Building Generalized Linear Models with Ultrahigh Dimensional Features: A Sequentially Conditional Approach

Qi Zheng

Department of Bioinformatics and Biostatistics, University of Louisville, Louisville, USA

email: qi.zheng@louisville.edu

and

Hyokyoung G. Hong

Department of Statistics and Probability, Michigan State University, East Lansing, USA

email: hhong@msu.edu

and

Yi Li

Department of Biostatistics, University of Michigan, Ann Arbor, USA

email: yili@umich.edu

SUMMARY: Conditional screening approaches have emerged as a powerful alternative to the commonly used marginal screening as they can identify marginally weak but conditionally important variables. However, most existing conditional screening methods need to fix the initial conditioning set, which may determine the ultimately selected variables. If the conditioning set is not properly chosen, the methods may produce false negatives and positives. Moreover, screening approaches typically need to involve tuning parameters and extra modeling steps in order to reach a final model. We propose a sequential conditioning approach by dynamically updating the conditioning set with an iterative selection process. We provide its theoretical properties under the framework of generalized linear models. Powered by an extended Bayesian information criterion as the stopping rule, the method will lead to a final model without the need to choose tuning parameters or threshold parameters. The practical utility of the proposed method is examined via extensive simulations and analysis of a real clinical study on predicting multiple myeloma patients' response to treatment based on their genomic profiles.

Biometrics 000, 000–000 DOI: 000

000 0000

Key words: extended Bayesian information criteria; high-dimensional predictors; predictive modeling; sequential conditioning; sure screening properties.

1. Introduction

With advent of treatment options in cancer, gene expression profiles have emerged as an important tool in predicting therapeutic response (Amin et al., 2014). In a multiple myeloma study that motivates this work (Mulligan et al., 2007), the prediction of the binary response status (1=complete response and 0=otherwise) based on massive genetic biomarkers requires the development of a parsimonious logistic regression model, or more broadly, a generalized linear model (GLM) (Paul et al., 2008). As the dimension of predictors defies any existing modeling approaches, feature screening has been commonly used for dimension reduction. The most popular screening approach is marginal screening (Fan and Lv, 2008), which selects variables based on their marginal associations with the response. However, marginal screening may miss signals which are marginally unimportant but conditionally important (Barut et al., 2016), resulting in biased predictive results (Li et al., 2019).

To resolve this issue, a number of authors have proposed conditional screening approaches under the GLM framework: Fan and Lv (2008) suggested an iterative procedure by repeatedly using the residuals from the previous iterations; Xu and Chen (2014) proposed a sparsity restricted maximum likelihood estimation method, which retains the virtues of the iterative procedure but is conceptually simpler and computationally efficient than the iterative procedures; Barut et al. (2016) proposed a conditional screening approach, given some important variables known *a priori*; Hong et al. (2016) further introduced a data-driven conditional screening approach in the absence of prior knowledge about the conditioning set.

These aforementioned methods have various drawbacks. First, most of the conditional screening approaches need to fix the initial choice of the conditioning set and the selected variables may depend on the conditioning set. In the absence of reliable information about the conditioning set, the methods may produce false negatives and positives. Second, the theoretical properties for the data-driven approaches are still elusive, making it difficult to

evaluate their generalizability. Third, most of these methods require the selection of tuning parameters, which is often computationally clumsy.

We propose a sequential conditioning (SC) approach, wherein variables sequentially enter the conditioning set according to the increment of likelihood. The procedure updates the conditioning set at each iteration based on the extended Bayesian information criterion (EBIC) (Chen and Chen, 2008), and constructs an offset term based on the variables in this set. In essence, this offset summarizes the information contained in the updated conditioning set and we search for a new variable that maximizes the likelihood given the offset term. We emphasize that the proposed sequential conditioning approach deviates fundamentally from the variable screening or selection approaches as it naturally leads to a final model when the procedure terminates.

In addition, our approach is innovative in several aspects. First, compared to the other conditional approaches, it is computationally efficient as it maximizes the likelihood with respect to only one covariate at each step given the offset. Second, the use of the EBIC accommodates natural selection of the final model without requiring tuning parameters or threshold parameters. Third, in contrast with marginal screening, the proposed method does not require restrictive faithfulness assumptions which stipulate that marginal models must reflect the original model. Fourth, we have established rigorous selection consistency results with the EBIC and showed that, if the dimension of the true model is finite, the proposed approach can discover all relevant predictors within a finite number of steps. The derived theoretical framework can accommodate a wide range of data types, such as binary, categorical and count data. Finally, the proposed approach starts with an empty model or some important variables identified *a priori* and then sequentially recruits more variables into the conditioning set, and our method is valid even in the absence of the prior information about which variables to condition on.

The rest of the paper is organized as follows. In Section 2, we introduce the proposed sequential conditioning procedure. In Section 3, we establish the sure screening property. Section 4 details the assessment of the finite sample performance of the proposed method and Section 5 illustrates our method by predicting treatment response based on myeloma patients' genomic profiles using the aforementioned data example. We conclude the paper with a brief discussion in Section 6 and relegate all the technical details, including lemmas, conditions and proofs, to the online Supporting Information.

2. Sequentially Conditional Modeling

Suppose that there are n independent samples $(\mathbf{X}_i, Y_i), i = 1, \ldots, n$, where Y_i is an outcome, $\mathbf{X}_i = (X_{i0}, X_{i1}, \ldots, X_{ip})^{\mathrm{T}}$ is a collection of $p+1$ predictors for the ith sample, and $X_{i0} = 1$ corresponds to the intercept. Assume without loss of generality that all the covariates have been standardized so that $E(X_{ij}) = 0$ and $E(X_{ij}^2) = 1$ for all $j \geqslant 1$. We focus on a class of GLMs by assuming that the conditional density of Y_i given $\mathbf{X}_i$ belongs to the linear exponential family:

$$\pi(Y_i \mid \mathbf{X}_i) = \exp\{Y_i \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta} - b(\mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta}) + A(Y_i)\}, \tag{1}$$

where $A(\cdot)$ and $b(\cdot)$ are some known functions, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^{\mathrm{T}}$ represents the coefficients of predictors, and β_0 is the intercept. Compared to the usual exponential family (McCullagh and Nelder, 1989), (1) adopts a canonical link function and an unit dispersion parameter for simplicity of presentation. We assume that function $b(\cdot)$ is twice continuously differentiable with a non-negative second derivative $b''(\cdot)$, and use $\mu(\cdot)$ and $\sigma(\cdot)$ to denote $b'(\cdot)$ and $b''(\cdot)$, respectively. For a non-random function $f(\cdot)$ and a sequence of independent random variables ξ_i $(i = 1, \ldots, n)$, let $\mathbb{E}_n\{f(\xi)\} = n^{-1}\sum_{i=1}^n f(\xi_i)$ be the mean of $\{f(\xi_i)\}_{i=1}^n$, which are independent replicates of $f(\xi)$. We also denote the empirical process by $\mathbb{G}_n\{f(\xi)\} = n^{-1/2}\sum_{i=1}^n (f(\xi_i) - E[f(\xi_i)])$. Further assume that $\{X_{ij}, \mathbf{X}_i, Y_i\}$ are

independently and identically distributed copies of $\{X_j, \mathbf{X}, Y\}$. We let $X_0 = 1$, corresponding to the intercept. When $p \geqslant n$, regularization estimation is often carried out under a sparsity assumption on the predictors. When p is on the exponential order of n, a popular approach for reducing the dimensionality is screening.

The log-likelihood function, apart from an additive constant, is

$$\frac{1}{n}\sum_{i=1}^{n} l(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}, Y_i) = \mathbb{E}_n\{l(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}, Y)\}, \tag{2}$$

where $l(\theta, y) = y\theta - b(\theta)$. For example, for logistic regression, $b(\theta) = \log\{1 + \exp(\theta)\}$ and the log-likelihood is equal to $n^{-1}\sum_{i=1}^{n}[Y_i\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta} - \log\{1 + \exp(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})\}]$. Denote by $\boldsymbol{\beta}_* = (\beta_{0*}, \beta_{1*}, \ldots, \beta_{p*})^{\mathrm{T}}$ the true values of $\boldsymbol{\beta}$, and denote the true model as $\mathcal{M} = \{j : \beta_{j*} \neq 0, j \geqslant 1\} \cup \{0\}$. We denote its estimate by $\widehat{\mathcal{M}}$.

More notation is introduced. For an index set $S \subset \{0, 1, \ldots, p\}$ and a p-dimensional vector $\mathbf{A}$, we use $\mathbf{A}_S = \{A_j : j \in S\}$ to denote the subvector of $\mathbf{A}$ corresponding to S. For example, $\mathbf{X}_{iS}$ denotes the collection of covariates for the ith individual corresponding to S. We use $|S|$ to denote the cardinality of S and S^c to denote the complement of S. We use $\ell_S(\boldsymbol{\beta}_S) := \mathbb{E}_n\{l(\mathbf{X}_S^{\mathrm{T}}\boldsymbol{\beta}_S, Y)\}$ to denote the average log-likelihood of the regression model of Y on $\mathbf{X}_S$ for a given $S \subset \{0, 1, \ldots, p\}$, and use $\widehat{\boldsymbol{\beta}}_S$ to denote the maximizer of $\ell_S(\boldsymbol{\beta}_S)$.

