entropy

MDPI

Article

Consistent Estimation of Generalized Linear Models with High Dimensional Predictors via Stepwise Regression

Alex Pijyan [1], Qi Zheng [2], Hyokyoung G. Hong [1,*] and Yi Li [3]

[1] Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824, USA; pijyanal@msu.edu
[2] Department of Bioinformatics and Biostatistics, University of Louisville, Louisville, KY 40202, USA; qi.zheng@louisville.edu
[3] Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA; yili@umich.edu
* Correspondence: hhong@msu.edu

Received: 1 August 2020; Accepted: 28 August 2020; Published: 31 August 2020

Abstract: Predictive models play a central role in decision making. Penalized regression approaches, such as least absolute shrinkage and selection operator (LASSO), have been widely used to construct predictive models and explain the impacts of the selected predictors, but the estimates are typically biased. Moreover, when data are ultrahigh-dimensional, penalized regression is usable only after applying variable screening methods to downsize variables. We propose a stepwise procedure for fitting generalized linear models with ultrahigh dimensional predictors. Our procedure can provide a final model; control both false negatives and false positives; and yield consistent estimates, which are useful to gauge the actual effect size of risk factors. Simulations and applications to two clinical studies verify the utility of the method.

Keywords: estimation consistency; generalized linear models; high dimensional predictors; model selection; stepwise regression

1. Introduction

In the era of precision medicine, constructing interpretable and accurate predictive models, based on patients' demographic characteristics, clinical conditions and molecular biomarkers, has been crucial for disease prevention, early diagnosis and targeted therapy [1]. When the number of predictors is moderate, penalized regression approaches such as least absolute shrinkage and selection operator (LASSO) by [2] have been used to construct predictive models and explain the impacts of the selected predictors. However, in ultrahigh dimensional settings where p is in the exponential order of n, penalized methods may incur computational challenges [3], may not reach globally optimal solutions and often generate biased estimates [4]. Sure independence screening (SIS) proposed by [5] has emerged as a powerful tool for modeling ultrahigh dimensional data. However, the method relies on a partial faithfulness assumption, which stipulates that jointly important variables must be marginally important, an assumption that may not be always realistic. To relieve this condition, some iterative procedures, such as ISIS [5], have been adopted to repeatedly screen variables based on the residuals from the previous iterations, but with heavy computation and unclear theoretical properties. Conditional screening approaches [see, e.g., [6]] have, to some extent, addressed the challenge. However, screening methods do not directly generate a final model, and post-screening regularization methods, such as LASSO, are recommended by [5] to produce a final model.

For generating a final predictive model in ultrahigh dimensional settings, recent years have seen a surging interest of performing forward regression, an old technique for model selection; see [7–9],

among many others. Under some regularity conditions and with some proper stopping criteria, forward regression can achieve screening consistency and sequentially select variables according to metrics such as AIC, BIC or R^2. Closely related to forward selection also, is least angle regression (LARS) [10], a widely used model selection algorithm for high-dimensional models. In the generalized linear model setting [11,12], proposed differential geometrical LARS (dgLARS) based on a differential geometrical extension of LARS.

However, these methods have drawbacks. First, once a variable is identified by the forward selection, it is not removable from the list of selected variables. Hence, false positives are unavoidable without a systematic elimination procedure. Second, most of the existing works focus on variable selection and are silent with respect to estimation accuracy.

To address the first issue, some works have been proposed to add backward elimination steps once forward selection is accomplished, as backward elimination may further eliminate false positives from the variables selected by forward selection. For example, ref. [13,14] proposed a stepwise selection for linear regression models in high-dimensional settings and proved model selection consistency. However, it is unclear whether the results hold for high-dimensional generalized linear models (GLMs); Ref. [15] proposed a similar stepwise algorithm in high-dimensional GLM settings, but with no theoretical properties on model selection. Moreover, none of the relevant works have touched upon the accuracy of estimation.

We extend a stepwise regression method to accommodate GLMs with high-dimensional predictors. Our method embraces both model selection and estimation. It starts with an empty model or pre-specified predictors, scans all features and sequentially selects features, and conducts backward elimination once forward selection is completed. Our proposal controls both false negatives and false positives in high dimensional settings: the forward selection steps recruit variables in an inclusive way by allowing some false positives for the sake of avoiding false negatives, while the backward selection steps eliminate the potential false positives from the recruited variables. We use different stopping criteria in the forward and backward selection steps, to control the numbers of false positives and false negatives. Moreover, we prove that, under a sparsity assumption of the true model, the proposed approach can discover all of the relevant predictors within a finite number of steps, and the estimated coefficients are consistent, a property still unknown to the literature. Finally, our GLM framework enables our work to accommodate a wide range of data types, such as binary, categorical and count data.

