response of j-th measurement of i-th subject, m is the number of subject, n_i is the number of measurement for subject i, X_{ij} and Z_{ij} are respectively the $p \times 1$ and $q \times 1$ covariate vectors of j-th measurement of i-th subject for fixed and random effects. A linear mixed effect model,

$$Y_{ij} = X_{ij}^T \beta + Z_{ij}^T b_i + \epsilon_{ij}, \ i = 1, \ldots, m, \ j = 1, \ldots, n_i \tag{2.1}$$

where β is the $p \times 1$ fixed effects coefficient, b_i is the $q \times 1$ random effect parameter and follows $\mathcal{N}(0, D)$, where $D = \psi^2 I_q$, ϵ_{ij} is the random error and $\epsilon_{ij} \sim \mathcal{N}(0, R)$, where $R = \sigma^2 I_n$. And b_i and ϵ_{ij} are independent with each other. The total number of observations is $n = \sum_{i=1}^{m} n_i$.

The vector-matrix version for model (2.1) is

$$Y = X\beta + Zb + \epsilon \tag{2.2}$$

where Y is the $n \times 1$ response vector, X is $n \times p$ and consists of the X_{ij}^T, Z is $n \times q$ and is composed of the Z_{ij}^T, and ϵ is the stacked vector of the ϵ_{ij}. The role of the three parts of the (2.2) can be explained as follows. The fixed effects are the population average coefficients for time variables and other predictors, which models the variations of mean growth change in data for the population. In contrast, the random effects account for the heterogeneity among the subjects by allowing differences from the overall average. Finally, the error accounts for the variation unexplained by the fixed and random effects.

2.2.2 SIS for longitudinal data

Independence screening with p-value

For single replicate data, marginal Pearson correlation or Spearman rank correlation between the response variable and each predictor variable are usually used in SIS to obtain the magnitudes of the predictor effects. However, these approaches are not applicable to longitudinal data, since marginal correlations between the response and individual predictors are not readily available due to the presence of within-subject correlations. Instead, the corresponding p-value of the fixed effect from the marginal linear mixed effect model,

$$Y_{ij} = \beta_{0,k}^M + \beta_k^M X_{ijk} + b_k^M + \epsilon_{ij}^M,\ i = 1, \ldots, m,\ j = 1, \ldots, n_i,\ k = 1, \ldots, p \tag{2.3}$$

is compared. where $\beta_{0,k}^M$ is the marginal intercept, β_k^M is the marginal slope, b_j^M is the marginal random intercept, and ϵ_{ij}^M is the marginal $i.i.d$ random error. We only consider the random intercept case here since we only focus on screening fixed effects. And it can be shown that using p-value is equivalent to using the marginal maximum likelihood in SIS [24]. That is, for each pair of Y and X_k, where X_k is the k-th column of X, a componentwise linear mixed effect model can be fitted and the p-value for testing on the significance of the coefficient, τ_k, can be recorded. We can then sort the magnitudes of all the τ_k in an increasing order and select a submodel

$$\mathcal{M}_{d_n} = \{1 \leq k \leq p : \tau_k \text{ is among the first } d_n \text{ smallest of all}\},$$

where d_n is a predefined threshold value (suggested to be $n/\log(n)$ by Fan and Lv[21]). This reduces the full model of size p to a submodel with the size d_n.

The p-value is defined as the probability, under the null hypothesis H_0, of obtaining a result

equal to or more extreme than what was actually observed. In each marginal linear mixed effect model, the null hypothesis is $H_0 : \beta_k^M = 0$. The smaller the marginal p-value is, the more significant the corresponding predictor variable is. Therefore, we can also say the coefficient of the fixed effect with smaller p-value is less likely to be zero and thus should be selected.

Independence screening by MMLE

Another selection criterion we consider is MMLE. Given the random sample from linear mixed effect model (2.2), the maximum marginal likelihood estimator $\hat{\beta}_k^M$ for $k = 1, \ldots, p$ is defined as the minimizer of the negative marginal likelihood

$$\hat{\beta}_k^M = \arg\min_{\beta_k^M} \left\{ ||Y - \beta_{0,k}^M 1_n - \beta_k^M X_k - b_k^M 1_n||^2 + b_k^{M^T} D_k^M b_k^M \right\} \tag{2.4}$$

It measures the strength of the marginal effect of X_k. And this can be rapidly computed and avoid numerical instability in NP-dimension problems. The marginal screening based on the estimated marginal magnitude of fixed effect is to keep the variables

$$\hat{\mathcal{M}}_\gamma = \left\{ 1 \leq k \leq p : |\hat{\beta}_k^M| \geq \gamma \right\}, \tag{2.5}$$

for a theoretical determined thresholding parameter γ. The importance of predictors is ranked by this independence learning through their magnitude of marginal regression coefficients, such that we are able to dramatically decrease the dimension of the parameter space from p to a much smaller number by the threshold parameter γ. In fact, sorting p-value is equivalent to MMLE screening in the sense that they both possess the sure independent property, the difference between these two methods is the decision of number of variables should be included. Practically, the threshold parameter γ, which is related to the minimum

strength of marginal signals in the data, needs to be estimated from the data. Overestimating γ would stop the screening while underestimating γ would result in large number of false positives. Barut, Fan and Verhaselt [5] suggest to select the γ by controlling false discovery rate (FDR). The details of selecting γ are described in Section 2.3.4.

2.2.3 Iterative SIS for longitudinal data

The key idea of SIS is to apply a single componentwise regression. Three potential issues, however, might arise with this approach. Fan and Lv [21] point out three potential problems of SIS:

1. some unimportant predictors that are highly correlated with the important predictors can have higher priority to be selected by SIS than other important predictors that are relatively weakly related to the response.

2. an important predictor that is marginally uncorrelated but jointly correlated with the response can not be picked by SIS and thus will not enter the estimated model

3. the issue of collinearity between predictors adds difficulty to the problem of variable selection.

These three issues will lead SIS procedure to miss some important predictors. To overcome this problem, an iterative SIS procedure is proposed to enhance the methodological power. It is an iterative applications of the SIS approach to feature screening.

The iterative SIS works as follows. In the first step, we select a subset of d_1 variables $\mathcal{A}_1 = \{X_{l_1}, ..., X_{l_{d_1}}\}$ using an SIS selection (either based on p-value or MMLE). Then we obtain an n-vector of the residuals from regressing the response Y over $X_{l_1}, ..., X_{l_{d_1}}$. In the next step, we treat these residuals as the new responses and apply the same method as in

the previous step to the remaining $p - d_1$ variables, which results in a subset of d_2 variables $\mathcal{A}_2 = \{X_{l_2}, ..., X_{l_{d_2}}\}$. The improvement is that fitting the residuals from the previous step on $\{X_1, ..., X_p\} \backslash \mathcal{A}_1$ can significantly weaken the priority of those unimportant variables that are highly correlated with the response through their associations with $X_{l_1}, ..., X_{l_{d_1}}$, since the residuals are uncorrelated with those selected variables in $\mathcal{A}_1$. Moreover, it enables some important predictors missed in the previous step to survive. We keep on doing this until we get L disjoint subsets $\mathcal{A}_1, ..., \mathcal{A}_L$, whose union $\mathcal{A} = \bigcup_{l=1}^{L} \mathcal{A}_l$, has a size d, which is less than n.

We now have the iterative SIS based model selection methods, which are extensions of SIS based model selection methods, for the problem of ultra-high dimensional variable selection.

2.3 Theoretical property

Sure independent screening approaches are simple and quick methods to screen out irrelevant features. In this section, we introduce a framework for establishing the sure screening property in longitudinal data.

2.3.1 Equivalence between p-value screening and MMLE screening

There are two types of sure screening criteria presented in this dissertation, one by sorting p-values, the other by sorting MMLE. In this section, we show that these two criteria are equivalent to each other and both of them possess the sure screening property in the later section.

Consider the marginal model(2.3) The null hypothesis is $H_0 : \beta_k^M = 0$ and the alternative

hypothesis is $H_a : \beta_k^M \neq 0$. The likelihood ratio test for the testing is

$$-2L_k = -2(\ell(\hat{\beta}_{0,k}^M, Y) - \ell((\hat{\beta}_{0,k}^M, \hat{\beta}_k^M), Y)),\ k = 1, \ldots, p, \tag{2.6}$$

where Y is the vector of response variable and $\ell(\cdot)$ is the log-likelihood function. Note that $\ell(\hat{\beta}_{0,k}^M, Y)$ is a constant since $\hat{\beta}_0{}^M = \arg\min_{\beta_{0,k}} \ell(\hat{\beta}_{0,k}^M, Y) = \bar{Y}$. And the p-value for the hypothesis testing is obtained by computing

$$p\text{-val}(\hat{\beta}_k^M) = 1 - \Phi(\sqrt{-2L_k}) \tag{2.7}$$

As can been seen from the above equation, the marginal p-value of $\hat{\beta}_k^M$, p-val$(\hat{\beta}_k^M)$, decreases as the marginal likelihood ratio L_k decreases. Such an independence learning ranks the importance of features according to their marginal contribution to the magnitudes of the likelihood function. The marginal p-value screening and MMLE screening share a common theoretical property.

2.3.2 Sure screening property

Since only fitting marginal regressions is a type of model misspecification to a joint regression, we need to figure out how the model information is preserved. Specifically for screening purposes, we are interested in the preservation of the nonsparsity from the joint regression to the marginal regression. Generally, the marginal regression coefficient β_k^M is different from the corresponding joint regression coefficient β_k^*. But we could expect that when $|\beta_k^*|$ exceeds a certain threshold, $|\beta_k^M|$ exceeds another threshold. The following theorem reveals that the marginal regression parameter is a measurement of the correlation between marginal covariate and the mean response function.

Theorem 1: For $k = 1, \ldots, p$, the marginal regression parameters $\beta_k^M = 0$ if and only if $cov(Y, X_k|b_k) = 0$

The proof of Theorem 1 is given in Appendix. The important variables $\{X_k : k \in \mathcal{M}_*\}$, where $\mathcal{M}_* = \{k : \beta_k \neq 0\}$, should be conditionally correlated with the response to have the sure screening property. Moreover, if X_k for $k \in M_*$ is conditionally correlated with the response, the regression coefficient β_k^M is not vanishing. This is the theoretical basis for the sure independent screening. This will be shown in Theorem 2 and requires Condition 1. And the proof will be in Appendix.

Condition 1: For some constant $c_1 > 0$ and $\kappa < \frac{1}{2}$, such that $\min_{k \in \mathcal{M}_*} |cov(Y, X_k|b_k)| \geq \frac{c_1}{n^\kappa}$.

As seen later, κ controls the rate of probability error in recovering the true sparse model. And this condition rules out the situation in which an important variable is marginally uncorrelated with Y, but jointly correlated with Y.

Theorem 2: If Condition 1 holds, for $k \in \mathcal{M}_*$, there exits a positive constant $c_2 > 0$ such that $|\beta_k^M| \geq \frac{c_2}{n^u}$.

2.3.3 Uniform convergence of MMLE

The uniform convergence of MMLE plays an important role in establishing the sure screening property of MMLE in order to control the maximum noise level relative to the true signal. In this section, we prove the convergence of marginal maximum likelihood estimator and the SIS property of longitudinal SIS method.

Since the function of log likelihood of marginal linear mixed effect model is concave, let $\mathbf{X}_k = (\mathbf{1}, X_k)$, $\mathbb{E}(\ell(\mathbf{X}_k\boldsymbol{\beta}_k, Y))$ has a unique minimizer over $\boldsymbol{\beta}_k = (\beta_{0,k}, \beta_k) \in \boldsymbol{\mathcal{B}}$ at an interior point $\boldsymbol{\beta}_k^M = (\beta_{0,k}^M, \beta_k^M)$, where $\boldsymbol{\mathcal{B}} = \left\{|\beta_{0,k}^M| \leq B, |\beta_k^M| \leq B\right\}$ for a sufficient large B is the set

over which the marginal likelihood is maximized. Several more conditions are required to obtain the uniform convergence for MMLE.

Condition 2:

(i) The operator norm of marginal Fisher's information $||I_k(\boldsymbol{\beta}_k^M)||_{\mathcal{B}}$ is bounded.