We elaborate on the idea of building model (1) with the proposed sequential conditioning approach. The key is to include an offset term which summarizes the information acquired from the previous selection steps and to search for a new candidate variable that maximizes the likelihood with such an offset.

Specifically, we denote by O_k the offset evaluated at the kth step and $S_k \subset \{0, 1, \ldots, p\}$ the set of indices of the covariates selected up to the kth step. Initializing $S_0 = \{0\}$, we set $O_0 = \widehat{\beta}_{S_0}$, where $\widehat{\beta}_{S_0}$ maximizes $\ell_{S_0}(\beta_{S_0})$ and is the estimated intercept without any other covariates. That is, we start from the null model with only an intercept term. We can also start with a set of given variables according to some *a priori* knowledge, which is in the

same spirit as conditional screening (Barut et al., 2016). However, as opposed to Barut et al. (2016), our procedure dynamically updates the conditioning set with a sequential selection process, which is detailed below.

First, with such an O_0, for $j \in \{1, \ldots, p\}$, we compute $\widehat{\beta}_j^{(1)} = \arg\max_\beta \ell_{O_0,j}(\beta)$, where $\ell_{O,j}(\beta) = \mathbb{E}_n\{l(O + \beta X_j, Y)\}$. Then $j_1 = \arg\max_{j\in\{1,\ldots,p\}} \ell_{O_0,j}(\widehat{\beta}_j^{(1)})$. Now set $S_1 = \{0, j_1\}$ and regress Y on $\mathbf{X}_{S_1}$ to obtain $\widehat{\boldsymbol{\beta}}_{S_1}$. Set $O_1 = \mathbf{X}_{S_1}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{S_1}$, which is embedded with information for the variable selected previously.

An iterative procedure follows naturally. For $k \geqslant 1$, given O_k and S_k, we compute $\widehat{\beta}_j^{(k+1)} = \arg\max_\beta \ell_{O_k,j}(\beta)$ for $j \in S_k^c$. Then $j_{k+1} = \arg\max_{j\in S_k^c} \ell_{O_k,j}(\widehat{\beta}_j^{(k+1)})$. Now set $S_{k+1} = S_k \cup \{j_{k+1}\}$ and regress Y on $\mathbf{X}_{S_{k+1}}$ to obtain $\widehat{\boldsymbol{\beta}}_{S_{k+1}}$ and let $O_{k+1} = \mathbf{X}_{S_{k+1}}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{S_{k+1}}$.

The procedure sequentially generates a series of covariate index sets: $S_0 \subset S_1 \subset \cdots \subset S_k \subset S_{k+1}$. To decide whether to end the procedure at the kth step or to recruit another variable j_{k+1} and proceed to the next step, we compute the following EBIC on set S_{k+1} with a tuning parameter η:

$$\begin{aligned}\mathrm{EBIC}(S_{k+1}) &= -2\ell_{S_{k+1}}(\widehat{\boldsymbol{\beta}}_{S_{k+1}}) + |S_{k+1}|(\log n + 2\eta \log p)/n \\ &= -2\ell_{S_{k+1}}(\widehat{\boldsymbol{\beta}}_{S_{k+1}}) + (k+1)(\log n + 2\eta \log p)/n. \end{aligned} \tag{3}$$

We terminate the algorithm if $\mathrm{EBIC}(S_{k+1}) > \mathrm{EBIC}(S_k)$ and declare $\widehat{\mathcal{M}} = S_k$, the final model; otherwise, the procedure will proceed to search a new variable. For more clarity, the following pseudocode captures the main thrust of the algorithm.

A sequential conditioning algorithm

1: **(Initialization)** Start with a set of a priori known S_0. Otherwise, initialize with $S_0 = \{0\}$.

Set $O_0 = \widehat{\beta}_{S_0}$, where $\widehat{\beta}_{S_0}$ maximizes $l_{S_0}(\beta_{S_0})$.

Compute $\widehat{\beta}_j^{(1)} = \arg\max_\beta \ell_{O_0,j}(\beta)$, where $\ell_{O,j}(\beta) = \mathbb{E}_n\{l(O + \beta X_j, Y)\}$.

Let $j_1 = \arg\max_{j\in\{1,\ldots,p\}} \ell_{O_0,j}(\widehat{\beta}_j^{(1)})$.

2: **(Repeat)** For $k \geqslant 1$, given $O_k = \mathbf{X}_{S_k}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_{S_k}$ and $S_k = S_k \cup \{j_k\}$,

compute $\widehat{\beta}_j^{(k+1)} = \arg\max_\beta \ell_{O_k,j}(\beta)$ for $j \in S_k^c$.

Set $j_{k+1} = \arg\max_{j\in S_k^c} \ell_{O_k,j}(\widehat{\beta}_j^{(k+1)})$.

3: **(Stop)** If $\mathrm{EBIC}(S_{k+1}) > \mathrm{EBIC}(S_k)$ and declare $\widehat{\mathcal{M}} = S_k$,

where $\mathrm{EBIC}(S_k) = -2\ell_{S_k}(\widehat{\boldsymbol{\beta}}_{S_k}) + k(\log n + 2\eta \log p)/n$.

The proposed SC approach simultaneously performs variable selection and model selection via EBIC, halting the procedure after including $k(< n)$ variables if the criterion of $\mathrm{EBIC}_k < \mathrm{EBIC}_{k+1}$ is met. In contrast, the typical screening approaches that do not internally incorporate the model selection procedure need to employ arbitrary cutoffs for termination, which may inflate the false positives or false negatives. We also treated the tuning parameter η as a fixed constant which may not vary by datasets. This is analogous to the constant "a" parameter in the SCAD penalty function (Fan and Li, 2001), and distinguishes our work from the screening approaches, which typically require data-driven tuning parameters and may incur much computational burden for finding them. To investigate the idea of fixing η, in Section 4, we numerically examined the results with different choices of η values.

3. Theoretical Properties

Let $\to_p$ and $\to_d$ denote convergence in probability and distribution, respectively. For a column vector $\boldsymbol{v}$, let $\boldsymbol{v}^{\otimes 2} = \boldsymbol{v}\boldsymbol{v}^{\mathrm{T}}$. For $q \geqslant 1$, denote its l_q-norm by $\|\boldsymbol{v}\|_q$, and, in particular, denote its

l_2-norm by $\|\boldsymbol{v}\|$. For any symmetric matrix $\mathbf{A}$, let $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ represent its smallest and largest eigenvalues. We impose the following regularity conditions.

(A) For a positive integer ρ satisfying $|\mathcal{M}| \leqslant \rho$ and $\rho \log p = o(n^{1/3})$, there exists a constant $L > 0$ such that $\sup_{|S| \leqslant \rho} \|\boldsymbol{\beta}_S^*\|_1 \leqslant L$, where $\boldsymbol{\beta}_S^* = \arg\max_{\boldsymbol{\beta}_S} E\{\ell_S(\boldsymbol{\beta}_S)\}$ denotes the least false value of model S.

(B) $\|\mathbf{X}\|_\infty \leqslant K$, where $K > 0$ is a constant.

(C) Let $\epsilon_i = Y_i - \mu(\boldsymbol{\beta}_*^{\mathrm{T}} \mathbf{X}_i)$. There exists a positive constant M such that the Cramer condition holds for all ϵ_i, i.e. $E[|\epsilon_i|^m] \leqslant m! M^m$ for all i and $m \geqslant 2$.

(D) There exist two positive constants $0 < \kappa_{\min} < \kappa_{\max} < \infty$, such that $\kappa_{\min} < \lambda_{\min}\left\{E\left(\mathbf{X}_S^{\otimes 2}\right)\right\}$ and $\lambda_{\max}\left\{E\left(\mathbf{X}_S^{\otimes 2}\right)\right\} < \kappa_{\max}$, uniformly in $S \subset \{0, 1, \ldots, p\}$ satisfying $|S| \leqslant \rho$.

(E) Let $D_S := \max_{j \in \mathcal{S}^c \cap \mathcal{M}} \left| E\left[\left\{Y - \mu(\boldsymbol{\beta}_S^{*\mathrm{T}} \mathbf{X}_S)\right\} X_j\right]\right|$. There exist some constants $C > 0$ and $\alpha > 0$ such that $\min_{S:|S| \leqslant \rho, \mathcal{M} \not\subseteq S} D_S > C n^{-\alpha}$ and $\rho n^{-1+4\alpha} \log p \to 0$.