To recap, our proposed method distinguishes from the existing stepwise approaches in high dimensional settings. For example, it improves [13,14] by extending the work to a more broad GLM setting and [15] by establishing the theoretical properties.

Compared with the other variable selection and screening works, our method produces a final model in ultrahigh dimensional settings, without applying a pre-screening step which may produce unintended false negatives. Under some regularity conditions, the method identifies or includes the true model with probability going to 1. Moreover, unlike the penalized approaches such as LASSO, the coefficients estimated by our stepwise selection procedure in the final model will be consistent, which are useful for gauging the real effect sizes of risk factors.

2. Method

Let $(\mathbf{X}_i, Y_i)$, $i = 1, \ldots, n$, denote n independently and identically distributed (i.i.d.) copies of $(\mathbf{X}, Y)$. Here, $\mathbf{X} = (1, X_1, \ldots, X_p)^{\mathrm{T}}$ is a $(p+1)$-dimensional predictor vector with $X_0 = 1$ corresponding to the intercept term, and Y is an outcome. Suppose that the conditional density of Y, given $\mathbf{X}$, belongs to a linear exponential family:

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

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is the vector of coefficients; β_0 is the intercept; and $\mathcal{A}(\cdot)$ and $b(\cdot)$ are known functions. Model (1), with a canonical link function and a unit dispersion parameter, belongs to a larger exponential family [16]. Further, $b(\cdot)$ is assumed twice continuously differentiable with a non-negative second derivative $b''(\cdot)$. We use $\mu(\cdot)$ and $\sigma(\cdot)$ to denote $b'(\cdot)$ and $b''(\cdot)$, i.e., the mean and variance functions, respectively. For example, $b(\theta) = \log(1 + \exp(\theta))$ in a logistic distribution and $b(\theta) = \exp(\theta)$ in a Poisson distribution.

Let $L(u, v) = uv - b(u)$ and $\mathbb{E}_n\{f(\xi)\} = n^{-1}\sum_{i=1}^n f(\xi_i)$ denote the mean of $\{f(\xi_i)\}_{i=1}^n$ for a sequence of i.i.d. random variables ξ_i $(i = 1, \ldots, n)$ and a non-random function $f(\cdot)$. Based on the i.i.d. observations, the log-likelihood function is

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

We use $\boldsymbol{\beta}_* = (\beta_{*0}, \beta_{*1}, \ldots, \beta_{*p})^{\mathrm{T}}$ to denote the true values of $\boldsymbol{\beta}$. Then the true model is $\mathcal{M} = \{j : \beta_{*j} \neq 0, j \geq 1\} \cup \{0\}$, which consists of the intercept and all variables with nonzero effects. Overarching goals of ultra-high dimensional data analysis are to identify $\mathcal{M}$ and estimate β_{*j} for $j \in \mathcal{M}$. While most of the relevant literature [8,9] is on estimating $\mathcal{M}$, this work is to accomplish both identification of $\mathcal{M}$ and estimation of β_{*j}.

When p is in the exponential order of n, we aim to generate a predictive model that contains the true model with high probability, and provide consistent estimates of regression coefficients. We further introduce the following notation. For a generic index set $S \subset \{0, 1, \ldots, p\}$ and a $(p+1)$-dimensional vector $\mathbf{A}$, we use S^c to denote the complement of a set S and $\mathbf{A}_S = \{A_j : j \in S\}$ to denote the subvector of $\mathbf{A}$ corresponding to S. For instance, if $S = \{2, 3, 4\}$, then $\mathbf{X}_{iS} = (X_{i2}, X_{i3}, X_{i4})^{\mathrm{T}}$. Moreover, denote by $\ell_S(\boldsymbol{\beta}_S) = \mathbb{E}_n\{L(\mathbf{X}_S^{\mathrm{T}}\boldsymbol{\beta}_S, Y)\}$ the log-likelihood of the regression model of Y on $\mathbf{X}_S$ and denote by $\hat{\boldsymbol{\beta}}_S$ the maximizer of $\ell_S(\boldsymbol{\beta}_S)$. Under model (1), we elaborate on the idea of stepwise (details in the supplementary materials) selection, consisting of the forward and backward stages.