(ii) There exits an $\epsilon_1 > 0$ such that

$$\sup_{\boldsymbol{\beta}_k \in \mathcal{B}, ||\boldsymbol{\beta}_k - \boldsymbol{\beta}_k^M|| \leq \epsilon_1} |\mathbb{E}\left\{(\mathbf{X}_k \boldsymbol{\beta}_k + Zb) I(|X_k| > K_n)\right\}| \leq o(n^{-1}) \ for \ k = 1, \ldots, p \tag{2.8}$$

where K_n is a constant such that for a given $\boldsymbol{\beta}_k \in \mathcal{B}$, the function $\ell(\mathbf{X}_k \boldsymbol{\beta}_k, Y)$ is Lipschitz for all $(\mathbf{X}_k, Y)$ in $\Lambda_n = \{\mathbf{X}_k, Y : ||\mathbf{X}_k||_\infty \leq K_n, |Y| \leq K_N^*\}$ with $K_n^* = r_0 K_n^\alpha / s_0$.

(iii) For all $\boldsymbol{\beta}_k \in \mathcal{B}$, we have $\mathbb{E}\left\{\ell(\mathbf{X}_k \boldsymbol{\beta}_k, Y) - \ell(\mathbf{X}_k \boldsymbol{\beta}_k^M, Y)\right\} \geq V||\boldsymbol{\beta}_k - \boldsymbol{\beta}_k^M||$ for some positive V, bounded from below uniformly over $k = 1, \ldots, p$.

(iv) There exit some positive constant r_0, r_1, s_0, s_1 and α such that

$$P(|X_k| > t) \leq r_1 \exp(-r_0 t^\alpha) \ for \ j = 1, \ldots, p \tag{2.9}$$

for sufficient large t and that

$$\mathbb{E}\left\{\exp(\frac{(X\beta^* + Zb + s_0)^2}{2} - \frac{(X\beta^* + Zb)^2}{2})\right\} + \mathbb{E}\left\{\exp(\frac{(X\beta^* + Zb - s_0)^2}{2} - \frac{(X\beta^* + Zb)^2}{2})\right\} \leq s_1 \tag{2.10}$$

The following theorem gives the uniform convergence result of MMLEs and the sure screening property of the procedure. However, the sure screening property is not directly based on the property of the covariance matrix of predictor variables. The proof of this theorem is in the Appendix.

Theorem 3: Under Condition 2, let $k_n = K_n B + B + r_0 K_n^{\alpha}/s_0$,

(i) If $n^{1-2\kappa} k_n^{-2} K_n^{-2} \to \infty$, for a positive c_3, there exists some positive constant c satisfying

$$P(\max_{1\le k\le p} |\hat{\beta}_k^M - \beta_k^M| \ge c_3 n^{-\kappa}) \le p \exp(-c_4 n^{1-2\kappa}/k_n K_n) + pns_1 \exp(-r_0 K_n^{\alpha}) \quad (2.11)$$

(ii) If Condition 1 holds, for some constant c_5, by making $\gamma = c_5 n^{-\kappa}$ for $c_5 \le c_3/2$, we can get

$$P(\mathcal{M}_* \subset \hat{\mathcal{M}}_\gamma) \ge 1 - s \exp(-c_4 n^{1-2\kappa} (k_n K_n)^{-2}) - sns_1 \exp(-r_0 K_n^{\alpha}) \quad (2.12)$$

and $s = |\mathcal{M}_*|$ is set of nonsparse element size.

Condition 3:

(i) There exist some $\tau \ge 0$ and $c_6 > 0$ such that $\lambda_{\max}(\Sigma) \le c_6 n^{\tau}$, where $\lambda_{\max}(\cdot)$ is the function of largest eigenvalue of a matrix.

(ii) $Var(Y) = O(1)$ and $Var(X\beta^*) = \beta^{*T}\Sigma\beta^* = O(1)$, where $\Sigma = cov(X)$

Condition 3(i) rules the case of strong collinearity, Condition 3(ii) implies both $Var(y)$ and $Var(X\beta^*)$ are bounded.

2.3.4 Selection of threshold parameter in MMLE

When using controlling FDR to select the threshold parameter γ, we use the fact that

$$\left[I_k(\hat{\beta}_k^M)\right]^{\frac{1}{2}} \hat{\beta}_k^M \sim \mathcal{N}(0,1) \quad (2.13)$$

where $I_k(\hat{\beta}_k^M)$ is the element corresponding to β_k^M in the information matrix. We can select variables based on high-criticism t-test by this property. Define $\hat{\mathcal{M}}_\psi = \left\{k : \left[I_k(\hat{\beta}_k^M)\right]^{\frac{1}{2}} \hat{\beta}_k^M > \psi\right\}$ which controls the false discovery rate $\mathcal{Q}$ defined by Zhao and Li[76].

Condition 4:

(i) For any k, let e_k be the vector of residuals from marginal linear mixed effect model for X_k. There exit $c_7 > 0$ such that $var(e_k) \geq c_7$ and $\sup \mathbb{E}|e_k|^{2+\chi} < \infty$ for some $\chi > 0$.

(ii) For $k \notin \mathcal{M}_*$, we have $cov(Y, X_k|b_k) = 0$.

Theorem 4: Under Condition 2, 3, and 4, if we take $\hat{\mathcal{M}}_\psi = \left\{k : \left[I_k(\hat{\beta}_k^M)\right]^{\frac{1}{2}} \hat{\beta}_k^M > \psi\right\}$, and $\psi = \Phi^{-1}(1 - \frac{f}{2p})$ and f is the allowable number of false positives, for some positive constant c_8 we have

$$\mathbb{E}(\mathcal{Q}) \leq \frac{f}{p} + \frac{c_8}{\sqrt{n}}. \tag{2.14}$$

2.3.5 p-value screening property

We have shown the screening property of MMLE screening and the equivalence between MMLE and p-value screening. One important difference between these two criteria is the number of variable they recruit each time. We will introduce a framework to specify the p-value sure screening property in this section.

Consider model (2.2) $Y = X\beta + Zb + \epsilon$, we can rewrite it to

$$Y = X\beta + \tilde{\epsilon} \tag{2.15}$$

where $\tilde{\epsilon} = Zb + \epsilon$, and $\tilde{\epsilon} \sim \mathcal{N}(0, V)$, $V = \sigma_b^2 ZZ^T + \sigma_\epsilon^2 I_n$. For this model (2.15), the off diagonal values of covariance matrix of residuals are not all zeros since there are correlations

among repeated measures. So to turn the off diagonal values to be zeros, we may multiply a matrix to both sides of the model (2.15), just like what we usually do when using weighted least squares (WLS)

$$Y^* = X^*\beta + \epsilon^* \tag{2.16}$$

where $Y^* = V^{-\frac{1}{2}}Y$, $X^* = V^{-\frac{1}{2}}X$, and $\epsilon^* = V^{-\frac{1}{2}}\tilde{\epsilon}$. Such that the off-diagonal entries in the variance matrix of ϵ^* are zeros. Besides, the transformation of model (2.2) does not affect the coefficients.

$$Z = X^*\Sigma^{-\frac{1}{2}} \tag{2.17}$$

We assume all predictor variables $X_1^*, \ldots, X_p^*$ are scaled to have mean equal zero and standard deviation equal one. And the design matrix X^* could be factored into $Z\Sigma^{\frac{1}{2}}$.

Recall the concentration property proposed by Fan and Lv [21]: the matrix Z has the concentration property for some constants c, $c_1 > 1$ as well as a constant matrix $C_1 > 0$, then we have

$$P(\lambda_{\max}(\tilde{p}^{-1}\tilde{Z}\tilde{Z}^T) > c_1 \; and \; \lambda_{\min}(\tilde{p}^{-1}\tilde{Z}\tilde{Z}^T) < \frac{1}{c}) \leq \exp -C_1 n \tag{2.18}$$

still holds for the submatrix $\tilde{Z}$ of Z, which is $n \times \tilde{p}$, with $cn < \tilde{p} < p$. Where the $\lambda_{\min}(\cdot)$ is the smallest eigenvalues of a matrix. This concentration property suggests the same order of n nonzero singular values of $\tilde{Z}$ with high possibility. By using conditions and properties provided before, we have our Theorem 6.

Theorem 5: Under Conditions 1–3, if $2u + \tau < 1$, there exists some $\theta < 1 - 2u - \tau$, if $\gamma \sim cn^\theta$ with some positive constant c, so that for some positive C we have

$$p(\mathcal{M}_* \subset \mathcal{M}_\gamma) = 1 - O(\exp(-Cn^{1-2u}/\log(n))) \tag{2.19}$$

The SIS has the sure screening property as shown in above theorem, as well as is able to decrease the exponentially growing dimensionality p downward to $d = [\gamma n] = O(n^{1-\theta}) < n$ for some $\theta > 0$, which is close to or smaller than the sample size. While the subset $\mathcal{M}_\gamma$ remains to keep all important variables with high probability. The proof of Theorem 6 is shown in the Appendix.

2.4 Numerical Analysis

We demonstrate the performance of SIS and iterative SIS in simulated data and empirical datasets in this section. We compare four types of screening methods in various settings.

2.4.1 Simulation

We compare the performance of the proposed p-value screening and the MMLE screening, as well as their iterative extensions (iterative SIS). So we have four types of screening methods in total. We keep the sample size 200 for different scenarios and the number of predictor variables from 5,000 to 40,000. We evaluate different screening methods according to the following criteria:

1. MMMS (median minimum model size): which is the selected model that is required to include all important variables. We measure the variation of MMMS by robust standard deviation (RSD), which is computed by associate interquartile range over 1.34, which meansures the sampling variability of minimum model size.

2. FP (false positive): average number of FPs

3. FN (false negative): average number of FNs

We generate the covariate following Fan and Song [25] from

$$X_k = \frac{\epsilon_k + a_k \epsilon}{\sqrt{1 + a_k^2}} \tag{2.20}$$

where ϵ and $\{\epsilon_k\}_{k=1}^{p/3}$ are $i.i.d$ standard Gaussian distributed variables; $\{\epsilon_k\}_{k=p/3+1}^{2p/3}$ are $i.i.d$ double Exponential random variables with location equalling to zero and scale being one; and $\{\epsilon_k\}_{k=2p/3+1}^{p}$ are $i.i.d$ mixture Gaussian variables: $0.5 \times \mathcal{N}(-1, 1) + 0.5 \times \mathcal{N}(1, 0.5)$. The variables are standardized to have mean zero and variance one before being fitted to the model.

We consider the following two settings to simulate data *Setting* 1. Number of variable $p = 5,000$, and number of true signal $s = 6$ and 12. $a_1, \ldots, a_{100}$ are some constant and the chosen to be the same, such that the correlation among first 100 variables are $\rho = corr(X_k, X_l) =$ 0, 0.5, and 0.8, and $a_{101} = \cdots = a_{5000} = 0$.

Setting 2. Number of variable $p = 40,000$ and number of true signal $s = 6$, 12, and 30. $a_1, \ldots, a_{50}$ are random and generated from a Normal distribution $a_k \sim \mathcal{N}(a, 1)$ for $k = 1, \ldots, 50$, and $a_{51} = \cdots = a_{40000} = 0$. a is constant and chosen to make correlation $E[corr(X_k, X_l)] = 0$, 0.5, and 0.8 among first 50 variables.

In both of the settings, the true signals of the regression coefficients are alternating sequence of 1 and 1.3. For each setting we simulate 200 data sets. The results are shown in Tables 2.1 and 2.2. As can be seen from these tables, for SIS cases, MMLE screening performs better in terms of false negatives but has larger false positives. While in iterative SIS cases, p-value screening has smaller false negatives. And since we limit the total number of selected variables, it is understandable that why iterative SIS has smaller number of false positives. And overall, iterative SIS performs better than SIS for both MMLE and p-value screening for MMMS, number of false positives and false negatives. Interestingly, the SIS becomes

more powerful when the correlation among the predictor variables is increasing. This might be due to only 50 or 100 predictor variables are set to be correlated, therefore SIS cannot take full advantage of its benefits.