Condition (A) differs from the Lipschitz assumption in van de Geer (2008), Fan and Song (2010), and Barut et al. (2016). A similar condition is assumed in Bühlmann (2006). The condition $\rho \log p = o(n^{1/3})$ is needed to ensure the consistency of EBIC, as required in Chen and Chen (2012). The parameter ρ is an upper bound of the model size, which is often required in joint-model-based selection or screening methods with various notation, such as "M" in Cheng et al. (2016), and "K" in Zhang and Huang (2008), Chen and Chen (2008), and Fan and Tang (2013). This condition is weaker than Assumption D in Cheng et al. (2016), which requires $\rho \log p = O(n^{1/5} \log n)$. Condition (B) has been commonly assumed in the literature for variable selection and screening (Zhao and Li, 2012; Li et al., 2016; Kwemou, 2016). The uniform boundedness of $\mathbf{X}$ is adopted to simplify our theoretical development and can be relaxed to Conditions B and D in Fan and Song (2010). In practice, data are often standardized at the pre-processing stage, which may warrant the reasonableness of this condition. Condition (C) is justified by Jiang and Zhang (2013) and Jiang et al. (2016) and

is similar to Condition 3 in Bradic et al. (2011). The condition ensures the light tail of the response variable Y and is satisfied by a wide range of outcome data, including Gaussian and discrete data (such as binary and count data). Condition (D) has been commonly assumed in literature (Wang, 2009; Zheng et al., 2015; Cheng et al., 2016) and represents the Sparse Riesz Condition (Zhang and Huang, 2008). Compared to those required by joint-model-based sequential screening methods in the literature, the signal condition (E) is not directly imposed on the regression coefficient. Instead, it is imposed on the conditional covariance between a covariate and the response, as in Barut et al. (2016). The condition can also be reviewed as an "strong irrepresentable" condition (Zhao and Yu, 2006) for model identifiability, stipulating that the true model $\mathcal{M}$ cannot be represented by a different set of variables that do not include the true model. It implies that the Kullback-Leibler divergence from a mis-specified model to the true model is large enough for mis-specified models to be detected; see Leroux (1992). The signal rate is comparable to those conditions required by other sequential methods in the literature, such as the rate $n^{-1/12}$ in Wang (2009) and the rate $n^{-1/5}$ in Cheng et al. (2016). Conditions (A) and (E) together indicate that the range of ρ depends on the true model size $|\mathcal{M}|$, the minimum signal strength, $n^{-\alpha}$, and the total number of covariates, p. The lower bound of ρ is $|\mathcal{M}|$, and the upper bound of ρ is $o((n^{1-4\alpha}/\log p) \wedge (n^{1/3}/\log p))$. For example, if $\alpha = 0$ and $|\mathcal{M}|$ is finite, ρ can be chosen as $O(n^{1/4}/\log p)$. If $\alpha = 1/6$ and $|\mathcal{M}| = o(n^{1/4}/\log p)$, ρ can be chosen as $O(n^{1/4+\delta}/\log p)$, for any $0 < \delta < 1/12$.

For any model S with cardinality $|S| \leqslant \rho$, Condition (A) implies that the parameter space under consideration can be restricted to $\mathbb{B} := \{\boldsymbol{\beta} \in \mathbb{R}^{p+1} : \|\boldsymbol{\beta}\|_1 \leqslant \tau L\}$ for some large constant τ. As $b(\cdot)$ is twice continuously differentiable, with a nonnegative second derivative $b''(\cdot)$, $b_{\max} := \max_{|t| \leqslant \tau KL} |b(t)|$, $\mu_{\max} := \max_{|t| \leqslant \tau KL} |b'(t)|$, and $\sigma_{\max} := \sup_{|t| \leqslant \tau KL} |b''(t)|$ are assumed to be bounded above, where L and K are defined in Conditions (A) and (B), respectively. In addition, $\sigma_{\min} := \inf_{|t| \leqslant \tau KL} |b''(t)|$ is bounded below.

Given any $\boldsymbol{\beta}_S$, when a variable $X_r, r \in S^c$ is added into the model S, we define the augmented log-likelihood as

$$\ell_{S\cup\{r\}}(\beta_r|\boldsymbol{\beta}_S) := \mathbb{E}_n\left\{l\left(\boldsymbol{\beta}_S^{\mathrm{T}}\mathbf{X}_S + \beta_r X_r, Y\right)\right\}. \tag{1}$$

In other words, $\ell_{S\cup\{r\}}(\beta_r|\boldsymbol{\beta}_S) = \ell_{S\cup\{r\}}((\boldsymbol{\beta}_S^{\mathrm{T}}, \beta_r)^{\mathrm{T}}) = \ell_{O,r}(\beta)$ with $O = \mathbf{X}_S^{\mathrm{T}}\boldsymbol{\beta}_S$, where $\ell_{O,r}(\beta)$ is defined as in Section 2.2. We use $\widehat{\beta}_{r|S}(\boldsymbol{\beta}_S)$ to denote the maximizer of $\ell_{S\cup\{r\}}(\beta_r|\boldsymbol{\beta}_S)$, which solves $\mathbb{E}_n\left[\{Y_i - \mu(\boldsymbol{\beta}_S^{\mathrm{T}}\mathbf{X}_{iS} + \beta_r X_{ir})\}X_{ir}\right] = 0$. In addition, denote the maximizer of $E\left\{\ell_{S\cup\{r\}}(\beta_r|\boldsymbol{\beta}_S^*)\right\}$ by $\beta_{r|S}^*$. Due to the concavity of the log-likelihood in GLMs with the canonical link, $\beta_{r\overline{S}}^*$ is unique and is an interior point over $[-L, L]$ (Fan and Song, 2010).

The following theorem establishes the lower bound of the increment of the log-likelihood provided by SC when the true model $\mathcal{M}$ is included in the selected model. Thus, it provides the feasibility foundation of the proposed SC.

THEOREM 3.1: *Under Conditions* (A)–(E)*, there exists some constant* C_1*, which does not depend on* n*, such that with probability at least* $1 - 24\exp(-6\rho\log p)$*,*

$$\max_{j\in S^c}\left[\ell_{S\cup\{j\}}\{\hat{\beta}_{j|S}(\hat{\boldsymbol{\beta}}_S)|\hat{\boldsymbol{\beta}}_S\} - \ell_S(\hat{\boldsymbol{\beta}}_S)\right] \geqslant C_1 n^{-2\alpha},$$

uniformly in S *satisfying* $|S| < \rho$ *and* $\mathcal{M} \not\subseteq S$*.*

Given a model S such that $\mathcal{M} \not\subseteq S$ and $|S| < \rho$, let r be the index of the variable selected by SC. As $(\widehat{\boldsymbol{\beta}}_S^{\mathrm{T}}, \widehat{\beta}_{r|S}(\widehat{\boldsymbol{\beta}}_S))^{\mathrm{T}}$ is suboptimal to $\widehat{\boldsymbol{\beta}}_{S\cup\{r\}}$ in terms of maximizing $\ell_{S\cup\{r\}}(\boldsymbol{\beta}_{S\cup\{r\}})$, we obtain $\ell_{S\cup\{r\}}(\widehat{\boldsymbol{\beta}}_{S\cup\{r\}}) \geqslant \ell_{S\cup\{r\}}\{(\widehat{\boldsymbol{\beta}}_S^{\mathrm{T}}, \widehat{\beta}_{r|S}(\widehat{\boldsymbol{\beta}}_S))^{\mathrm{T}}\} = \ell_{S\cup\{r\}}\{\widehat{\beta}_{r|S}(\widehat{\boldsymbol{\beta}}_S)|\widehat{\boldsymbol{\beta}}_S\}$. Thus, Theorem 3.1 implies that that the increment of the log-likelihood provided by SC is at least $C_1 n^{-2\alpha}$ with probability tending to 1, if $\mathcal{M} \not\subseteq S$.

In fact, the lower bound of increment from Theorem 3.1 also guarantees that the proposed SC will stop in steps of polynomial size and thus provides the validity of SC. Since the maximum increment is bounded by $(\sqrt{2}M + 2\mu_{\max})\tau KL + b_{\max}$ with probability tending to

1 (Lemma 3), we naturally obtain an upper bound on the number of steps for SC, which is stated in the next corollary.

COROLLARY 3.1: *Under Conditions* (A)–(E), *if* $N := 2C_1^{-1}\{(\sqrt{2}M + 2\mu_{\max})\tau KL + b_{\max}\}n^{2\alpha} < \rho$ *and* $0 \leqslant \alpha < 1/6$, *then* $\mathcal{M} \subset S_k$, *for some* S_k *selected by SC with* $k \leqslant N$, *with probability at least* $1 - 26\exp(-6\rho \log p)$.

Corollary 3.1 establishes the screening consistency of SC. It follows a similar idea in Fan and Song (2010) and Cheng et al. (2016). The condition $N < \rho$ is sufficient but not necessary, as the upper bound N is obtained based on the lower bound on the increment of the log-likelihood and is not tight. With certain additional conditions, the bound can be improved significantly. The following theorem establishes an upper bound of the number of steps by assessing how likely a signal variable will be selected at each step.