Forward stage: We start with F_0, a set of variables that need to be included according to some *a priori* knowledge, such as clinically important factors and conditions. If no such information is available, F_0 is set to be $\{0\}$, corresponding to a null model. We sequentially add covariates as follows:

$$F_0 \subset F_1 \subset F_2 \subset \cdots \subset F_k,$$

where $F_k \subset \{0, 1, \ldots, p\}$ is the index set of the selected covariates upon completion of the kth step, with $k \geq 0$. At the $(k+1)$th step, we append new variables to F_k one at a time and refit GLMs: for every $j \in F_k^c$, we let $F_{k,j} = F_k \cup \{j\}$, obtain $\hat{\boldsymbol{\beta}}_{F_{k,j}}$ by maximizing $\ell_{F_{k,j}}(\boldsymbol{\beta}_{F_{k,j}})$, and compute the increment of log-likelihood,

$$\ell_{F_{k,j}}(\hat{\boldsymbol{\beta}}_{F_{k,j}}) - \ell_{F_k}(\hat{\boldsymbol{\beta}}_{F_k}).$$

Then the index of a new candidate variable is determined to be

$$j_{k+1} = \arg\max_{j \in F_k^c} \ell_{F_{k,j}}(\hat{\boldsymbol{\beta}}_{F_{k,j}}) - \ell_{F_k}(\hat{\boldsymbol{\beta}}_{F_k}).$$

Additionally, we update $F_{k+1} = F_k \cup \{j_{k+1}\}$. We then need to decide whether to stop at the kth step or move on to the $(k+1)$th step with F_{k+1}. To do so, we use the following EBIC criterion:

$$\mathrm{EBIC}(F_{k+1}) = -2\ell_{F_{k+1}}(\hat{\boldsymbol{\beta}}_{F_{k+1}}) + |F_{k+1}|n^{-1}(\log n + 2\eta_1 \log p), \tag{3}$$

where the second term is motivated by [17] and $|F|$ denotes the cardinality of a set F.

The forward selection stops if $\mathrm{EBIC}(F_{k+1}) > \mathrm{EBIC}(F_k)$. We denote the stopping step by k^* and the set of variables selected so far by F_{k^*}.

Backward stage: Upon the completion of forward stage, backward elimination, starting with $B_0 = F_{k^*}$, sequentially drops covariates as follows:

$$B_0 \supset B_1 \supset B_2 \supset \cdots \supset B_k,$$

where B_k is the index set of the remaining covariates upon the completion of the kth step of the backward stage, with $k \geq 0$. At the $(k+1)$th backward step and for every $j \in B_k$, we let $B_{k/j} = B_k \setminus \{j\}$, obtain $\hat{\boldsymbol{\beta}}_{B_{k/j}}$ by maximizing $\ell(\boldsymbol{\beta}_{B_{k/j}})$, and calculating the difference of the log-likelihoods between these two nested models:

$$\ell_{B_k}(\hat{\boldsymbol{\beta}}_{B_k}) - \ell_{B_{k/j}}(\hat{\boldsymbol{\beta}}_{B_{k/j}}).$$

The variable that can be removed from the current set of variables is indexed by

$$j_{k+1} = \arg\min_{j \in B_k} \ell_{B_k}(\hat{\boldsymbol{\beta}}_{B_k}) - \ell_{B_{k/j}}(\hat{\boldsymbol{\beta}}_{B_{k/j}}).$$

Let $B_{k+1} = B_k \setminus \{j_{k+1}\}$. We determine whether to stop at the kth step or move on to the $(k+1)$th step of the backward stage according to the following BIC criterion:

$$\mathrm{BIC}(B_{k+1}) = -2\ell_{B_{k+1}}(\hat{\boldsymbol{\beta}}_{B_{k+1}}) + \eta_2 n^{-1} |B_{k+1}| \log n. \tag{4}$$

If $\mathrm{BIC}(B_{k+1}) > \mathrm{BIC}(B_k)$, we end the backward stage at the kth step. Let k^{**} denote the stopping step and we declare the selected model $B_{k^{**}}$ to be the final model. Thus, $\hat{\mathcal{M}} = B_{k^{**}}$ is the estimate of $\mathcal{M}$. As the backward stage starts with the k^* variables selected by forward selection, k^{**} cannot exceed k^*.

A strength of our algorithm, termed STEPWISE hereafter, is the added flexibility with η_1 and η_2 in the stopping criteria for controlling the false negatives and positives. For example, a smaller value of η_1 close to zero in the forward selection step will likely include more variables, and thus incur more false positives and less false negatives, whereas a larger value of η_1 will recruit too few variables and cause too many false negatives. Similarly, in the backward selection step, a large η_2 would eliminate more variables and therefore further reduce more false positives, and vice versa for a small η_2. While finding optimal η_1 and η_2 is not trivial, our numerical experiences suggest a small η_1 and a large η_2 may well balance the false negatives and positives. When $\eta_2 = 0$, no variables can be dropped after forward selection; hence, our proposal includes forward selection as a special case.

Moreover, [8] proposed a sequentially conditioning approach based on offset terms that absorb the prior information. However, our numerical experiments indicate that the offset approach may be suboptimal compared to our full stepwise optimization approach, which will be demonstrated in the simulation studies.