			SIS					
p	s	ρ	$MMMS_{fdr}$	FP_{fdr}	FN_{fdr}	$MMMS$	FP	FN
5000	6	0	7(42)	43.20	0.06	7(42)	32.07	0.07
5000	6	0.5	8(8.9)	137.12	0	8(8.9)	32.02	0.02
5000	6	0.8	11(10.86)	136.90	0	11(10.86)	32.04	0.04
5000	12	0	200.50(722.22)	42.86	1.51	200.50(722.22)	27.78	0.78
5000	12	0.5	48(18.72)	130.20	0	48(18.72)	27.16	0.16
5000	12	0.8	55(17.50)	129.83	0	55(17.50)	27.52	0.52
			iterative SIS					
p	s	ρ	$MMMS_{fdr}$	FP_{fdr}	FN_{fdr}	$MMMS$	FP	FN
5000	6	0	NA	31.17	0	NA	32	0
5000	6	0.5	NA	31.23	0	NA	31.93	0
5000	6	0.8	NA	31.43	0.43	NA	32.09	0.09
5000	12	0	NA	25.57	0.01	NA	26	0
5000	12	0.5	NA	25.60	0	NA	25.97	1.03
5000	12	0.8	NA	25.98	0.75	NA	28.66	2.68

Table 2.1: MMMS, and its RSD (in parentheses), FP, and FN over 200 replicates for Setting 1, $\epsilon \sim \mathcal{N}(0,1), N = 200$

2.4.2 SLE Data

In this section, we apply our methods to do feature screening with the SLE data. The data is open source and accessible at $https://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-65391/$. After screening by four types of screening methods, we applied a linear mixed effect model with an ℓ_1-penalty for further variable selection. The final selected predictor variables are shown in Table 2.5 and we can see that iterative SIS with p-value retains the most number of variables followed by iterative SIS with MMLE, and two types

			SIS					
p	s	ρ	$MMMS_{fdr}$	FP_{fdr}	FN_{fdr}	$MMMS$	FP	FN
40000	6	0	11(22.39)	47.91	0.22	11(22.39)	32.27	0.27
40000	6	0.5	8(4.47)	142.87	0	8(4.47)	32.03	0.03
40000	6	0.8	11(6.90)	142.44	0	11(6.90)	32.05	0.05
40000	12	0	23292.5(15468.84)	66.29	4.04	23292.5(15468.84)	4.54	30.54
40000	12	0.5	4491.5(15386.94)	82.16	1.40	4491.5(15386.94)	2.73	28.73
40000	12	0.8	49(429.66)	85.36	0.35	48(32.46)	2.70	28.70
40000	30	0	33848.5(8494.03)	58.19	12.38	33007.5(9707.28)	13.90	21.90
40000	30	0.5	21919.5(13507.28)	65.50	3.47	13121(15272)	7.11	15.11
40000	30	0.8	352.5(8178.17)	67.56	0.85	57(3046.27)	7.06	15.06
			iterative SIS					
p	s	ρ	$MMMS_{fdr}$	FP_{fdr}	FN_{fdr}	$MMMS$	FP	FN
40000	6	0	NA	33.82	0.63	NA	32.40	0.46
40000	6	0.5	NA	33.45	0.15	NA	32.17	0.24
40000	6	0.8	NA	33.07	0.03	NA	33.13	1.17
40000	12	0	NA	30.79	3.53	NA	29.48	3.48
40000	12	0.5	NA	27.15	0.13	NA	27.14	1.26
40000	12	0.8	NA	27.14	0.08	NA	30.22	4.22
40000	30	0	NA	25.48	15.98	NA	26.21	18.21
40000	30	0.5	NA	11.45	1.26	NA	21.68	13.68
40000	30	0.8	NA	10.25	0.30	NA	23.40	15.40

Table 2.2: MMMS, and its RSD (in parentheses), FP, and FN over 200 replicates for Setting 2, $\epsilon \sim \mathcal{N}(0,1), N = 200$

of iterative SIS keep more than two types of SIS. Then we conduct leave-one-out-cross-validation (LOOCV) based on the final selected predictor variables using linear mixed effect model, and calculate the mean squared error (MSE) for each type of screening method. Here we leave one subject out other than one observation out. Table 2.5 shows that two types of iterative SIS have smaller MSE, and p-value iterative SIS has smaller MSE than estimator iterative SIS. And as can be seen that the method with more variables tends to have smaller MSE, this also illustrate the ability of keeping important predictor variables, which meet with the result in the simulation study that iterative SIS has smallest false negatives most

			SIS					
p	s	ρ	$MMMS_{fdr}$	FP_{fdr}	FN_{fdr}	$MMMS$	FP	FN
5000	6	0	384.5(1008.58)	117.82	1.28	384.5(1008.58)	115.28	1.28
5000	6	0.5	28.5(24.07)	210.36	0	28.5(24.07)	114	0
5000	6	0.8	54.5(26.68)	211	0	54.5(26.68)	114	0
5000	12	0	1246(669.96)	119.44	2.86	1246(669.96)	110.92	2.92
5000	12	0.5	51.5(12.69)	204.24	0	51.5(12.69)	108	0
5000	12	0.8	70.5(13.06)	206.36	0	70.5(13.06)	108	0
			iterative SIS					
p	s	ρ	$MMMS_{fdr}$	FP_{fdr}	FN_{fdr}	$MMMS$	FP	FN
5000	6	0	NA	100.48	2.16	NA	116.24	2.24
5000	6	0.5	NA	102.84	3.74	NA	114.06	3.2
5000	6	0.8	NA	104.48	4.9	NA	106.52	4.92
5000	12	0	NA	97.86	4.68	NA	112.66	4.66
5000	12	0.5	NA	101.12	6.98	NA	113.8	6.06
5000	12	0.8	NA	102.92	9.52	NA	117.2	9.84

Table 2.3: MMMS, and its RSD (in parentheses), FP, and FN over 200 replicates for Setting 1, $\epsilon \sim \mathcal{N}(0, 5), N = 800$

of the time.

2.5 Discussion

This chapter explores the problem of variable screening in ultra-high dimensional longitudinal data analysis. We proposed two types of independent screening criteria by selecting certain amount of variables or by a threshold parameter. Sure independent screening (SIS) has been proved to the capability to reduce the dimensionality from the ultra-high to a relatively moderate scale that is similar or smaller than the sample size, and we try to introduce a similar procedure to longitudinal data analysis. The success of marginal screening could be an alternative option for variable screening problem as long as it is able to preserve the nonsparsity structure of the true model. SIS could also be combined with other lower

			SIS					
p	s	ρ	$MMMS_{fdr}$	FP_{fdr}	FN_{fdr}	$MMMS$	FP	FN
40000	6	0	19993.5(17840.49)	139.14	2.78	19993.5(17840.49)	116.78	2.78
40000	6	0.5	47.5(2913.06)	163.94	0.56	47.5(2913.06)	114.64	0.64
40000	6	0.8	41(10.26)	168.94	0.1	41(10.26)	114.12	0.12
40000	12	0	24718(14903.92)	142.8	5.36	24718(14903.92)	113.4	5.4
40000	12	0.5	426(11151.49)	160.72	0.98	426(11151.49)	109.02	1.02
40000	12	0.8	47(3.54)	164.06	0.12	48(32.46)	108.14	0.14
40000	30	0	31363(9784.52)	139.38	10.26	31363(9784.52)	100.54	10.54
40000	30	0.5	13223.5(18200.19)	144.22	3.47	13121(15272)	91.68	1.68
40000	30	0.8	50(23.51)	145.7	0.34	50(23.51)	90.34	0.34
			iterative SIS					
p	s	ρ	$MMMS_{fdr}$	FP_{fdr}	FN_{fdr}	$MMMS$	FP	FN
40000	6	0	NA	115.8	4.52	NA	118.4	4.4
40000	6	0.5	NA	115.18	3.94	NA	117.34	3.4
40000	6	0.8	NA	115.6	4.48	NA	118.2	4.2
40000	12	0	NA	113.86	8.84	NA	116.3	8.3
40000	12	0.5	NA	112.82	7.42	NA	114.06	6.06
40000	12	0.8	NA	113.9	8.52	NA	115.98	8.04
40000	30	0	NA	108.02	20.82	NA	109.42	19.42
40000	30	0.5	NA	102.64	15.18	NA	103.66	13.66
40000	30	0.8	NA	106.82	19.98	NA	108.76	18.76

Table 2.4: MMMS, and its RSD (in parentheses), FP, and FN over 200 replicates for Setting 2, $\epsilon \sim \mathcal{N}(0, 5), N = 800$

dimensional technique for better estimating accuracy and further variable selection.

The simulation results indicate that MMLE screening performs better than p-value screening for the SIS case in terms of false negative which makes sense, since the screening property of MMLE is not directly based on the covariance matrix of the predictor variables, for instance the operator norm growth, which, whereas, p-value does depend on. This is an advantage of MMLE screening over using the p-value. The larger false positives of MMLE screening implies it is more conservative than p-value screening, since MMLE screening recruits more variables on average and should have smaller false negative but larger false positives. And since they

Gene	SIS_{fdr}	$iter_S IS_{fdr}$	SIS	$iter_S IS$
NM_005224.2	✓		✓	
XR_037946.1				✓
XR_037042.1		✓		✓
NM_194448.1		✓		
NM_022757.3	✓	✓	✓	
XM_937023.1				✓
NR_023915.1		✓		✓
XM_497717.2				✓
NM_152236.1				✓
NM_014801.2		✓		
CB529853	✓		✓	
NM_153020.1		✓	✓	✓
NM_015596.1		✓		✓
AK098676		✓		✓
XM_945614.1		✓		✓
NM_004314.1		✓		✓
NM_002448.3				✓
XM_941900.1		✓		✓
XR_039674.1		✓		✓
NM_005570.2		✓		
NM_003343.4				✓
NM_173529.3	✓	✓	✓	✓
DA645971	✓		✓	
NM_016095.1				✓
NM_018301.2				✓
XM_929340.3		✓		✓
NM_003015.2			✓	
XM_933679.1		✓		✓
NM_025138.3				✓
XM_933070.1		✓		✓
XM_934038.2		✓		✓
NM_001125.2				✓
XM_499058.2				✓
NM_145701.1	✓	✓	✓	✓
NM_003804.3		✓		✓
XM_001713750.1		✓	✓	✓
NM_015149.3		✓		
XM_001726935.1		✓		✓
NM_003565.1	✓		✓	
NM_033256.1			✓	
BM978703	✓		✓	
NM_002021.1	✓		✓	
NM_018454.5		✓		✓
NM_001025159.1		✓		✓
XR_042416.1			✓	
NM_019086.3	✓	✓	✓	✓
XM_938701.1		✓		
XM_001133210.1				✓
BF508595		✓		
NM_022067.2		✓		✓
NM_152402.2	✓			
NM_032821.2		✓		✓
XM_496953.3		✓		✓
XM_926698.1			✓	
NM_007269.2				✓
XM_001134320.1				✓
XM_930049.1		✓		✓
NM_003450.1			✓	
$\lvert\hat{s}\rvert$	23	26	29	24
MSE	92.29	80.27	90.51	90.19

Table 2.5: The number of selected variable ($|S|$) and MSE by four types of screening methods.

both are based on the likelihood, the MMMS and RSD are exactly the same in SIS case. The iterative SIS is an extension of SIS to enhance the performance of screening. Generally speaking, the computation cost of iterative SIS is much larger than SIS, which could be one disadvantage. But iterative SIS has better ability of keeping important variables, as well as excluding unimportant variables (lower false positives). These would be some trade-offs between choosing SIS or iterative SIS. We would recommend iterative SIS if the computation time is allowed.

This study focus only on linear mixed effect model. The extension to generalized linear mixed model case could be expected. Then this linear version will be a special case with identical canonical link function. The p-value screening should be a good option since the marginal p-value should be easier to be computed.

Chapter 3

Graphical modeling for longitudinal data

In this chapter, we focus on graphical modeling for longitudinal data and aim to develop a conditional independence graph for a bunch of longitudinal variables. The proposed methodology could be used to establish graphical models for immune systems based on experimental data, and help biomedical researchers to capture complex interactions and relationships between various immunological factors.

3.1 Introduction

Graphical model is a frequently used to explore the relationship networks among a set of variables, where the variables could be genes or any other objects under study based on the research. The idea underlying graphical models is to verify the conditional independence between nodes. More specifically, the lack of edge between two nodes indicates a conditional independence between the two variables represented by the nodes given the other variables in a real system. When the graph is directed, an edge in such a graph indicates a causal relationship between the two variables or one variable "controls" the other.

The motivating example of this dissertation comes from a study of systemic lupus erythematosus (SLE). SLE is a chronic disease that affects young women with a breakdown of nucleic acid tolerance and highly varied clinical manifestations. The study was to compare transcriptomes of SLE patients at various stages of the disease, as well as to evaluate the transcriptome for individual patients over time while taking a large number of gene expression values into account. The blood transcriptome of 158 patients was collected and profiled for up to 1,412 days, representing 924 visits, in order to assess its molecular heterogeneity. 26 gene were measured for the gene expression values at each visit. The number of visits per patient ranged from 1 to 22. The repeated measures within each subject (patient) warrants the need for a longitudinal data analysis approach.