THEOREM 3.2: *Under Conditions (A) – (E), if* $\max_{j \in \mathcal{S}^c \cap \mathcal{M}^c} \left| E\left[\left\{Y - \mu(\boldsymbol{\beta}_S^{*\mathrm{T}} \mathbf{X}_S)\right\} X_j\right]\right| = o(n^{-\alpha})$ *uniformly over* S *with* $|S| \leqslant \rho$, *then there exists some constant* $C_2 > 2$ *such that* $\mathcal{M} \subset S_k$, *for some* S_k *selected by SC and* $k \leqslant C_2|\mathcal{M}|$, *with probability at least* $1 - 36\exp(-3\rho \log p)$.

The "max" condition in the theorem is similar to a condition in Section 5.3 of Fan and Song (2010). It is a generalization of the partial orthogonality assumption that $\mathbf{X}_{\mathcal{M}^c}$ are independent of $\mathbf{X}_{\mathcal{M}}$. This condition ensures that a signal variable brings more increment in log-likelihood than a noise variable, with probability tending to 1 uniformly over all model $S : |S| < \rho, \mathcal{M} \not\subseteq S$. Therefore, the proposed procedures have a large probability to select a signal variable at each step.

Since EBIC is a consistent model selection criterion (Luo and Chen, 2014; Luo et al., 2015), we expect the proposed SC to stop early with $\mathcal{M} \subset S_k$ for some finite k as shown in the following theorem. Thus, the final model $\widehat{\mathcal{M}}$ provided by SC may not include too many noise variables.

THEOREM 3.3: *Suppose the conditions in Corollary 3.1 or Theorem 3.2 hold. If* $\mathcal{M} \not\subset S_{k-1}$ *and* $\mathcal{M} \subset S_k$*, then the procedure stops at the* k*th step with probability going to 1.*

4. Numerical Studies

We conducted simulation studies to compare the proposed sequential conditioning (SC) approach with some competing methods, including sure independence screening (SIS) of Fan and Lv (2008), and conditional sure independence screening (CSIS) of Barut et al. (2016).

The competing screening approaches typically rely on some arbitrary cutoffs when determining the number of selected variables, which may inflate the false positives. Therefore, to make fair comparisons, we first applied these methods to select the top $[n/\log n]$ variables as suggested by Fan and Lv (2008) and then applied Lasso (Tibshirani, 1996), SCAD (Fan and Li, 2001), and MCP (Zhang, 2010) penalties to arrive at the final models. In the tables, we used method+penalty to denote the corresponding procedure.

Example 1: We set $\boldsymbol{\beta} = c \times (1, -1, 1, -1, 1, -\iota + \iota^2 - \iota^3 + \iota^4 - \iota^5, \mathbf{0}_{p-6})^{\mathrm{T}}$, where $\iota = 0.5$. $\mathbf{X}$ were generated under a multivariate normal distribution with mean 0, variance 1, and $\mathrm{cor}(X_j, X_{j'}) = 0.5^{|j-j'|}$, for $1 \leqslant j \neq j' \leqslant p$. Here c (and hereafter) is a positive constant, which will be chosen to maintain a pre-specified signal-to-noise ratio.

Example 2: We set $\boldsymbol{\beta} = c \times (1, 1, 1, 1, 1, -2.5, \mathbf{0}_{p-6})^{\mathrm{T}}$. $\mathbf{X}$ were generated under a multivariate normal distribution with mean 0, variance 1, and $\mathrm{cor}(X_j, X_{j'}) = 0.5$, for $1 \leqslant j \neq j' \leqslant p$.

Example 3: We set $\boldsymbol{\beta} = c \times (\mathbf{1}_{15}, \mathbf{0}_{p-15})^{\mathrm{T}}$ and generated $\mathbf{X}$ from the independent standard normal distribution.

Example 4: We set $\boldsymbol{\beta} = c \times (\mathbf{1}_{15}, \mathbf{0}_{p-15})^{\mathrm{T}}$. A total of 15 active variables were generated by a zero-mean multivariate normal distribution, where the covariance matrix had a block-diagonal structure with 3 equal-sized blocks. The inverses of this covariance matrix corresponds to 3 independent star-shaped graphs. Within each graph, 4 nodes are connected to

a hub node with no other connections. Specifically, the covariance matrix S for each block can be formulated as: $\Sigma_{ii} = 1$, $\Sigma_{ij} = 0.3$ if (i, j) is an edge and $\Sigma_{ij} = 0.3^2$ otherwise. The other $p - 15$ variables were independently generated from the standard normal distribution.

For each example, we considered linear regression, logistic regression, and Poisson regression models. For linear regression model, we generated $Y = \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} + \epsilon$, where the random error ϵ follows $N(0, \sigma^2)$; for binary outcomes, Y were generated as independent Bernoulli variables with probability of success $\exp(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta})/\{1 + \exp(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta})\}$, and for Poisson regression model, we generated Y as independent Poisson variables with mean $\exp(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta})$.

We set the magnitude of the coefficients in various GLMs according to a pre-specified signal-to-noise ratio (SNR), the ratio of the variance of a signal to the variance of the noise (Czanner et al., 2008). Specifically, we set c to produce an SNR of 2.

We considered p=1000 and varied sample sizes n=200 and 400. For each parameter configuration, we simulated 200 independent datasets. We evaluated the performance of the methods by the criteria of true positives (TP) and false positives (FP). Tables 1 – 3, which report the results for the logistic and Poisson regression models respectively, present several interesting observations.

First, Examples 1 and 2 were designed in such a way that X_6, though active, has a 0 marginal correlation with the outcome and, therefore, is not detectable by marginal screening methods, such as unconditional SIS. Indeed, SIS was found to produce fewer true positives, whereas by conditioning on an active variable X_1, CSIS increased true positives but at the price of increasing false positives. In contrast, because SC selects variables sequentially and is able to detect such a "hidden" variable, the proposed SC recruited almost all the active variables with the average TP close to the true model size.

Second, even with Lasso, SCAD and MCP to further reduce false positives, the competing

screening methods still resulted in many false positives. In contrast, the proposed SC with the EBIC-based stopping rule had fewer false positives.

Third, although the covariates generated from multivariate normal distributions which were unbounded, our proposed methods worked well, hinting at robustness of the methods toward the boundness assumption on covariates.

Fourth, as shown in the results of Examples 1–2, the performance of CSIS tends to depend on the conditioning set. Even compared to the case in which CSIS used the known prior information, SC works competently well without any prior information.

Fifth, when the number of active variables was relatively large as in Examples 3–4, the performance of the proposed method deteriorated especially for a smaller n. This might be due to poor fitting of models with larger model sizes and smaller sample sizes. However, as the sample size increased, the performance improved and was fairly robust toward the choice of η values.

Lastly, we observed that EBIC with larger values of η tended to select fewer variables, especially for the binary and count data. We also noted that the choice of $\eta = 1 - \{\log n/(3 \log p)\}$ well balanced the true positives and false positives among all the scenarios examined.

[Table 1 about here.]

[Table 2 about here.]

[Table 3 about here.]

5. Analysis of a Multiple Myeloma Trial

We demonstrate the utility of the proposed method by predicting the responses to treatment among a group of multiple myeloma patients. Multiple myeloma is an incurable malignancy that originates in the antibody-secreting bone marrow plasma cells, and genomics has important prognostic values for this disease. The practicality and utility of using genomic research

to predict the outcome for a specific therapy remain unclear. We apply the proposed approach to identify genes that are relevant to clinical response in a trial conducted by Mulligan et al. (2007). In the study, patients were classified as achieving complete response (CR), partial response (PR), minimal response (MR), no change (NC), or progressive disease (PD), using the European Group for Bone Marrow Transplantation criteria. In brief, CR, PR, and MR require at least 100%, 50%, and 25% decreases in paraprotein respectively, whereas PD requires at least 25% increases. We applied the proposed methods to the binary response [CR vs. PR/MR/NC/PD]. A total of 76 patients achieved CR among a total of 264 patients. Amin et al. (2014) gave clinical justifications on and importance of identifying relevant genomic profiles to predict complete response, among many other possible choices of endpoints. Our analysis was to identify a set of important genes that could predict CR among a total of 44,928 gene probes.

Equipped with the EBIC stopping rule, our method sequentially selected probe sets, which formed the final model for predicting treatment response. To evaluate the impact of the different choices of η on variable selection, we applied sequential conditioning (SC) approach with various values of $\eta_1 = 0.5, \eta_2 = 1,\ \eta_3 = 1 - \{\log n/(3\log p)\}$. The results were fairly robust, though as expected SC with smaller η tends to impose less penalty and include more variables. Specifically, genes, such as TNFRSF11A (TNF receptor superfamily member 11a), FAM127A (Family with sequence similarity 127, member A), and STRBP (spermatid perinuclear RNA binding protein), were selected for all η values, while additionally Crabp1 (cellular retinoic acid binding protein I) and PPFIBP1 (PPFIA binding protein 1) were selected only for $\eta = 0.5$. Our findings have biological interpretations. For example, Gene TNFRSF11A was one of the oncogenes selected by the recursive feature addition and gradient based Leave-one-out gene selection based on the MAQC-II breast cancer data (Liu et al., 2009). Gene FAM127A was identified as one of the genes significantly different among HCV

cirrhotic tissue compared to normal liver tissue (Fassnacht, 2010). Gene STRBP was found to be a transcriptional signature for mutations in chromatin modifying genes (Green et al., 2015).