Multivariate Gaussian models have been the most popular approach for graphical models. Magwene and Kim [46], Wille and Bühlmann [67] proposed to estimate the first-order partial correlation coefficient instead of the full-order partial correlation correlation. Meinshausen and Bühlmann [51] estimated the sparse conditional independence graph using a nodewise regression method by fitting a LASSO model to each variable with other variables as predictors. Yuan and Lin [73] focused on the concentration matrix directly by minimizing a penalized likelihood function. Based on Yuan and Lin, fast algorithms (Friedman, Hastie and Tibshirani [30]; Banerjoo, Ghaoui, and D'Aspremont [4]) have also been developed using blockwise coordinate descent. Although these methods are well-developed with high accuracy and small computational cost, they are only designed for single replicate data but not for longitudinal data. One distinct feature in our data is the repeated measurements, which are collected longitudinally at a small number of time points. None of the aforementioned papers considered longitudinal data. Some recent developments (Perrin et al [54], Zou and Conzen [82]) adopted the dynamic Bayesian network approach. But this approach considered the longitudinal measurements as time series, which often requires many more sampling time points. The time-varying dynamic Bayesian networks (Song et al [58], Wang el al [66]) need even higher sampling frequencies since they aim to estimating the evolution of the network structure. But in our real data, the number of visits per patient ranged from 1 to 22, which is not able to meet the requirement of these methods. Thus, they are not so appealing to our application.

In this dissertation, we propose a nodewise graphical model for longitudinal data. We develop a penalized likelihood approach to identify the edges in the conditional independence graph for longitudinal data. We use pairwise coordinate descent combined with second order cone programming to optimize the penalized likelihood and estimate the parameters. We also extend the nodewise regression method developed by Meinshausen and Bühlmann [51] with

the node-wise model extended from an ordinary lasso regression to linear mixed-effect lasso regression. Besides, we also use a stepwise linear mixed effect model as a surrogate of LASSO in the nodewise regression method for comparison.

In our simulations, we compare the performance of our penalized likelihood method, as well as the two nodewise regression methods by checking the average number of false positive (FP), false negative (FN), true negative rate (TNR), false discovery rate (FDR), and true positive rate (TPR). The application of our methods to the SLE data provide critical information about this disease. To the best of our knowledge, it should be the first graphical model developed specifically for longitudinal data where correlation between within-subject observations cannot be ignored. The proposed methodology could be used to establish graphical models for immune systems based on experimental data, and help biomedical researchers to capture complex interactions and relationships between various immunological factors.

The rest of this article is organized as follows. We introduce the penalized likelihood method as well as the nodewise regression method of graphical modeling in Section 2. In Section 3, we conducted numerical analysis of the penalize likelihood and nodewise regression methods on simulated and real data. Section 4 is a summary of our conclusion and discussions.

3.2 Method

3.2.1 Models and notation

Consider a longitudinal study on n subjects where subject i has n_i measurements, and $N = \sum_{i=1}^{n} n_i$ is the total number of observations. Each measurement consists of p variables. Therefore, the observations for the ith subject are $\{X_{ijk}, j = 1, \ldots, p, k = 1, \ldots, n_i\}$. Then the conditional independence graph of the p variables could be conveniently expressed by a

graphical model $\mathcal{G} = (\mathcal{D}, E)$, where $\mathcal{D} = (1, \ldots, p)$ is the set of variables or the nodes and E is the set of edges in $\mathcal{D} \times \mathcal{D}$. Then the neighborhood ne_j of a node X_j, $j \in \mathcal{D}$, is the smallest subset of $\mathcal{D}_{\setminus j}$ such that X_j is conditionally independent of all other variables. In a edge set E, a pair X_u and X_v are conditionally dependent given all remaining variables $X_{\mathcal{D}_{\setminus\{u,v\}}} = \{X_j : j \in \mathcal{D} \setminus \{u, v\}\}$. Each pair of variables contained in the edge set E is conditionally dependent given all other variables. Correspondingly, this would be a zero entry in the inverse covariance matrix (or the precision matrix, $C = \Sigma^{-1}$) [39], and which also holds in longitudinal data. For each variable X_j $(1 \leq j \leq p)$, we consider a conditional mixed effects model, by treating X_j as the response variable and the remaining variables, denoted by $X_{\setminus j}$, as the predictor variables. Particularly, the model for observation X_{ijk} (the kth observation of the jth variable for subject i) is

$$X_{ijk} | \left\{ X_{\setminus j}, b_{ij} \right\} = \beta_{0j} + \sum_{l=1, l \neq j}^{p} \beta_{jl} X_{ilk} + b_{ij} + \epsilon_{ijk}, \tag{3.1}$$

$$p(X_{ijk} | \left\{ X_{\setminus j}, b_{ij} \right\}) = \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left\{ -\frac{1}{2\sigma_j^2} (X_{ijk} - \mu_{ij})^2 \right\} \tag{3.2}$$

where $\mu_{ij} = \beta_{0j} + \sum_{l=1, l \neq j}^{p} \beta_{jl} X_{ilk} + b_{ij}$ is the conditional mean whose fixed effects part contains the unknown intercept β_{0j} and the unknown coefficients β_{jl}, and random effects part is the random intercept $b_{ij} \sim N(0, \delta_j^2)$. And the random error $\epsilon_{ijk} \sim \mathcal{N}(0, \sigma_j^2)$ is independent of b_{ij}.

3.2.2 Neighborhood selection

In a conditional independence graph, there are p nodes with each representing one of the p variables. An edge between two nodes j_1 and j_2 indicates that the corresponding two variables X_{j_1} and X_{j_2} are dependent, and no edge between the two nodes means the two variables are independent. The model represented by equations (3.1) and (3.2) basically

regresses each variable against the other variables. Hence, two variables X_{j_1} and X_{j_2} are conditionally independent if and only if $\beta_{j_1 j_2} = \beta_{j_2 j_1} = 0$. Thus the neighborhood estimate of a node/variable is defined as the nonzero regression coefficient: $\beta^{(j_1)}_{\setminus j_1}$, which is the vector of fixed effect coefficients when treating X_j as the response variable with others as predictor variables. And the nonzero elements in $\beta^{(j_1)}_{\setminus j_1}$, which are the same with the corresponding nonzero elements $\{j_1 \in \mathcal{D} \setminus \{j_1\} : C_{j_1 j_2} \neq 0\}$ of the row in the precision matrix, which determines the neighbors of node j_1 precisely. Therefore, only some linear form of other variables in the neighborhood set of the variable j_1 is the best predictor for X_{j_1}. Then the set of neighbors of a node $j \in \mathcal{D}$ can thus be written as

$$\Omega_j = \left\{ j \in \mathcal{D} : \hat{\beta}^{(j)}_{\setminus j} \neq 0 \right\}. \tag{3.3}$$

This set corresponds to the set of important fixed effects in (3.1). Therefore, it is reasonable to try to exploit this relation, since the nonzero coefficients for variables in the neighborhood of the variable j precisely defines the optimal linear prediction of X_j.

When incorporating regularization regression, such as LASSO, for each node, the neighborhood estimate is defined as the nonzero regression coefficient in a penalized regression, which includes a mean squared error and a penalty function parameterized by λ. Thus, the set of neighbors of a node $j \in \mathcal{D}$ can be expressed in:

$$\hat{\Omega}^{\lambda}_j = \left\{ j \in \mathcal{D} : \hat{\beta}^{(j,\lambda)}_{\setminus j} \neq 0 \right\}. \tag{3.4}$$

where $\hat{\beta}^{(j,\lambda)}_{\setminus j}$ is the vector of fixed effect coefficients when treating X_j as the response variable and others as predictor variables with λ, each of which specifies a resulting neighborhood Ω^{λ}_j of node $j \in \mathcal{D}$. Larger value of λ leads to a sparse neighborhood for a node, while more variables would be contained in $\hat{\Omega}^{\lambda}_j$ when λ takes a very small value.

3.2.3 SOCP estimation

We shall develop a penalized pseudo-likelihood approach to simultaneously estimate the parameters and thus identify the edges in the conditional independence graph. Treating the random intercepts as observed, the complete log pseudo-likelihood takes the form

$$P_{PL}(X, b, \Theta) = \sum_{j=1}^{p} \sum_{k=1}^{n_i} \sum_{i=1}^{n} \log p(X_{ijk} | \{X_{ijk} : (u, v) \in E\}, b, \Theta) \tag{3.5}$$

where b is the vector of all the random intercepts, and Θ represents all the fixed-effects parameters (β's) in our model (3.1) and (3.2). Generally E is a much smaller set than the complete edge set $D \times D$. So we impose a penalty to enforce such edge sparsity. Then the penalized log-likelihood is

$$\sum_{j=1}^{p} \sum_{k=1}^{n_i} \sum_{i=1}^{n} (X_{ijk} - \beta_{0j} - \sum_{l=1, l \neq j}^{p} \beta_{jl} X_{ilk} - b_{ij})^2 + \lambda \sum_{1 \leq j_1 < j_2 \leq p} \sqrt{\beta_{j_1 j_2}^2 + \beta_{j_2 j_1}^2} \tag{3.6}$$

The first part of (3.6) is the mean squared errors and the second part is a penalty enforced on the paired coefficients. The form of the penalty generalizes the lasso penalty to encourage simultaneous zeros or nonzeros for $\beta_{j_1 j_2}$ and $\beta_{j_2 j_1}$. While this penalty is meaningful in checking the independence between each pair of variables, it also creates some optimization difficulty when minimizing the (3.6) directly, as there is no closed form solution for the fixed effect coefficients. Hence, we use coordinate descent algorithm to optimize each pair of $\beta_{j_1 j_2}$ and $\beta_{j_2 j_1}$ at the same time. For each pair of $\beta_{j_1 j_2}$ and $\beta_{j_2 j_1}$, the objective function is

$$\sum_{k=1}^{n_i}\sum_{i=1}^{n}(X_{ij_1k}-\beta_{0j_1}-\beta_{j_1j_2}X_{ij_2k}-\sum_{l=1,l\neq j_1,l\neq j_2}^{p}\beta_{j_1l}X_{ilk}-b_{ij_1})^2+$$
$$\sum_{k=1}^{n_i}\sum_{i=1}^{n}(X_{ij_2k}-\beta_{0j_2}-\beta_{j_2j_1}X_{ij_1k}-\sum_{l=1,l\neq j_1,l\neq j_2}^{p}\beta_{j_2l}X_{ilk}-b_{ij_2})^2+ \tag{3.7}$$
$$\lambda\sqrt{\beta_{j_1j_2}^2+\beta_{j_2j_1}^2}$$

The estimator of $b_j=(b_{1j},\ldots,b_{nj})'$ is $\hat{b_j}=D_jZ'V_j^{-1}(X_j-X_{\setminus j}\beta^{(j)})$, where $D_j=\delta_j^2I_n$, $V_j=ZD_jZ'+R_j$, and $R_j=\sigma_j^2I_N$.

In the minimization of (3.7), we need an initial estimate of $(\beta_{\setminus j_1}^{(j_1)},\beta_{\setminus j_2}^{(j_2)})$ to calculate the estimate of b_{j_1}, b_{j_2}. As we are only interested in estimating $(\beta_{j_1j_2},\beta_{j_2j_1})$, we consider other fixed effect coefficients as known, and form a new response variable d_{ij_1k}, $d_{ij_1k}=X_{ij_1k}-\beta_{0j_1}-\sum_{l=1,l\neq j_1,l\neq j_2}^{p}\beta_{j_1l}X_{ilk}-b_{ij_1}$, which enable us to consider the observations in the data are independent. We apply the second order cone programming (SOCP) to minimize

$$\sum_{k=1}^{n_i}\sum_{i=1}^{n}(d_{ij_1k}-\beta_{j_1j_2}X_{ij_2k})^2+\sum_{k=1}^{n_i}\sum_{i=1}^{n}(d_{ij_2k}-\beta_{j_2j_1}X_{ij_1k})^2+\lambda\sqrt{\beta_{j_1j_2}^2+\beta_{j_2j_1}^2} \tag{3.8}$$

With the estimation of all pairs of $(\beta_{j_1j_2},\beta_{j_2j_1})$, we can use the Best Linear Unbiased Prediction (BLUP) estimating equation to compute and update the random effect b_j's. We simultaneously get the fixed effect and random effect coefficients estimated by iterating between calculating estimates of all $\beta^{(j)}$'s and b_j's. The estimating process for a specific λ is accomplished after the convergence of fixed effect coefficients. And this process is repeated for a grid of $\lambda's$, the way of selecting the best λ will be described later. The $\beta^{(j)}$'s corresponding to the best λ is reported.

Selection of λ

The above estimation procedure is conducted for a fixed value of λ. In reality, λ is usually chosen based on a series of values and for each value of λ can we obtain a estimate of fixed and random effect coefficient. But there would be an vest λ_{op} that will provide the optimal estimates of $\beta^{(j)}$ and $b^{(j)}$. This could be achieved by minimizing a criterion such as AIC [1], BIC [57], or via k-fold Cross-Validation [26]. But AIC or BIC are not well-suited, and Cross-Validation is too time consuming. Here the criteria we use for choosing λ is the conditional AIC (cAIC), which is proposed by Hodges and Sargent [63]. cAIC is very similar to the ordinary AIC with a different way of counting the degrees of freedom in the linear model. We extend the cAIC to our longitudinal data setting by defining

$$cAIC_\lambda = -2P_{PL} + 2K \tag{3.9}$$

where K is given by

$$K = \frac{N(N-p-1)}{(N-p)(N-p-2)}(\rho+1) + \frac{N(p+1)}{(N-p)(N-p-2)} \tag{3.10}$$

and ρ is the effective number of degrees of freedom in the mixed effect model also proposed by Hodges and Sargent [33]. This is a better estimate of the degrees of freedom for the linear mixed effect model, which is also adopted for this setting. We select the solution that minimizes the $cAIC_\lambda$ criterion.

3.2.4 Nodewise LASSO

We also propose a nodewise LASSO method, which extends the work by Meinshausen and Bühlmann [51], as an alternative way to conduct longitudinal data graphical model. But we

need a linear mixed effect version of LASSO instead of ordinary LASSO [62] to accommodate for the longitudinal data. The nodewise LASSO works in the algorithm below. And the following subsections describe the way how we develop the linear mixed effect LASSO.

Nodewise LASSO procedure	
Step 1:	Fit the linear mixed effect LASSO for the j^{th} node: $LASSO(X_j, X_{\setminus j})$
Step 2:	Collect the nonzero coefficients from all the nodewise LASSO models
Step 3:	Identify edges for the j^{th} node based on variable selected in step 2
Step 4:	Repeat Step 1 - 3 for all variables

Estimation of fixed effect for LASSO

To obtain the penalized log-likelihood function, we may start with the unpenalized log-likelihood for linear mixed effect model by considering the j^{th} variable, X_j, as response variable, and other variables $X_{\setminus j}$ as predictors:

$$\mathcal{L}_0 = -\frac{1}{2\sigma_j^2}||X_j - \beta_{0j}1_N - X_{\setminus j}\beta^{(j)} - Zb^{(j)}||_2^2 - \frac{1}{2}\log|V_j| \tag{3.11}$$

where $\beta^{(j)}$ and $b^{(j)}$ are fixed and random effect coefficients when consider X_j as the response variable, $V_j = R_j + Z\Lambda_j Z^T$, $\Lambda_j = \delta_j^2 I_n$ and $R = \sigma_j^2 I_N$. σ^2 and δ^2 are respectively the residual and within-subject variances. By treating $b^{(j)}$ as observed and dropping constant term $-\frac{1}{2}ln|V_j|$ since it does not affect on selecting fixed effects, we can write the penalized complete joint log-likelihood function of (β, b) as

$$\mathcal{L}_1 = -\frac{1}{2\sigma_j^2}||X_j - \beta_{0j}1_N - X_{\setminus j}\beta^{(j)} - Zb^{(j)}||_2^2 - b^{(j)^T}\Lambda_j b^{(j)} \tag{3.12}$$

Now, the variable could be selected by maximizing conditional expectation along with a penalty function on β. By including the LASSO penalty function we obtain equation

$$\mathcal{L}_p = -\frac{1}{2\sigma_j^2}||X_j - \beta_{0j}1_N - X_{\setminus j}\beta^{(j)} - Zb^{(j)}||_2^2 - b^{(j)^T}\Lambda_j b^{(j)} - \lambda||\beta^{(j)}||_1 \tag{3.13}$$

where $0 \leq \lambda < \infty$ and $||\beta_l^{(j)}||_1 = \sum_{l=1,l\neq j}^{p} |\beta^{(j)}|$ is the ℓ_1-norm of the coefficient vector $\beta^{(j)}$. Then the variable could be selected through maximizing the penalized expectation of this joint likelihood.

The Expectation-Maximization (EM) algorithm was applied by Laird, Lange and Stram [38] in linear mixed effect model, and the observed part, the response variable, as well as unobserved part, the random effect form the complete data. The EM algorithm is applied here and the conditional expectation of $\mathcal{L}_p$ assuming the random effects are unobserved is computed. Then the penalized likelihood is minimized to update the estimates of the fixed and random effect coefficients. These steps will be repeated iteratively till the convergence of the parameters.

Given (3.13), the estimation of $\mathbf{b^{(j)}}$ is

$$\hat{b}^{(j)} = \Lambda_j Z^T V_j^{-1}(X_j - X_{\setminus j}\beta^{(j)}) \tag{3.14}$$

An initial estimate of $\beta^{(j)}$ is needed to calculate an estimate of $b^{(j)}$, $\hat{b^{(j)}}$ in the joint likelihood maximization. A new response variable $\tilde{X}_j$, which is equal to $X_j - b^{(j)}$, could be created since we are only interested in estimating $\beta^{(j)}$. It is reasonable to consider each observation in the dataset is independent from the other due to the subtraction of the random effect, which allows us to employ the LASSO variable selection procedure. Then the coordinate descent algorithm is used to estimate $\beta^{(j)}$ by fitting $\tilde{X}_j = X_{\setminus j}\beta^{(j)} + e, e \sim \mathcal{N}(0, \sigma^2)$. $\beta^{(j)}$ could be estimate by coordinate descent algorithm [29] on a fixed λ. With the estimation

of $\beta^{(j)}$, the estimating equation (3.14) is used to calculate an updated estimate of $b^{(j)}$. The estimates of fixed effects and random effects coefficients could be obtained by iterating between calculating estimates of $\beta^{(j)}$ and $b^{(j)}$. The estimation procedure is repeated for a grid of $\lambda's$, the way of choosing the optimal λ is will be described later. The $\hat{\beta}^{(j)}$ corresponding to the best λ is reported as the LASSO estimator.

Estimation of variance components

The restricted maximum likelihood estimator (REML) of θ_j, where $\theta_j = (\sigma_j^2, \delta_j^2)$, is used here, because it is more robust comparing to the maximum likelihood estimator. The log-likelihood needs to be approximated since the penalty function 3.13 does not have continuous second derivatives. By Fan and Li [26], we obtain,

$$\begin{aligned}\ell_{approx} = -\frac{1}{2}\log|V_j(\theta_j)| - \frac{1}{2}(X_j - X^c_{\setminus j}\hat{\beta}^{(j)}_c)^T V_j(\theta_j)^{-1}(X_j - X^C_{\setminus j}\hat{\beta}^{(j)}_c) \\ - \frac{1}{2}(\beta - \hat{\beta}^{(j)}_c)^T(X^c_{\setminus j}V_j(\theta_j)^{-1}X^c_{\setminus j})(\beta - \hat{\beta}^{(j)}_c) - \frac{1}{2}\beta^{(j)^T}\Sigma_\lambda(\hat{\beta}^{(j)}_c)\beta^{(j)} \quad (3.15)\end{aligned}$$

where $\hat{\beta}^{(j)}_c$ is the non-zero element in the $\beta^{(j)}$ estimate from LASSO, and $X^c_{\setminus j}$ is the subset of $X_{\setminus j}$ that corresponds to $\hat{\beta}^{(j)}_c$. And

$$\Sigma_\lambda(\hat{\beta}^{(j)}_c) = diag(\frac{p'_\lambda(|\beta^{(j)}_{1c}|)}{|\beta^{(j)}_{1c}|}, \ldots, \frac{p'_\lambda(|\beta^{(j)}_{dc}|)}{|\beta^{(j)}_{dc}|}) \quad (3.16)$$

where $\beta^{(j)}_{1c}, \ldots, \beta^{(j)}_{dc}$ are elements in $\beta^{(j)}_c$. And $p_\lambda(|\beta^{(j)}_{jc}|) = \lambda|\beta^{(j)}_{jc}|$. After some algebra, the REML log-likelihood is approximately equal to

$$\begin{aligned}\ell_{REML} = -\frac{1}{2}\log|V_j(\theta_j)| - \frac{1}{2}\log|X^c_{\setminus j}V_k(\theta_j)^{-1}X^c_{\setminus j} + \Sigma_\lambda(\hat{\beta}^{(j)}_c)| \\ - \frac{1}{2}(X_j - X^c_{\setminus j}\hat{\beta}^{(j)}_c)^T V_j(\theta_j)^{-1}(X_j - X^C_{\setminus j}\hat{\beta}^{(j)}_c) \quad (3.17)\end{aligned}$$

then the $\hat{\theta}_k$ could be estimated by Newton-Ralphson algorithm.

Selection of λ

Similar to the way of selecting the regularization parameter, λ is selected on a grid and the optimal λ will give the best estimates of $\beta^{(k)}$ and $b^{(k)}$. Here the criteria we use for choosing λ is BIC, since it is well known that BIC is consistent for model selection under general conditions. While AIC is not consistent for selection [71], though is minimax optimal. So the BIC-type criterion we apply is

$$BIC_\lambda = -2\hat{\mathcal{L}}_0 + log(n) \times (df_\lambda) \tag{3.18}$$

where the $\hat{\mathcal{L}}_0$ is the observed value of $\mathcal{L}_0$ (3.11). And We let the degrees of freedom df_λ be the total number of non-zero fixed effect coefficients in $\hat{\beta}^{(k)}$. We adopt it as an unbiased estimate of the degrees of freedom [81]. Then we select the solution that minimizes the BIC_λ criterion criterion

3.3 Numerical Analysis

We demonstrate the performance of the penalized likelihood and nodewise regression methods on simulated data and empirical datasets in this section. We add one more nodewise regression method, which is to replace the LASSO by a backwards stepwise linear mixed effect model. We compare the three graphical modeling methods in a variaty of settings.

3.3.1 Simulation

Recall that $C = Sigma^{-1}$ is the precision matrix, or the inverse of the variance covariance matrix Sigma. And the off-diagonal element of C indicates an edge between two nodes, which means these two variables represented by these two nodes are conditionally dependent.

We consider three different models following Yuan and Lin [73] in our simulation

Model 1. An AR(2) model with $C_{jj} = 1$, $C_{j,j-1} = C_{j-1,j} = 0.5$, and $C_{j,j-2} = C_{j-2,j} = 0.25$. The true number of edges is $2p - 3$.

Model 2. An AR(3) model with $C_{jj} = 1$, $C_{j,j-1} = C_{j-1,j} = 0.5$, $C_{j,j-2} = C_{j-2,j} = 0.25$, and $C_{j,j-3} = C_{j-3,j} = 0.2$. The true number of edges is $3p - 6$.

Model 3. An AR(4) model with $C_{jj} = 1$, $C_{j,j-1} = C_{j-1,j} = 0.4$, $C_{j,j-2} = C_{j-2,j} = C_{j,j-3} = C_{j-3,j} = 0.2$, and $C_{j,j-4} = C_{j-4,j} = 0.1$. The true number of edges is $4p - 10$.

For each model, we also have three different settings

Setting 1. 20 subjects with 10 measurement for each, and dimension $p = 10$

Setting 2. 50 subjects with 10 measurement for each, and dimension $p = 25$

Setting 3. 80 subjects with 10 measurement for each, and dimension $p = 40$

We compared the three approaches in terms of the number of false positives (FPs) which is incorrectly identified edges, the number of false negatives (FNs) which is incorrectly missed edges. In order to make the results comparable across models, we also compare the true negative rate ($TNR = TN/(TN + FP)$), the false discovery rate ($FDR = FP/(FP + TP)$), and true positive rate ($TPR = TP/(TP + FN)$), where TP and TN are respectively the number of true positive and true negative dependencies detected. Table 3.1, Table 3.2 and Table 3.3 document the mean, and standard deviation, in parentheses, from 200 replications

Model		NW LASSO	NW step	SOCP
	FP	11.65(4.65)	5.75(3.79)	3.63(3.60)
	FN	0.82(1.21)	0.30(0.84)	0.29(0.79)
1	TPR	0.98(0.04)	0.99(0.02)	0.99(0.02)
	FDR	0.25(0.08)	0.14(0.08)	0.09(0.08)
	TNR	0.79(0.08)	0.90(0.07)	0.94(0.06)
	FP	8.21(4.42)	4.33(2.91)	3.96(2.88)
	FN	9.84(5.05)	4.46(3.42)	4.88(3.67)
2	TPR	0.80(0.11)	0.91(0.07)	0.90(0.08)
	FDR	0.17(0.07)	0.09(0.06)	0.08(0.06)
	TNR	0.80(0.11)	0.90(0.07)	0.91(0.07)
	FP	4.58(3.42)	3.19(2.43)	2.88(2.52)
	FN	21.89(7.45)	13.60(5.38)	14.79(6.46)
3	TPR	0.64(0.12)	0.77(0.09)	0.75(0.11)
	FDR	0.10(0.06)	0.06(0.05)	0.06(0.05)
	TNR	0.85(0.11)	0.89(0.08)	0.90(0.08)

Table 3.1: Comparison of average (standard deviation) over 200 replications with $p = 10$ and $n = 200$

for each combination. The penalized likelihood procedure is referred to be SOCP in the table because of its usage of SOCP in convex optimization.