For comparisons, we also applied the different methods introduced in Section 4 to screen out irrelevant genes and reached the final predictive models through various penalties. Table 4 presents the selected genes by different methods. For brevity, among a total of 31 unique genes were chosen from various methods, we elected to display the genes that were selected by more than five methods. Our SC algorithm did identify genes that were not identified by the multiple myeloma literature, though external validation is needed to further confirm the findings.

Finally, we reported the leave-one-out prediction error for different η in Table 5 for each subject, which is the average of the squared difference between the response and the predicted probability. For a fair comparison, we assumed the prior of CSIS is unknown and used 10 different randomly selected conditioning sets. Table 5 indicates that, although our proposed method yielded the smallest model size, the associated prediction error was very comparable to that with a larger model size. The average prediction error of CSIS from 10 such random prior was 0.19 with a range between 0.17 to 0.20, indicating the choice of conditioning set did influence the performance of the conditional screening.

[Table 4 about here.]

[Table 5 about here.]

6. Concluding Remarks

Marginal screening approaches, though widely used, have often been challenged for restrictive faithfulness assumptions and lack of clear rules for the final model selection. This article fills the gap by investigating a sequential conditioning approach, which utilizes an offset

term to aggregate information obtained from the previous steps. The approach is promising with computationally and theoretically useful results. We have demonstrated that, if the dimension of the true model is finite, our approach can discover the true model within a finite number of steps. As our method is likelihood based, we envision the theoretical framework will facilitate a wide range of outcome data.

There are several directions for future research. We employed an EBIC (with an added penalty term, quantified by η, to the usual BIC) to select the final models. Although it worked well under our simulations, it tends to be conservative in real data analysis and recruits too few variables. It would be interesting to investigate the optimal η in the EBIC penalty term to strike a balance between false positives and negatives.

In addition, drawing inferences on top of a variable selection procedure remains challenging, though our asymptotic results could be a very first step. There are some other approaches, such as debiased Lasso estimators (van de Geer, 2008), for drawing inferences for high-dimensional linear regression models and GLMs. Extensions of these approaches to accommodate the proposed sequential conditioning approach are of substantial interest and, perhaps, require the development of new theory and algorithms.

References

Amin, S. B., Yip, W.-K., Minvielle, S., Broyl, A., Li, Y., Hanlon, B., Swanson, D., Shah, P. K., Moreau, P., van der Holt, B., and van Duin M (2014). Gene expression profile alone is inadequate in predicting complete response in multiple myeloma. *Leukemia*, 28(11):2229–2234.

Barut, E., Fan, J., and Verhasselt, A. (2016). Conditional sure independence screening. *Journal of the American Statistical Association*, 111(515):1266–1277.

Bradic, J., Fan, J., and Jiang, J. (2011). Regularization for Cox's proportional hazards model with NP-dimensionality. *The Annals of Statistics*, 39(6):3092–3120.

Bühlmann, P. (2006). Boosting for high-dimensional linear models. *The Annals of Statistics*, 34(2):559–583.

Chen, J. and Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771.

Chen, J. and Chen, Z. (2012). Extended BIC for small-n-large-p sparse GLM. *Statistica Sinica*, 22:555–574.

Cheng, M.-Y., Honda, T., and Zhang, J.-T. (2016). Forward variable selection for sparse ultra-high dimensional varying coefficient models. *Journal of the American Statistical Association*, 111(515):1209–1221.

Czanner, G., Sarma, S. V., Eden, U. T., and Brown, E. N. (2008). A signal-to-noise ratio estimator for generalized linear model systems. In *Proceedings of the World Congress on Engineering, July 2 - 4, London, U.K.*, volume II.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911.

Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *The Annals of Statistics*, 38(6):3567–3604.

Fan, Y. and Tang, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(3):531–552.

Fassnacht, R. (2010). *Molecular mechanisms involved in the interaction effects of HCV and ethanol on liver cirrhosis*. PhD dissertation, Virginia Commonwealth University.

Green, M. R., Kihira, S., Liu, C. L., Nair, R. V., Salari, R., Gentles, A. J., Irish, J.,

Stehr, H., Vicente-Dueñas, C., Romero-Camarero, I., and Sanchez-Garcia, I. (2015). Mutations in early follicular lymphoma progenitors are associated with suppressed antigen presentation. *Proceedings of the National Academy of Sciences*, 112(10):E1116–E1125.

Hong, H. G., Wang, L., and He, X. (2016). A data-driven approach to conditional screening of high-dimensional variables. *Stat*, 5(1):200–212.

Jiang, Y., He, Y., and Zhang, H. (2016). Variable selection with prior information for generalized linear models via the prior Lasso method. *Journal of the American Statistical Association*, 111(513):355–376.

Jiang, Y. and Zhang, C. (2013). High-dimensional regression and classification under a class of convex loss functions. *Statistics and Its Interface*, 6(2):285–299.

Kwemou, M. (2016). Non-asymptotic oracle inequalities for the Lasso and group Lasso in high dimensional logistic model. *ESAIM: Probability and Statistics*, 20:309–331.

Leroux, B. G. (1992). Consistent estimation of a mixing distribution. *The Annals of Statistics*, 20(3):1350–1360.

Li, J., Zheng, Q., Peng, L., and Huang, Z. (2016). Survival impact index and ultrahigh-dimensional model-free screening with survival outcomes. *Biometrics*, 72(4):1145–1154.

Li, Y., Hong, H. G., Ahmed, S. E., and Li, Y. (2019). Weak signals in high-dimensional regression: Detection, estimation and prediction. *Applied Stochastic Models in Business and Industry*, 35(2):283–298.

Liu, Q., Sung, A. H., Chen, Z., Liu, J., Huang, X., and Deng, Y. (2009). Feature selection and classification of MAQC-II breast cancer and multiple myeloma microarray gene expression data. *PloS one*, 4(12):e8250.

Luo, S. and Chen, Z. (2014). Sequential Lasso cum EBIC for feature selection with ultra-high dimensional feature space. *Journal of the American Statistical Association*,

109(507):1229–1240.

Luo, S., Xu, J., and Chen, Z. (2015). Extended bayesian information criterion in the Cox model with a high-dimensional feature space. *Annals of the Institute of Statistical Mathematics*, 67(2):287–311.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models, Second Edition*. Chapman & Hall: London-New York.

Mulligan, G., Mitsiades, C., Bryant, B., Zhan, F., Chng, W. J., Roels, S., Koenig, E., Fergus, A., Huang, Y., Richardson, P., and Trepicchio, W. (2007). Gene expression profiling and correlation with outcome in clinical trials of the proteasome inhibitor bortezomib. *Blood*, 109(8):3177–3188.

Paul, D., Bair, E., Hastie, T., and Tibshirani, R. (2008). "Preconditioning" for feature selection and regression in high-dimensional problems. *The Annals of Statistics*, 36:1595–1618.

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

van de Geer, S. A. (2008). High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36(2):614–645.

Wang, H. (2009). Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association*, 104(488):1512–1524.

Xu, C. and Chen, J. (2014). The sparse mle for ultrahigh-dimensional feature screening. *Journal of the American Statistical Association*, 109(507):1257–1269.

Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942.

Zhang, C.-H. and Huang, J. (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics*, 36(4):1567–1594.

Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine learning research*, 7:2541–2563.

Zhao, S. D. and Li, Y. (2012). Principled sure independence screening for Cox models with ultra-high-dimensional covariates. *Journal of Multivariate Analysis*, 105(1):397–411.

Zheng, Q., Peng, L., and He, X. (2015). Globally adaptive quantile regression with ultra-high dimensional data. *The Annals of Statistics*, 43(5):2225–2258.

Supporting Information

All the technical details, including lemmas, conditions and proofs, in Section 3 are available with this paper at the Biometrics website on Wiley Online Library. The code, example data, and README file are also available.