As shown in Table 3.1, Table 3.2 and Table 3.3, the penalized likelihood methods outperform the other two nodewise methods in general. Both methods of nodewise LASSO and nodewise step tend to have larger FP, which may partly explain they could sometimes enjoy a as high TPR as SOCP method, but always suffer from high FDR than SOCP method.

3.3.2 Real data example

We apply our methods to conduct graphical modeling with the SLE data in this section. The data is open source and accessible at $https://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-65391/$. We use the $p = 26$ variables that are selected by the iterative SIS because

Model		NW LASSO	NW step	SOCP
	FP	111.42(33.78)	48.17(11.82)	6.36(6.38)
	FN	1.38(1.66)	0(0)	0(0)
1	TPR	0.99(0.02)	1(0)	1(0)
	FDR	0.53(0.08)	0.33(0.06)	0.06(0.06)
	TNR	0.78(0.07)	0.90(0.02)	0.98(0.01)
	FP	52.12(31.31)	41.86(10.64)	17.91(6.71)
	FN	39.87(7.87)	0.21(0.61)	1.40(1.74)
2	TPR	0.71(0.06)	0.99(0.01)	0.99(0.01)
	FDR	0.32(0.11)	0.23(0.05)	0.11(0.04)
	TNR	0.89(0.07)	0.91(0.02)	0.96(0.01)
	FP	38.70(29.10)	37.79(10.15)	16.28(11.50)
	FN	92.39(15.93)	11.37(4.57)	26.01(8.89)
3	TPR	0.49(0.09)	0.94(0.03)	0.85(0.05)
	FDR	0.27(0.12)	0.18(0.04)	0.09(0.05)
	TNR	0.91(0.07)	0.91(0.02)	0.96(0.03)

Table 3.2: Comparison of average (standard deviation) over 200 replications with $p = 25$ and $n = 500$

it yielded the smallest testing error 2.5 in the variables selection section.

We applied our SOCP method to the data, and Figure 3.1 shows the identified networks which consists of 54 edges. The nodewise lasso and nodewise step methods were also applied to the SLE data for comparison. Figure 3.1 show the networks resulted from the two methods. The two networks are respectively composed of 68 and 27 edges. The nodewise step method is denser while the nodewise lasso is sparser than the network resulted from SOCP method.

3.4 Discussion

We have proposed a maximizing penalized likelihood method to identify the edges in graphical models for longitudinal data. Furthermore, two nodewise regression methods are also

Model		NW LASSO	NW step	SOCP
	FP	208.21(57.14)	130.92(21.60)	7.51(10.57)
	FN	0.85(1.31)	0(0)	0(0)
1	TPR	0.99(0.01)	1(0)	1(0)
	FDR	0.57(0.07)	0.46(0.04)	0.04(0.06)
	TNR	0.85(0.04)	0.91(0.02)	0.99(0.01)
	FP	95.47(58.83)	115.9(18.81)	10.93(14.38)
	FN	67.81(8.86)	0(0)	1.84(1.96)
2	TPR	0.70(0.04)	1(0)	0.99(0.01)
	FDR	0.34(0.13)	0.34(0.04)	0.04(0.05)
	TNR	0.93(0.04)	0.91(0.01)	0.99(0.01)
	FP	73.36(53.49)	107.57(17.82)	57.15(10.62)
	FN	168.31(19.47)	7.79(3.79)	18.27(5.76)
3	TPR	0.44(0.06)	0.97(0.01)	0.94(0.02)
	FDR	0.32(0.13)	0.27(0.03)	0.17(0.03)
	TNR	0.94(0.04)	0.91(0.01)	0.95(0.01)

Table 3.3: Comparison of average (standard deviation) over 200 replications with $p = 40$ and $n = 800$

proposed. The numerical examples indicate that the SOCP method outperforms the nodewise regression methods.

One important advantage of SOCP method is that its likelihood function incorporate the penalty enforced on the paired coefficients, which is a form of the penalty generalizes the lasso penalty to encourage simultaneous zeros for $\beta_{j_1 j_2}$ and $\beta_{j_2 j_1}$. Whereas, the other two nodewise methods suffer from the potential problem of asymmetry: $\beta_{j_1 j_2}$ equals to zero but $\beta_{j_2 j_1}$ does not, or vice versa. In this condition, we have to force both of them to be zero or nonzero. And the nodewise method often leads to a dense network, which can also be seen in our simulation study (large number of FPs for nodewise methods in Table 3.1, Table 3.2, and Table 3.3). Liang et al. [42] said that the unexplained signal variables will pull in other edges that would otherwise not be included because of the shrinkage of the regression coefficients on the true edge towards 0 by ℓ_1 penalty

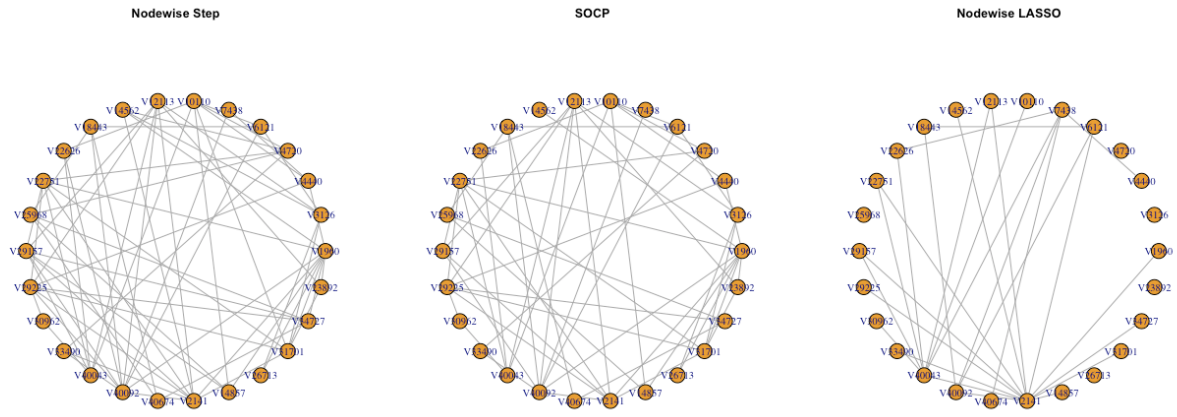

Figure 3.1: Networks identified for the SLE data by nodewise step method, SOCP method and nodewise LASSO method

One important innovation of our method is that it is the first graphical modeling method that is applicable to the longitudinal data. The within cluster variation, as well as cluster-to-cluster variation in longitudinal data add the difficulties to solve the graphical modeling methods, especially for those incorporated sample variance matrix in the penalized likelihood (for example, Yuan and Lin [73], and Friedman et al [30]). Instead of working on penalized likelihood based on precision matrix, our method focus on the regression coefficients to avoid estimating the variance matrix in longitudinal data. But graphical modeling methods using variance matrix (for either within cluster or cluster-to-cluster or both) in longitudinal data can also be expected in the future.

Appendices

Appendix A

Appendix for SIS

A.1 Proof of Theorem 1

Sufficient part:

$$\begin{aligned} Cov(Y, X_k|b_k) &= \mathbb{E}[(Y - \mathbb{E}(Y|b_k))(X_k - \mathbb{E}(X_k|b_k))] \\ &= \mathbb{E}[X_k(Y - \mathbb{E}(Y|b_k))] - \mathbb{E}[X_k|b_K]\mathbb{E}(Y - \mathbb{E}(Y|b_k))] \\ &= \mathbb{E}[X_k Y] - \mathbb{E}[X_k \mathbb{E}[Y|b_k]] \\ &= \mathbb{E}[X_k Y] - \mathbb{E}[X\beta^* X_k] = 0 \end{aligned}$$

And thus, $\mathbb{E}[X\beta^* X_k] = \mathbb{E}[X_k Y] = \mathbb{E}[X_k \beta_k^M X_k]$. The second equality is coming from score equation of the marginal regression. And $\beta_k^m = 0$ should be a solution for this equation. Due to the concaveness of likelihood, the solution to the equation should be unique. Thus,

$$Cov(Y, X_k|b_k) = 0 \;\Rightarrow\; \beta_k^M = 0$$

Necessary part:

When $\beta_k^M = 0$, the score equations now take the form

$$\mathbb{E}[\beta_{0,k}^M] = \mathbb{E}[X\beta^*],\ and \tag{A.1}$$

$$\mathbb{E}[\beta_{0,k}^M X_k] = \mathbb{E}[Y X_k] = \mathbb{E}[X\beta^* X_k] \tag{A.2}$$

$\beta_{0,k}^M$ is a constant. By the first equality of A.2, we have

$$\mathbb{E}[X_k(Y - \mathbb{E}(Y|b_k))] = 0$$

Then,

$$Cov(Y, X_k|b_k) = \mathbb{E}[(Y - \mathbb{E}(Y|b_k))(X_k - \mathbb{E}(X_k|b_k))] = 0$$

A.2 Proof of Theorem 2

By Lipschitz continuity of $\beta_{0,k}^M + \beta_{1,k}^M X_k$, we have

$$|\left\{(\beta_{0,k}^M + \beta_{1,k}^M X_k) - \beta_{0,k}^M\right\} X_k| \leq |\beta_k^M| X_k^2$$

Take expectation on both sides,

$$\begin{aligned} &|\mathbb{E}\left\{(\beta_{0,k}^M + \beta_{1,k}^M X_k) - \beta_{0,k}^M\right\} X_k| \leq |\beta_k^M| \\ &\Rightarrow \ |\beta_k^M| \geq |Cov[(\beta_{0,k}^M + \beta_{1,k}^M X_k), X_k]| \end{aligned}$$

By Condition 1, we have

$$|\beta_k^M| \geq c_1 n^{-\kappa}$$

A.3 Proof of Theorem 3

It can be noted that Condition 2(ii) is met with k_n. In fact

$$\begin{aligned} &\mathbb{E}[\ell(\beta_k, Y) - \ell(\beta_k^M, Y)](1 - I_n(X_k, Y)) \\ &\qquad \leq |\mathbb{E}[X_k\beta_k + Zb]I(|X_k| \leq K_n)| + |\mathbb{E}[X_k\beta_k^M + Zb]I(|X_k| \leq K_n)| + B(\beta_k) + B(\beta_k^M) \end{aligned}$$

where $B(\beta_k) = |\mathbb{E}[YX_k\beta_k(1-I_n(X_k,Y))]|$. By our assumption, the first two terms are in the order of $o(1/n)$, and the exponential tail Condition 2(iv) and the Cauchy-Schwarz inequality can bound the last two terms.

Using this and Theorem 1 of Fan and Song[25], letting $1+t = c_3 V n^{1/2-\kappa}/(16k_n)$ we have

$$P(|\beta_k^M - \beta_k| > c_3 n^{-\kappa}) \leq \exp(-c_4 n^{1-2\kappa}/(k_n K_n)^2) + nm_1 \exp(-m_0 K_n^\alpha)$$

Then first conclusion is shown.

Consider

$$\mathcal{A}_n \equiv \left\{ \max_{k\in\mathcal{M}_*} |\hat{\beta_k^M} - \beta_K^M| \leq c_2 n^{-\kappa}/2 \right\}$$

then the second conclusion is proved.

By Theorem 3, it holds that for all $k \in \mathcal{M}_*$,

$$|\beta_k^M| \geq c_3 n^{-\kappa}/2$$

Thus by the choice of $\gamma \leq c_3 n^{-\kappa}/2$ on the event $\mathcal{A}_n$. Using the first result and Bonferroni's inequality for all selected k could show the second statement:

$$P(\mathcal{A}_n^c) \leq s[\exp(-c_4 n^{1-2\kappa}/(k_n K_n)^2) + nm_1 \exp(-m_0 K_n^\alpha)].$$

A.4 Proof of Theorem 4

THe false discovery rate defined by Zhao and Li[76] could be expressed

$$\mathcal{Q} = \mathbb{E}(\frac{|\hat{\mathcal{M}}_\delta \cap (\mathcal{M}_*)^c|}{|(\mathcal{M}_*)^c|}) = \frac{1}{d - |\mathcal{M}_*|} \sum_{k \in (\mathcal{M}_*)^c} \mathbb{P}(\Big[I_k(\hat{\beta}_k^M)\Big]^{\frac{1}{2}} |\hat{\beta}_k^M| > \delta)$$

With the given conditions, by Theorem 1, we have $\beta_k = 0$. It is known that $(\Big[I_k(\hat{\beta}_k^M)\Big]^{\frac{1}{2}} |\hat{\beta}_k^M|$ has an asymptotically standard normal distribution. Then, it follows that for a $c_7 > 0$

$$\sup_z |\mathbb{P}(\Big[I_k(\hat{\beta}_k^M)\Big]^{\frac{1}{2}} |\hat{\beta}_k^M| > z) - \Phi(z)| \leq c_7 n^{-1/2}$$

Combining both equations, we obtain

$$\mathbb{E}(\mathcal{Q}) \leq \frac{1}{d - |\mathcal{M}_*|} \sum_{k \in (\mathcal{M}_*)^c} (2(1 - \Phi(\delta)) + c_7 n^{-1/2}).$$

Setting $\delta = \Phi^{-1}(1 - \frac{f}{2d})$ gives the result.

A.5 Proof of Theorem 5

Step 1 Consider model 2.16, data fitted to this model now have independent observations. As we know the p-value of a regression coefficient β_k^M has some relationship with the correlation between X_k and Y. The Pearson's correlation test has the expression:

$$t = \frac{\omega_k(n-2)}{\sqrt{1-\omega^2}} \to \infty \; as \; \omega_k \to 1,$$

which means,

$$p - value \to 0 \ as \ \omega_k \to 1$$

So the estimated marginal regression coefficient tends to be more significant as the correlation increases. Throughout, let $S = (Z^T Z)^+ Z^T Z$ and denote e_k as a unit vector in $\mathbb{R}^p$ with the k-th entry 1 and 0 elsewhere, $k = 1, \ldots, p$. $X = Z\Sigma^{1/2}$, by singular vector decomposition (SVM) we have

$$X^T X = p\Sigma^{1/2}\tilde{U}^T diag(\mu_1, \ldots, \mu_n)\tilde{U}\Sigma \tag{A.3}$$

where $\mu_1, \ldots, \mu_n$ are n eigenvector of $p^{-1}ZZ^T$, $\tilde{U} = (I_n, 0)_{n\times p}U$, and $U = \mathcal{O}(p)$. Since $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_p) = X^{*T}Y^*$, we have

$$\boldsymbol{\omega} = X^{*T}X^*\beta + X^{*T}\epsilon^* \hat{=} \boldsymbol{\xi} + \boldsymbol{\eta} \tag{A.4}$$

We are going to separately learn the two vectors $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$.

For vector $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_p) = X^{*T}X\beta$. When bounding $||\boldsymbol{\xi}||$, it is obvious that

$$diag(\mu_1^2, \ldots, \mu_n^2) \leq [\lambda_{\max}(p^{-1}ZZ^T)]^2 I_n$$

and $\tilde{U}\Sigma\tilde{U}^T \leq \lambda_{\max}(\Sigma)I_n$. These and A.3 lead to

$$||\xi||^2 \leq p^2\lambda_{\max}(\Sigma)[\lambda_{\max}(p^{-1}ZZ^T)]^2\beta^T\Sigma^{1/2}\tilde{U}^T\Sigma^{1/2}\beta \tag{A.5}$$

If we have $Q \in \mathcal{O}(p)$ then $\Sigma^{1/2}\beta = ||\Sigma^{1/2}\beta||Qe_1$. By Lemma 1,

$$\beta^T\Sigma^{1/2}\tilde{U}^T\Sigma^{1/2}\beta = ||\Sigma^{1/2}\beta||^2\langle Q^T SQe_1, e_1\rangle \stackrel{d}{=} ||\Sigma^{1/2}\beta||^2\langle Se_1, e_1\rangle$$

we express identical in distribution as $\stackrel{d}{=}$ for simplicity. By Condition 1 and 3, as well as Theorem 2, Lemma 4, for some $C > 0$ we have

$$P(\beta^T \Sigma^{1/2} \tilde{U}^T \Sigma^{1/2} \beta > O(\frac{n}{p})) \leq O(e^{-Cn}) \tag{A.6}$$

By Condition 3(i), $\lambda_{\max}(\Sigma) = O(n^\tau)$ and $P(\lambda_{\max}(p^{-1}ZZ^T) > c_1) \leq e^{-C_1 n}$, (A.5) and (A.6) combining with the Bonferroni's inequality give

$$P(||\xi||^2 > O(n^{1+\tau}p)) \leq O(e^{-Cn}) \tag{A.7}$$

Bounding $|\xi_k|, k \in \mathcal{M}_*$ from below. Now fix an arbitrary $k \in \mathcal{M}_*$, by A.3 we have

$$\xi_k = pe_k^t \Sigma^{1/2} \tilde{U}^T diag(\mu_1, \ldots, \mu_n) \tilde{U} \Sigma^{1/2} \beta$$

Note that $||\Sigma^{1/2} e_k|| = \sqrt{Var(X_k)} = 1$, $||\Sigma^{1/2}\beta|| = O(1)$. By Condition 1 and Theorem 2, there exits some $c > 0$ such that

$$|\langle \Sigma^{1/2}\beta, \Sigma^{1/2} e_k \rangle| = |\beta_k| |Cov(\beta_k^{-1} Y^*, X_k^*)| \geq cn^{-\kappa} \tag{A.8}$$

Hence, there exists some $Q \in \mathcal{O}(p)$ so that $\Sigma^{1/2} e_k = Qe_1$

$$\Sigma^{1/2}\beta = \langle \Sigma^{1/2}\beta, \Sigma^{1/2} e_k \rangle Qe_1 + O(1)Qe_2$$

By Lemma 1 $(\mu_1, \ldots, \mu_n)^T$ is independent from $\tilde{U}$, the uniform distribution is invariant on the orthogonal group $\mathcal{O}(p)$, we have that

$$\xi_k \stackrel{d}{=} p\langle \Sigma^{1/2}\beta\Sigma^{1/2} e_k \rangle R_1 + O(p)R_2 \hat{=} \xi_{k,1} + \xi_{k,2} \tag{A.9}$$

and $\mathbf{R} = (R_1, \ldots, R_p)^T = \tilde{U}^T diag(\mu_1, \ldots, \mu_n)\tilde{U}e_1$. We are going to check the two random vector $\xi_{k,1}$ and $\xi_{k,2}$.

$$R_1 \geq e_1^T \tilde{U}^T \lambda_{\min}(p^{-1}ZZ^T) I_n \tilde{U} e_1 = \lambda_{\min}(p^{-1}ZZ^T)\langle Se_1, e_1\rangle,$$

and thus by Lemma 4, Condition Property C, and Bonferroni's inequality, for some positive c and C we have,

$$P(R_1 < cn/p) \leq O(e^{-Cn})$$

Combining this with (A.8), for some positive constant c we have,

$$P(|\xi_{k,1}| < cn^{1-\kappa}) \leq O(e^{-Cn}) \tag{A.10}$$

Likewise, we can have

$$P(||\mathbf{R}||^2 > O(n/p)) \leq O(e^{-Cn}) \tag{A.11}$$

by Lemma 1 $(\mu_1, \ldots, \mu_n)^T$ is independent from $\tilde{U}$, and $\tilde{\mathbf{R}} = (R_2, \ldots, R_p)^T$ is invariant using the argument used in proving Lemma 5 given orthogonal group $\mathcal{O}(p-1)$. Hence, we have that $\tilde{\mathbf{R}} \stackrel{d}{=} ||\tilde{\mathbf{R}}||\mathbf{W}/||\mathbf{W}||$, $\mathbf{W} = (W_1, \ldots, W_{p-1})^T \sim \mathcal{N}(0, I_{p-1})$, independent from $||\mathbf{R}||$. Therefore,

$$R_2 \stackrel{d}{=} ||\tilde{\mathbf{R}}||W_1/||\mathbf{W}||. \tag{A.12}$$

For (A.11) and (A.12), and $\xi_{k,2} = O(pR_2)$, for some positive c, using the the argument in proving Lemma 5, we have

$$P(|\xi_{k,2}| > c\sqrt{n}|W|) \leq O(e^{-Cn}), \tag{A.13}$$

where W is a standard Normal distributed random variable. Based in the classical Gaussian

tail bound, by letting $x_n = c\sqrt{2C}n^{1-\kappa}/\sqrt{\log(n)}$ give

$$P(c\sqrt{n}|W| > x_n) \leq \sqrt{2/\pi}\frac{\exp(-Cn^{1-2\kappa}/\log(n))}{\sqrt{2C}n^{1/2-\kappa}/\sqrt{\log(n)}} = O(\exp(-Cn^{1-2\kappa}/\log(n)))$$

combining with (A.13) and Bonferroni's inequality yield

$$P(|\xi_{k,2}| > x) \leq O(\exp(-Cn^{1-2\kappa}/\log(n))) \tag{A.14}$$

Therefore, by Bonferroni's inequality, combining A.9, A.10, along with A.14, for some positive c we have,

$$P(|\xi_k| < cn^{1-\kappa}) \leq O(\exp(-Cn^{1-2\kappa}/\log(n))),\ k \in \mathcal{M}^*. \tag{A.15}$$

Then we will check the vector $\boldsymbol{\eta} = (eta_1, \ldots, eta_p)^T = X^{*T}\epsilon^*$.

Bounding $||\boldsymbol{eta}||$ we have

$$X^*X^{*T} = Z\Sigma Z^T \leq Z\lambda_{\max}(\Sigma)I_pZ^T = p\lambda_{\max}(\Sigma)\lambda_{\max}(p^{-1}ZZ^T)I_n.$$

we have that

$$||\boldsymbol{\eta}||^2 = \epsilon^{*T}X^*X^{*T}\epsilon^* \leq p\lambda_{\max}(\Sigma)\lambda_{\max}(p^{-1}ZZ^T)||\epsilon^*||^2 \tag{A.16}$$

And we know that $\epsilon_1^{*2}/\sigma^2, \ldots, \epsilon_n^{*2}/\sigma^2$ are *i.i.d.* χ^2 variables. By Lemma 3, for some positive c and C we have

$$P(||\epsilon^*||^2 > cn\sigma^2) \leq e^{-Cn}$$

combining with A.16, Bonferroni's inequality, and Condition 3(i) we have

$$P(||\boldsymbol{\eta}||^2 > O(n^{1+tau}p)) \leq O(e^{-Cn}). \tag{A.17}$$

Bounding $|\eta_k|$ from above, $\boldsymbol{\eta} = X^{*T}\epsilon^* \sim \mathcal{N}(0, \sigma^2 X^{*T}X^*)$. Thus $\eta_k|X \sim \mathcal{N}(0, var(\eta_k|X))$ and

$$Var(\eta_k|X) = \sigma^2 e_k^T X^{*T} X e_k.$$

Denote $\mathcal{E}$ as $\{Var(\eta_k|X) \leq cn\}$ for some positive c. Likewise, we can prove that for some positive C,

$$P(\mathcal{E}^c) \leq O(e^{-Cn}) \tag{A.18}$$

yields

$$P(|\eta_k| > x|X) \leq P(\sqrt{cn}|W| > x) \ for \ any \ x > 0, \tag{A.19}$$

on the event $\mathcal{E}$, where W is a standard Normal random variable. Hence, by (A.18) and (A.19) we have

$$P(|\eta_k| > x) \leq O(e^{-Cn}) + P(\sqrt{cn}|W| > x). \tag{A.20}$$

Let $x'_n = \sqrt{2cC}n^{1-\kappa}/\sqrt{\log(n)}$, such that

$$P(\sqrt{cn}|W| > x'_n) = O(\exp(-Cn^{1-2\kappa}/\log(n)))$$

by the classical Gaussian tail bound again, combining which with (A.20) do we have

$$P(|\eta_k| > o(n^{1-\kappa})) \leq O(\exp(-Cn^{1-2\kappa}/\log(n))) \tag{A.21}$$

Step 1.3

Combining the result we've got in Step 1.1 and Step 1.2, and Bonferroni's inequality, we have from A.4, A.7, A.15, A.21 that for some positive c_1, c_2, C,

$$P(\min_{k\in\mathcal{M}_*} |\omega_k| < c_1 n^{1-\kappa} or ||\omega||^2 > c_2 n^{1+\tau} p) \leq O(s\exp(-Cn^{1-2\kappa}/\log(n))) \tag{A.22}$$

which shows with high probability of $1 - O(s\exp(-Cn^{1-2\kappa}/\log(n)))$, the magnitudes of $\omega_k, k \in \mathcal{M}_*$, are in the order of $n^{1-2\kappa}$ and, for some positive c,

$$\sharp\left\{1 \leq k \leq p : |\omega_k| \geq \min_{k\in\mathcal{M}_*} |\omega_k|\right\} \leq c\frac{n^{1+\tau}p}{n^{1-\kappa^2}} \tag{A.23}$$

where $\sharp\{\cdot\}$ denotes the number of elements in a set.