Table 1
Comparisons of competing methods with linear regression models

		$(n, p) = (200, 1000)$		$(n, p) = (400, 1000)$	
Example	Method	TP	FP	TP	FP
1A	SC_{η_1}	5.61 (0.49)	1.25 (0.64)	5.96 (0.19)	1.10 (0.37)
	SC_{η_2}	5.55 (0.50)	0.89 (0.52)	5.95 (0.21)	0.95 (0.33)
	SC_{η_3}	5.53 (0.50)	0.67 (0.55)	5.93 (0.26)	0.60 (0.49)
	SIS	3.46 (0.64)	33.54 (0.64)	4.15 (0.77)	61.85 (0.77)
	SIS+Lasso	3.10 (1.02)	13.49 (13.85)	4.08 (0.86)	6.48 (9.08)
	SIS+MCP	3.13 (0.95)	11.00 (9.27)	4.00 (1.00)	0.66 (2.26)
	SIS+SCAD	3.20 (0.92)	12.27 (11.43)	4.06 (0.89)	0.77 (2.52)
	$\text{CSIS}_{\mathcal{A}}$	4.42 (0.58)	32.58 (0.58)	5.00 (0.36)	61.00 (0.36)
	$\text{CSIS}_{\mathcal{A}}$+Lasso	4.38 (0.67)	27.76 (6.90)	5.00 (0.35)	14.80 (14.60)
	$\text{CSIS}_{\mathcal{A}}$+MCP	4.41 (0.58)	7.83 (8.39)	5.00 (0.36)	0.63 (1.77)
	$\text{CSIS}_{\mathcal{A}}$+SCAD	4.41 (0.58)	10.38 (10.84)	5.00 (0.35)	0.25 (0.76)
	$\text{CSIS}_{\mathcal{I}}$	3.47 (0.62)	33.53 (0.62)	4.16 (0.76)	61.84 (0.76)
	$\text{CSIS}_{\mathcal{I}}$+Lasso	3.12 (0.99)	13.52 (14.09)	4.08 (0.86)	6.68 (9.35)
	$\text{CSIS}_{\mathcal{I}}$+MCP	3.15 (0.94)	11.08 (9.40)	4.01 (0.99)	0.68 (2.53)
	$\text{CSIS}_{\mathcal{I}}$+SCAD	3.21 (0.91)	12.54 (11.54)	4.07 (0.88)	0.75 (2.40)
2A	SC_{η_1}	6.00 (0.00)	1.14 (0.41)	6.00 (0.00)	1.12 (0.37)
	SC_{η_2}	6.00 (0.00)	1.01 (0.08)	6.00 (0.00)	1.01 (0.13)
	SC_{η_3}	6.00 (0.00)	1.00 (0.06)	6.00 (0.00)	1.00 (0.00)
	SIS	4.87 (0.38)	32.13 (0.38)	5.00 (0.00)	61.00 (0.00)
	SIS+Lasso	4.78 (0.46)	3.63 (2.80)	5.00 (0.00)	0.76 (1.10)
	SIS+MCP	4.07 (0.77)	0.25 (0.57)	4.79 (0.43)	0.46 (1.60)
	SIS+SCAD	4.49 (0.65)	0.96 (1.42)	4.89 (0.34)	0.10 (0.49)
	$\text{CSIS}_{\mathcal{A}}$	5.87 (0.36)	31.13 (0.36)	6.00 (0.04)	60.00 (0.04)
	$\text{CSIS}_{\mathcal{A}}$+Lasso	5.86 (0.42)	10.89 (3.46)	6.00 (0.04)	12.63 (3.82)
	$\text{CSIS}_{\mathcal{A}}$+MCP	5.81 (0.57)	0.97 (1.33)	6.00 (0.04)	0.70 (1.08)
	$\text{CSIS}_{\mathcal{A}}$+SCAD	5.83 (0.51)	0.42 (1.22)	6.00 (0.04)	0.14 (0.59)
	$\text{CSIS}_{\mathcal{I}}$	4.99 (0.51)	32.01 (0.51)	5.17 (0.38)	60.83 (0.38)
	$\text{CSIS}_{\mathcal{I}}$ +Lasso	4.91 (0.58)	4.85 (3.93)	5.17 (0.38)	2.87 (4.60)
	$\text{CSIS}_{\mathcal{I}}$+MCP	4.35 (0.94)	0.65 (0.96)	5.03 (0.57)	0.96 (3.09)
	$\text{CSIS}_{\mathcal{I}}$+SCAD	4.67 (0.76)	1.26 (1.62)	5.09 (0.52)	0.46 (0.78)
3A	SC_{η_1}	14.97 (0.21)	0.91 (1.01)	15.00 (0.00)	0.68 (0.66)
	SC_{η_2}	14.90 (0.48)	0.36 (0.71)	15.00 (0.00)	0.20 (0.42)
	SC_{η_3}	14.79 (0.65)	0.25 (0.61)	15.00 (0.00)	0.02 (0.13)
	SIS	12.73 (1.16)	24.27 (1.16)	14.96 (0.20)	51.04 (0.20)
	SIS+Lasso	12.73 (1.16)	20.77 (2.54)	14.96 (0.20)	21.00 (9.91)
	SIS+MCP	12.73 (1.17)	6.72 (4.30)	14.96 (0.20)	0.79 (1.64)
	SIS+SCAD	12.73 (1.16)	10.78 (4.80)	14.96 (0.20)	1.91 (1.86)
	$\text{CSIS}_{\mathcal{A}}$	13.14 (1.07)	23.86 (1.07)	14.98 (0.13)	51.02 (0.13)
	$\text{CSIS}_{\mathcal{A}}$+Lasso	13.14 (1.07)	20.44 (2.42)	14.98 (0.13)	21.75 (10.39)
	$\text{CSIS}_{\mathcal{A}}$+MCP	13.14 (1.07)	6.38 (4.39)	14.98 (0.13)	0.83 (1.72)
	$\text{CSIS}_{\mathcal{A}}$+SCAD	13.14 (1.07)	10.52 (4.94)	14.98 (0.13)	1.82 (1.73)
	$\text{CSIS}_{\mathcal{I}}$	12.65 (1.21)	24.35 (1.21)	14.96 (0.20)	51.04 (0.20)
	$\text{CSIS}_{\mathcal{I}}$+Lasso	12.65 (1.21)	20.40 (2.72)	14.96 (0.20)	20.63 (9.66)
	$\text{CSIS}_{\mathcal{I}}$+MCP	12.65 (1.22)	6.63 (4.37)	14.96 (0.20)	0.76 (1.58)
	$\text{CSIS}_{\mathcal{I}}$+SCAD	12.65 (1.21)	10.65 (4.84)	14.96 (0.20)	1.85 (1.85)
4A	SC_{η_1}	12.87 (1.93)	2.16 (1.75)	14.99 (0.09)	0.93 (0.43)
	SC_{η_2}	11.94 (2.06)	1.25 (1.31)	14.99 (0.13)	0.47 (0.50)
	SC_{η_3}	10.83 (2.22)	0.71 (0.94)	14.97 (0.19)	0.10 (0.30)
	SIS	14.49 (0.69)	22.51 (0.69)	15.00 (0.04)	51.00 (0.04)
	SIS+Lasso	14.49 (0.69)	16.11 (3.66)	15.00 (0.04)	9.24 (6.37)
	SIS+MCP	14.36 (0.80)	5.26 (3.27)	14.99 (0.08)	1.99 (2.37)
	SIS+SCAD	14.43 (0.75)	8.56 (3.78)	15.00 (0.06)	4.48 (2.81)
	$\text{CSIS}_{\mathcal{A}}$	13.54 (0.99)	23.46 (0.99)	14.87 (0.34)	51.13 (0.34)
	$\text{CSIS}_{\mathcal{A}}$+Lasso	13.53 (1.01)	17.23 (4.34)	14.87 (0.34)	9.61 (7.24)
	$\text{CSIS}_{\mathcal{A}}$+MCP	13.36 (1.15)	6.96 (3.91)	14.87 (0.36)	2.32 (2.59)
	$\text{CSIS}_{\mathcal{A}}$+SCAD	13.49 (1.04)	11.17 (4.41)	14.87 (0.35)	5.21 (3.23)
	$\text{CSIS}_{\mathcal{I}}$	14.45 (0.70)	22.55 (0.70)	15.00 (0.04)	51.00 (0.04)
	$\text{CSIS}_{\mathcal{I}}$+Lasso	14.45 (0.70)	15.70 (3.74)	15.00 (0.04)	9.63 (6.48)
	$\text{CSIS}_{\mathcal{I}}$+MCP	14.32 (0.82)	5.05 (3.16)	14.99 (0.09)	1.96 (2.37)
	$\text{CSIS}_{\mathcal{I}}$+SCAD	14.40 (0.75)	8.36 (3.76)	15.00 (0.06)	4.56 (2.93)