We can see that from (A.23) if δ satisfies $\delta n^{1-2\kappa-\tau} \to \infty$ ad $n \to \infty$, then Theorem 1 is valid for some positive $C > 0$ greater than those in (A.22).

Step 2

For $r \in (0, 1)$, and choose $(\frac{n}{p})^{\frac{1}{h}-r}$, for some integer $h \geq 1$. In addition, fix an arbitrary $\theta_1 \in (0, 1 - 2\kappa - \tau)$, also take $r < 1$ that is near 1, then $\theta_0 = \theta_1/r < 1 - 2\kappa - \tau$. We can select a grid of integers $h \geq 1$ so that

$$\delta n^{1-2\kappa-\tau} \to \infty \; and \; \delta n^{\theta_0} \to 0 \; as \; n \to \infty, \tag{A.24}$$

where $\delta = (\frac{n}{p})^{\frac{1}{h}-r}$. Condition 3 (i) guarantee that

$$\lambda_{\max}(\Sigma^0) \leq \lambda_{\max}(\Sigma) \leq c_4 n^{\tau}$$

for Σ^0 of Σ to a corresponding subset of variables. By Property C, and follows from A.24 that $h = O(\log(p)/\log(n)$, in the order of $O(n^{\xi}/\log(n))$ by Condition 1. Hence, an increase of $C > 0$ gives

$$P(\mathcal{M}_* \subset \mathcal{M}_\gamma) = 1 - O(\exp(-Cn^{1-2\kappa}/\log(n))).$$

This probability bound hold for $\gamma \sim cn^{-\theta}$, and $\theta < 1 - 2\kappa - r$ and $c > 0$.

Bibliography

[1] H. Akaike. *Information theory and an extension of the maximum likelihood principle. In: Second International Symposium on Information Theory, vol. 1, 267-281.* Akademinai Kiado, Budapest, 1973.

[2] F. Bach. Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008.

[3] Romain Banchereau, Seunghee Hong, Brandi Cantarel, Nicole Baldwin, Jeanine Baisch, Michelle Edens, Alma-Martina Cepika, Peter Acs, Jacob Turner, Esperanza Anguiano, et al. Personalized immunomonitoring uncovers molecular networks that stratify lupus patients. *Cell*, 165(3):551–565, 2016.

[4] Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine learning research*, 9(Mar):485–516, 2008.

[5] Emre Barut, Jianqing Fan, and Anneleen Verhasselt. Conditional sure independence screening. *Journal of the American Statistical Association*, 111(515):1266–1277, 2016.

[6] Howard D Bondell, Arun Krishna, and Sujit K Ghosh. Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics*, 66(4):1069–1077, 2010.

[7] Peter Bühlmann, Markus Kalisch, and Marloes H Maathuis. Variable selection in high-dimensional linear models: partially faithful distributions and the pc-simple algorithm. *Biometrika*, 97(2):261–278, 2010.

[8] Robert Castelo and Alberto Roverato. A robust procedure for gaussian graphical model search from microarray data with p larger than n. *Journal of Machine Learning Research*, 7(Dec):2621–2650, 2006.

[9] Robert Castelo and Alberto Roverato. Reverse engineering molecular regulatory networks from microarray data with qp-graphs. *Journal of Computational Biology*, 16(2):213–227, 2009.

[10] J. Chen and Z. Chen. Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95:759–771, 2008.

[11] Zhen Chen and David B Dunson. Random effects selection in linear mixed models. *Biometrics*, 59(4):762–769, 2003.

[12] H. Cui, R. Li, and W. Zhong. Model-free feature screening for ultra-high dimensional discriminant analysis. *Journal of the American Statistical Association*, 110:630–641, 2015.

[13] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, 57:1413–1457, 2004.

[14] Arthur P Dempster. Covariance selection. *Biometrics*, pages 157–175, 1972.

[15] Peter Diggle. *Analysis of longitudinal data*. Oxford University Press, 2002.

[16] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression (with discussion). *The Annals of Statistics*, 32:407–499, 2004.

[17] J. Fan and Y. Fan. High dimensional classification using features annealed independence rules. *The Annals of Statistics*, 36:2605–2637, 2008.

[18] J. Fan, Y. Feng, and R. Song. Nonparametric independence screening in sparse ultra-high dimensional additive models. *Journal of the American Statistical Association*, 106:544–557, 2011.

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

[20] J. Fan and R. Li. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20:101–148, 2010.

[21] J. Fan and J. Lv. Sure independence screening for ultra-high dimensional feature space. *Journal of the Royal Statistical Society, Series B*, 70:849–911, 2008.

[22] J. Fan and J. Lv. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20:101–148, 2010.

[23] J. Fan, Y. Ma, and W. Dai. Nonparametric independence screening in sparse ultra-high-dimensional varying coefficient models. *Journal of the American Statistical Association*, 109:1270–1284, 2014.

[24] J. Fan, R. Samworth, and Y. Wu. Ultra-high dimensional feature selection: beyond the linear model. *Journal of Machine Learning Research*, 10:2013–2038, 2009.

[25] J. Fan and R. Song. Sure independence screening in generalized linear models with np-dimensionality. *The Annals of Statistics*, 38:3567–3604, 2010.

[26] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

[27] Yingying Fan and Runze Li. Variable selection in linear mixed effects models. *Annals of statistics*, 40(4):2043, 2012.

[28] L. Frank and J. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35:109–135, 1993.

[29] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

[30] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[31] W. Fu. Penalized regressions: the bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7:397–416, 1998.

[32] P. Hall and H. Miller. Using generalized correlation to effect variable selection in very high dimensional problems. *Journal of Computational and Graphical Statistics*, 18:533–550, 2009.

[33] James S Hodges and Daniel J Sargent. Counting degrees of freedom in hierarchical and other richly-parameterised models. *Biometrika*, 88(2):367–379, 2001.

[34] D. Huang, R. Li, and H. Wang. Feature screening for ultrahigh-dimensional categorical data with applications. *Journal of Business & Economic Statistics*, 32:237–244, 2014.

[35] J. Huang, S. Ma, and C. Zhang. Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica*, 18:1603–1618, 2008.

[36] Joseph G Ibrahim, Hongtu Zhu, Ramon I Garcia, and Ruixin Guo. Fixed and random effects selection in mixed effects models. *Biometrics*, 67(2):495–503, 2011.

[37] Y. Kim, J. Kim, and Y. Kim. Blockwise sparse regression. *Statistica Sinica*, 16:375–390, 2006.

[38] Nan Laird, Nicholas Lange, and Daniel Stram. Maximum likelihood computations with repeated measures: application of the em algorithm. *Journal of the American Statistical Association*, 82(397):97–105, 1987.

[39] Steffen L Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.

[40] G. Li, H. Peng, J. Zhang, and L. Zhu. Robust rank correlation based screening. *The Annals of Statistics*, 40:1846–1877, 2012.

[41] R. Li, W. Zhong, and L. Zhu. Feature screening via distance correlation learning. *Journal of the American Statistical Association*, 107:1129–1139, 2012.

[42] Faming Liang, Qifan Song, and Peihua Qiu. An equivalent measure of partial correlation coefficients for high-dimensional gaussian graphical models. *Journal of the American Statistical Association*, 110(511):1248–1265, 2015.

[43] Hua Liang, Hulin Wu, and Guohua Zou. A note on conditional aic for linear mixed-effects models. *Biometrika*, 95(3):773–778, 2008.

[44] J. Liu, R. Li, and R. Wu. Feature selection for varying coefficient models with ultrahigh-dimensional covariates. *Journal of the American Statistical Association*, 109:266–274, 2014.

[45] Y. Liu and Y. Wu. Variable selection via a combination of the l_0 and l_1 penalties. *Journal of Computational and Graphical Statistics*, 16:782–798, 2007.

[46] Paul M Magwene and Junhyong Kim. Estimating genomic coexpression networks using first-order conditional independence. *Genome biology*, 5(12):R100, 2004.

[47] Q. Mai and H. Zou. The kolmogorov filter for variable screening in high-dimensional binary classification. *Biometrika*, 100:229–234, 2013.

[48] C. Mallows. Some comments on cp. *Technometrics*, 15:661–675, 1973.

[49] Rahul Mazumder and Trevor Hastie. The graphical lasso: New insights and alternatives. *Electronic journal of statistics*, 6:2125, 2012.

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

[51] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, pages 1436–1462, 2006.

[52] Y. Nardi and A. Rinaldo. On the asymptotic properties of the group lasso estimator for linear models. *Electronic Journal of Statistics*, 2:605–633, 2008.

[53] M. Osborne, B. Presnell, and B. Turlach. A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis*, 20:389–404, 2000.

[54] Bruno-Edouard Perrin, Liva Ralaivola, Aurelien Mazurie, Samuele Bottani, Jacques Mallet, and Florence d'Alche Buc. Gene networks inference using dynamic bayesian networks. *Bioinformatics*, 19(suppl_2):ii138–ii148, 2003.

[55] Wenji Pu and Xu-Feng Niu. Selecting mixed-effects models based on a generalized information criterion. *Journal of multivariate analysis*, 97(3):733–758, 2006.

[56] Jürg Schelldorfer, Lukas Meier, and Peter Bühlmann. Glmmlasso: an algorithm for high-dimensional generalized linear mixed models using l1-penalization. *Journal of Computational and Graphical Statistics*, 23(2):460–477, 2014.

[57] H. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.

[58] Le Song, Mladen Kolar, and Eric P Xing. Time-varying dynamic bayesian networks. In *Advances in neural information processing systems*, pages 1732–1740, 2009.

[59] Peter Spirtes, Clark N Glymour, Richard Scheines, David Heckerman, Christopher Meek, Gregory Cooper, and Thomas Richardson. *Causation, prediction, and search*. MIT press, 2000.

[60] C. Stone. Additive regression and other nonparametric models. *The Annals of Statistics*, 13:6897–705, 1985.

[61] Gbor J Szkely, Maria L Rizzo, and Nail K Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, 2007.

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

[63] Florin Vaida and Suzette Blanchard. Conditional akaike information for mixed-effects models. *Biometrika*, 92(2):351–370, 2005.

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

[65] L. Wang, G. Chen, and H. Li. Group scad regression analysis for microarray time course gene expression. *Bioinformatics*, 23:1486–1494, 2007.

[66] Zhaowen Wang, Ercan E Kuruoglu, Xiaokang Yang, Yi Xu, and Thomas S Huang. Time varying dynamic bayesian network for nonstationary events modeling and online inference. *IEEE Transactions on Signal Processing*, 59(4):1553–1568, 2011.

[67] Anja Wille and Peter Bühlmann. Low-order conditional independence graphs for inferring genetic networks. *Statistical applications in genetics and molecular biology*, 5(1), 2006.

[68] S. Wu, X. Shen, and C. Geyer. Adaptive regularization using the entire solution surface. *Biometrika*, 96:513–527, 2009.

[69] T. Wu and K. Lange. Coordinate descent procedures for lasso penalized regression. *The Annals of Applied Statistics*, 2:224–244, 2008.

[70] C. Xu and J. Chen. The sparse mle for ultrahigh-dimensional feature screening. *Journal of the American Statistical Association*, 109:1257–1265, 2014.

[71] Yuhong Yang. Can the strengths of aic and bic be shared? a conflict between model indentification and regression estimation. *Biometrika*, 92(4):937–950, 2005.

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

[73] Ming Yuan and Yi Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

[74] C. Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38:894–942, 2010.

[75] P. Zhao, G. Rocha, and B. Yu. Grouped and hierarchical model selection through composite absolute penalties. *The Annals of Statistics*, 37:3468–3497, 2009.

[76] Sihai Dave Zhao and Yi Li. Principled sure independence screening for cox models with ultra-high-dimensional covariates. *Journal of multivariate analysis*, 105(1):397–411, 2012.

[77] L. Zhu, L. Li, R. Li, and L. Zhu. Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, 106:1464–1475, 2011.

[78] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429, 2006.

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

[80] H. Zou and R. Li. One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics*, 36:1509–1533, 2008.

[81] Hui Zou, Trevor Hastie, Robert Tibshirani, et al. On the "degrees of freedom" of the lasso. *The Annals of Statistics*, 35(5):2173–2192, 2007.

[82] Min Zou and Suzanne D Conzen. A new dynamic bayesian network (dbn) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics*, 21(1):71–79, 2004.