Table 2
Comparisons of competing methods with logistic regression models

		$(n, p) = (200, 1000)$		$(n, p) = (400, 1000)$	
Example	Method	TP	FP	TP	FP
1B	SC_{η_1}	5.52 (0.62)	5.81 (2.44)	5.96 (0.19)	2.58 (3.53)
	SC_{η_2}	5.47 (0.71)	3.31 (2.63)	5.96 (0.21)	1.15 (1.09)
	SC_{η_3}	5.42 (0.74)	1.56 (1.81)	5.94 (0.24)	0.76 (0.44)
	SIS	3.35 (0.63)	33.65 (0.63)	4.18 (0.82)	61.82 (0.82)
	SIS+Lasso	2.85 (1.12)	14.87 (14.85)	4.07 (0.91)	12.62 (20.18)
	SIS+MCP	2.96 (0.97)	16.51 (7.31)	4.01 (1.05)	4.50 (8.98)
	SIS+SCAD	3.03 (0.95)	19.00 (8.38)	4.09 (0.90)	5.07 (9.05)
	$CSIS_{\mathcal{A}}$	4.27 (0.60)	32.73 (0.60)	4.99 (0.39)	61.01 (0.39)
	$CSIS_{\mathcal{A}}$+Lasso	4.17 (0.82)	28.05 (7.91)	4.99 (0.38)	35.37 (24.07)
	$CSIS_{\mathcal{A}}$+MCP	4.23 (0.66)	14.95 (5.73)	4.99 (0.38)	10.04 (11.79)
	$CSIS_{\mathcal{A}}$+SCAD	4.24 (0.63)	16.98 (6.23)	4.98 (0.39)	12.15 (12.78)
	$CSIS_{\mathcal{I}}$	3.35 (0.62)	33.65 (0.62)	4.16 (0.81)	61.84 (0.81)
	$CSIS_{\mathcal{I}}$+Lasso	2.80 (1.13)	14.24 (15.26)	4.05 (0.90)	11.43 (19.28)
	$CSIS_{\mathcal{I}}$+MCP	2.97 (0.94)	16.94 (7.45)	4.00 (1.03)	4.47 (9.17)
	$CSIS_{\mathcal{I}}$+SCAD	3.03 (0.94)	19.21 (8.76)	4.08 (0.90)	5.13 (9.44)
2B	SC_{η_1}	4.92 (1.12)	7.45 (4.12)	5.99 (0.10)	2.17 (3.43)
	SC_{η_2}	4.54 (1.24)	2.90 (3.32)	5.99 (0.12)	0.88 (0.62)
	SC_{η_3}	4.14 (1.19)	1.17 (1.62)	5.98 (0.13)	0.42 (0.51)
	SIS	4.20 (0.80)	32.80 (0.80)	4.96 (0.19)	61.04 (0.19)
	SIS+Lasso	3.12 (1.27)	5.22 (3.93)	4.79 (0.46)	3.83 (2.91)
	SIS+MCP	2.82 (1.02)	2.88 (2.24)	4.40 (0.69)	0.69 (1.11)
	SIS+SCAD	3.36 (1.12)	6.65 (3.61)	4.79 (0.44)	3.15 (2.87)
	$CSIS_{\mathcal{A}}$	5.22 (0.80)	31.78 (0.80)	5.92 (0.27)	60.08 (0.27)
	$CSIS_{\mathcal{A}}$+Lasso	4.89 (1.28)	13.81 (5.89)	5.90 (0.39)	17.29 (5.19)
	$CSIS_{\mathcal{A}}$+MCP	4.63 (1.34)	4.39 (2.96)	5.87 (0.47)	1.44 (1.77)
	$CSIS_{\mathcal{A}}$+SCAD	4.80 (1.22)	8.19 (3.93)	5.90 (0.38)	4.35 (3.73)
	$CSIS_{\mathcal{I}}$	4.34 (0.89)	32.66 (0.89)	5.14 (0.43)	60.86 (0.43)
	$CSIS_{\mathcal{I}}$+Lasso	3.50 (1.41)	7.10 (5.71)	5.01 (0.61)	6.46 (6.31)
	$CSIS_{\mathcal{I}}$+MCP	3.23 (1.28)	3.37 (2.59)	4.70 (0.85)	1.07 (1.43)
	$CSIS_{\mathcal{I}}$+SCAD	3.64 (1.29)	7.44 (4.05)	5.01 (0.60)	3.76 (3.00)
3B	SC_{η_1}	9.62 (3.23)	6.01 (3.55)	15.00 (0.00)	6.53 (2.64)
	SC_{η_2}	4.78 (3.33)	0.84 (1.76)	15.00 (0.00)	4.08 (3.02)
	SC_{η_3}	1.86 (1.47)	0.07 (0.27)	15.00 (0.04)	2.07 (2.51)
	SIS	11.22 (1.29)	25.78 (1.29)	14.73 (0.51)	51.27 (0.51)
	SIS+Lasso	11.15 (1.38)	23.62 (2.61)	14.73 (0.51)	43.84 (5.01)
	SIS+MCP	10.54 (1.53)	12.05 (3.64)	14.73 (0.51)	12.57 (5.54)
	SIS+SCAD	10.59 (1.48)	12.99 (4.08)	14.73 (0.51)	13.61 (5.99)
	$CSIS_{\mathcal{A}}$	11.70 (1.27)	25.30 (1.27)	14.78 (0.46)	51.22 (0.46)
	$CSIS_{\mathcal{A}}$+Lasso	11.62 (1.34)	23.14 (2.42)	14.78 (0.46)	43.83 (4.58)
	$CSIS_{\mathcal{A}}$+MCP	11.05 (1.48)	11.17 (3.80)	14.78 (0.47)	11.98 (5.10)
	$CSIS_{\mathcal{A}}$+SCAD	11.01 (1.49)	11.79 (4.09)	14.78 (0.47)	12.75 (5.69)
	$CSIS_{\mathcal{I}}$	11.11 (1.30)	25.89 (1.30)	14.72 (0.51)	51.28 (0.51)
	$CSIS_{\mathcal{I}}$+Lasso	11.04 (1.38)	23.87 (2.34)	14.72 (0.51)	44.07 (5.04)
	$CSIS_{\mathcal{I}}$+MCP	10.44 (1.50)	12.32 (3.78)	14.72 (0.52)	12.56 (5.62)
	$CSIS_{\mathcal{I}}$+SCAD	10.47 (1.46)	13.31 (4.18)	14.72 (0.51)	13.61 (5.95)
4B	SC_{η_1}	7.25 (2.02)	6.03 (3.72)	14.60 (0.88)	6.77 (3.70)
	SC_{η_2}	5.01 (1.73)	0.81 (1.92)	14.10 (1.57)	2.74 (3.32)
	SC_{η_3}	3.59 (1.11)	0.04 (0.33)	11.09 (3.63)	0.48 (1.38)
	SIS	13.72 (0.97)	23.28 (0.97)	14.98 (0.13)	51.02 (0.13)
	SIS+Lasso	13.56 (1.01)	19.94 (2.12)	14.95 (0.25)	35.87 (14.65)
	SIS+MCP	11.35 (1.58)	7.04 (2.84)	14.85 (0.44)	12.89 (4.14)
	SIS+SCAD	11.23 (1.61)	7.28 (3.06)	14.84 (0.42)	13.41 (4.50)
	$CSIS_{\mathcal{A}}$	12.64 (1.10)	24.36 (1.10)	14.56 (0.63)	51.44 (0.63)
	$CSIS_{\mathcal{A}}$+Lasso	12.42 (1.20)	21.26 (2.21)	14.53 (0.70)	35.97 (15.63)
	$CSIS_{\mathcal{A}}$+MCP	10.29 (1.63)	8.76 (3.11)	14.38 (0.86)	14.13 (4.65)
	$CSIS_{\mathcal{A}}$+SCAD	10.22 (1.67)	8.87 (3.22)	14.36 (0.90)	14.44 (4.96)
	$CSIS_{\mathcal{I}}$	13.71 (0.97)	23.29 (0.97)	14.98 (0.13)	51.02 (0.13)
	$CSIS_{\mathcal{I}}$+Lasso	13.54 (1.04)	20.07 (2.16)	14.96 (0.22)	36.22 (14.14)
	$CSIS_{\mathcal{I}}$+MCP	11.35 (1.59)	7.04 (2.73)	14.85 (0.42)	12.89 (4.12)
	$CSIS_{\mathcal{I}}$+SCAD	11.27 (1.62)	7.27 (3.06)	14.84 (0.44)	13.21 (4.33)

Table 3
Comparisons of competing methods with Poisson regression models

		$(n, p) = (200, 1000)$		$(n, p) = (400, 1000)$	
Example	Method	TP	FP	TP	FP
1C	SC_{η_1}	5.12 (0.89)	2.14 (1.86)	5.90 (0.30)	1.12 (0.51)
	SC_{η_2}	5.06 (0.93)	1.77 (1.57)	5.87 (0.34)	0.88 (0.42)
	SC_{η_3}	5.02 (0.99)	1.52 (1.32)	5.85 (0.36)	0.51 (0.52)
	SIS	3.13 (0.58)	33.87 (0.58)	3.95 (0.76)	62.05 (0.76)
	SIS+Lasso	3.00 (0.70)	26.19 (4.90)	3.90 (0.80)	35.26 (15.54)
	SIS+MCP	2.69 (0.83)	14.76 (4.37)	3.84 (0.90)	17.25 (12.80)
	SIS+SCAD	2.80 (0.80)	18.99 (5.14)	3.87 (0.85)	18.86 (17.25)
	$\text{CSIS}_{\mathcal{A}}$	3.97 (0.62)	33.03 (0.62)	4.82 (0.51)	61.18 (0.51)
	$\text{CSIS}_{\mathcal{A}}$+Lasso	3.95 (0.63)	25.91 (3.51)	4.80 (0.50)	31.11 (12.39)
	$\text{CSIS}_{\mathcal{A}}$+MCP	3.89 (0.69)	10.79 (5.83)	4.81 (0.50)	4.48 (8.29)
	$\text{CSIS}_{\mathcal{A}}$+SCAD	3.91 (0.68)	14.74 (7.35)	4.80 (0.50)	3.74 (8.83)
	$\text{CSIS}_{\mathcal{I}}$	3.14 (0.59)	33.86 (0.59)	3.93 (0.77)	62.07 (0.77)
	$\text{CSIS}_{\mathcal{I}}$+Lasso	3.01 (0.69)	26.39 (5.07)	3.89 (0.81)	35.24 (15.93)
	$\text{CSIS}_{\mathcal{I}}$+MCP	2.70 (0.84)	14.71 (4.54)	3.82 (0.91)	17.51 (12.71)
	$\text{CSIS}_{\mathcal{I}}$+SCAD	2.84 (0.79)	19.02 (5.36)	3.85 (0.86)	19.19 (17.39)
2C	SC_{η_1}	4.30 (1.29)	2.40 (1.57)	5.97 (0.19)	1.20 (0.60)
	SC_{η_2}	4.06 (1.33)	1.84 (1.34)	5.96 (0.21)	0.97 (0.48)
	SC_{η_3}	3.82 (1.32)	1.55 (1.25)	5.93 (0.28)	0.63 (0.57)
	SIS	3.47 (1.12)	33.53 (1.12)	4.80 (0.45)	61.20 (0.45)
	SIS+Lasso	2.92 (1.10)	8.57 (3.51)	4.58 (0.63)	7.00 (3.51)
	SIS+MCP	2.08 (1.02)	2.49 (1.60)	3.92 (0.83)	1.57 (1.67)
	SIS+SCAD	2.24 (1.07)	3.42 (2.42)	3.99 (0.84)	1.82 (2.26)
	$\text{CSIS}_{\mathcal{A}}$	4.84 (0.95)	32.16 (0.95)	5.84 (0.40)	60.16 (0.40)
	$\text{CSIS}_{\mathcal{A}}$+Lasso	4.51 (1.34)	12.88 (3.93)	5.79 (0.54)	15.14 (4.38)
	$\text{CSIS}_{\mathcal{A}}$+MCP	3.99 (1.63)	2.38 (1.53)	5.72 (0.74)	1.16 (1.08)
	$\text{CSIS}_{\mathcal{A}}$+SCAD	4.08 (1.58)	3.32 (2.45)	5.73 (0.68)	1.05 (1.73)
	$\text{CSIS}_{\mathcal{I}}$	3.88 (1.03)	33.12 (1.03)	5.03 (0.51)	60.97 (0.51)
	$\text{CSIS}_{\mathcal{I}}$+Lasso	3.40 (1.18)	10.42 (4.13)	4.82 (0.71)	9.28 (4.73)
	$\text{CSIS}_{\mathcal{I}}$+MCP	2.59 (1.35)	3.09 (1.76)	4.34 (0.99)	1.85 (1.56)
	$\text{CSIS}_{\mathcal{I}}$+SCAD	2.71 (1.32)	4.07 (2.71)	4.39 (0.98)	2.29 (2.32)
3C	SC_{η_1}	8.47 (2.90)	5.38 (2.63)	15.00 (0.04)	0.83 (0.63)
	SC_{η_2}	7.59 (2.74)	3.85 (2.21)	14.99 (0.08)	0.42 (0.56)
	SC_{η_3}	6.75 (2.54)	2.90 (1.90)	14.98 (0.15)	0.16 (0.42)
	SIS	9.05 (1.69)	27.95 (1.69)	14.02 (1.08)	51.98 (1.08)
	SIS+Lasso	9.00 (1.71)	20.70 (3.22)	14.02 (1.08)	29.76 (5.88)
	SIS+MCP	8.61 (1.84)	9.33 (3.44)	14.02 (1.08)	3.25 (3.52)
	SIS+SCAD	8.81 (1.79)	12.96 (3.70)	14.02 (1.08)	6.50 (4.95)
	$\text{CSIS}_{\mathcal{A}}$	9.63 (1.60)	27.37 (1.60)	14.10 (1.07)	51.90 (1.07)
	$\text{CSIS}_{\mathcal{A}}$+Lasso	9.58 (1.61)	20.21 (3.20)	14.10 (1.07)	29.38 (6.10)
	$\text{CSIS}_{\mathcal{A}}$+MCP	9.19 (1.73)	8.95 (3.31)	14.10 (1.07)	3.09 (3.32)
	$\text{CSIS}_{\mathcal{A}}$+SCAD	9.39 (1.69)	12.51 (3.67)	14.10 (1.07)	6.36 (4.99)
	$\text{CSIS}_{\mathcal{I}}$	8.94 (1.68)	28.06 (1.68)	13.97 (1.09)	52.03 (1.09)
	$\text{CSIS}_{\mathcal{I}}$+Lasso	8.91 (1.70)	20.60 (3.33)	13.97 (1.09)	29.61 (6.00)
	$\text{CSIS}_{\mathcal{I}}$+MCP	8.49 (1.84)	9.47 (3.35)	13.97 (1.09)	3.43 (3.60)
	$\text{CSIS}_{\mathcal{I}}$+SCAD	8.72 (1.74)	13.11 (3.70)	13.97 (1.09)	6.47 (4.75)
4C	SC_{η_1}	8.68 (2.56)	6.62 (2.98)	15.00 (0.00)	1.53 (0.89)
	SC_{η_2}	8.16 (2.43)	5.06 (2.51)	15.00 (0.09)	1.42 (0.89)
	SC_{η_3}	7.70 (2.33)	4.02 (2.14)	15.00 (0.09)	1.25 (0.96)
	SIS	10.94 (2.02)	26.06 (2.02)	14.46 (0.88)	51.54 (0.88)
	SIS+Lasso	10.90 (2.05)	14.89 (3.47)	14.46 (0.88)	18.63 (4.66)
	SIS+MCP	10.40 (2.16)	5.22 (2.70)	14.46 (0.88)	1.48 (2.60)
	SIS+SCAD	10.51 (2.19)	7.21 (3.36)	14.46 (0.88)	2.01 (3.18)
	$\text{CSIS}_{\mathcal{A}}$	10.27 (1.74)	26.73 (1.74)	13.46 (1.14)	52.54 (1.14)
	$\text{CSIS}_{\mathcal{A}}$+Lasso	10.24 (1.76)	15.67 (3.51)	13.46 (1.14)	20.57 (5.29)
	$\text{CSIS}_{\mathcal{A}}$+MCP	9.72 (1.91)	6.02 (2.77)	13.45 (1.15)	4.28 (4.18)
	$\text{CSIS}_{\mathcal{A}}$+SCAD	9.86 (1.89)	7.97 (3.34)	13.45 (1.15)	5.35 (5.17)
	$\text{CSIS}_{\mathcal{I}}$	11.10 (1.89)	25.90 (1.89)	14.43 (0.92)	51.57 (0.92)
	$\text{CSIS}_{\mathcal{I}}$+Lasso	11.07 (1.92)	14.61 (3.38)	14.43 (0.92)	18.65 (4.67)
	$\text{CSIS}_{\mathcal{I}}$+MCP	10.51 (2.07)	5.11 (2.73)	14.43 (0.92)	1.56 (2.58)
	$\text{CSIS}_{\mathcal{I}}$+SCAD	10.68 (2.07)	6.87 (3.25)	14.43 (0.92)	1.97 (3.07)

Table 4
Selected genes by different methods

	SC_{η_1}	SC_{η_2}	SC_{η_3}	SIS	SIS+ Lasso	SIS+ MCP	SIS+ SCAD	CSIS	CSIS+ Lasso	CSIS+ MCP	CSIS+ SCAD	Frequency
FAM127A	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	11
TNFRSF11A	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	10
PPFIBP1	✓			✓	✓	✓	✓	✓	✓	✓	✓	9
BLVRA				✓	✓	✓	✓	✓	✓	✓	✓	8
C11orf82				✓	✓	✓	✓	✓	✓	✓	✓	8
STRBP	✓	✓	✓	✓	✓	✓	✓					7
IDS				✓	✓		✓	✓	✓	✓	✓	7
RHPN1-AS1				✓	✓		✓	✓	✓	✓	✓	7
SYN1				✓	✓			✓	✓	✓	✓	6
NCAN				✓				✓	✓	✓	✓	5
EIF2S1				✓				✓	✓	✓	✓	5
KHDRBS1				✓			✓	✓	✓		✓	5

Table 5
Final model size and leave-one-out prediction errors

	SC_{η_1}	SC_{η_2}	SC_{η_3}	SIS	SIS+ Lasso	SIS+ MCP	SIS+ SCAD	CSIS	CSIS+ Lasso	CSIS+ MCP	CSIS+ SCAD
Prediction error	0.17	0.18	0.18	0.19	0.17	0.17	0.17	0.19	0.18	0.17	0.17
Model size	5	3	3	23	14	9	13	23	17	13	14