3. Theoretical Properties

With a column vector $\mathbf{v}$, let $\|\mathbf{v}\|_q$ denote the L_q-norm for any $q \geq 1$. For simplicity, we denote the L_2-norm of $\mathbf{v}$ by $\|\mathbf{v}\|$, and denote $\mathbf{v}\mathbf{v}^{\mathrm{T}}$ by $\mathbf{v}^{\otimes 2}$. We use $C_1, C_2, \ldots,$ to denote some generic constants that do not depend on n and may change from line to line. The following regularity conditions are set.

1. There exist a positive integer q satisfying $|\mathcal{M}| \leq q$ and $q \log p = o(n^{1/3})$ and a constant $K > 0$ such that $\sup_{|S| \leq q} \|\boldsymbol{\beta}_S^*\|_1 \leq K$, where $\boldsymbol{\beta}_S^* = \arg\max_{\boldsymbol{\beta}_S} E\left[\ell_S(\boldsymbol{\beta}_S)\right]$ is termed the least false value of model S.
2. $\|\mathbf{X}\|_\infty \leq K$. In addition, $E(X_j) = 0$ and $E(X_j^2) = 1$ for $j \geq 1$.
3. 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, i.e., $E[|\epsilon_i|^m] \leq m! M^m$ for all $m \geq 1$.
4. $|\sigma(a) - \sigma(b)| \leq K|a - b|$ and $\sigma_{\min} := \inf_{|t| \leq K^3} |b''(t)|$ is bounded below.

5. There exist two positive constants, $\kappa_{\min}$ and $\kappa_{\max}$ such that $0 < \kappa_{\min} < \Lambda\left(E\left(\mathbf{X}_S^{\otimes 2}\right)\right) < \kappa_{\max} < \infty$, uniformly in $S \subset \{0, 1, \ldots, p\}$ satisfying $|S| \leq q$, where $\Lambda(\mathbf{A})$ is the collection of all eigenvalues of a square matrix $\mathbf{A}$.
6. $\min_{S:\mathcal{M}\not\subseteq S, |S|\leq q} D_S > Cn^{-\alpha}$ for some constants $C > 0$ and $\alpha > 0$ that satisfies $qn^{-1+4\alpha}\log p \to 0$, where $D_S = \max_{j\in S^c\cap\mathcal{M}} \left|E\left[\left(\mu(\boldsymbol{\beta}_*^{\mathrm{T}}\mathbf{X}) - \mu(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S)\right)X_j\right]\right|$.

Condition (1), as assumed in [8,18], is an alternative to the Lipschitz assumption [5,19]. The bound of the model size allowed in the selection procedure or q is often required in model-based screening methods see, e.g., [8,20–22]. The bound should be large enough so that the correct model can be included, but not too large; otherwise, excessive noise variables would be included, leading to unstable and inconsistent estimates. Indeed, Conditions (1) and (6) reveal that the range of q depends on the true model size $|\mathcal{M}|$, the minimum signal strength, $n^{-\alpha}$ and the total number of covariates, p. The upper bound of q is $o((n^{1-4\alpha}/\log p) \wedge (n^{1/3}/\log p))$, ensuring the consistency of EBIC [17]. Condition (1) also implies that the parameter space under consideration can be restricted to $\mathbb{B} := \{\boldsymbol{\beta} \in \mathbb{R}^{p+1} : \|\boldsymbol{\beta}\|_1 \leq K^2\}$, for any model S with $|S| \leq q$. Condition (2), as assumed in [23,24], reflects that data are often standardized at the pre-processing stage. Condition (3) ensures that Y has a light tail, and is satisfied by Gaussian and discrete data, such as binary and count data [25]. Condition (4) is satisfied by common GLM models, such as Gaussian, binomial, Poisson and gamma distributions. Condition (5) represents the sparse Riesz condition [26] and Condition (6) is a strong "irrepresentable" condition, suggesting that $\mathcal{M}$ cannot be represented by a set of variables that does not include the true model. It further implies that adding a signal variable to a mis-specified model will increase the log-likelihood by a certain lower bound [8]. The signal rate is comparable to the conditions required by the other sequential methods, see, e.g., [7,22].

Theorem 1 develops a lower bound of the increment of the log-likelihood if the true model $\mathcal{M}$ is not yet included in a selected model S.

Theorem 1. *Suppose Conditions (1)–(6) hold. There exists some constant C_1 such that with probability at least 1–6*$\exp(-6q\log p)$,

$$\min_{S:\mathcal{M}\not\subseteq S, |S|<q}\left\{\max_{j\in S^c}\ell_{S\cup\{j\}}(\hat{\boldsymbol{\beta}}_{S\cup\{j\}}) - \ell_S(\hat{\boldsymbol{\beta}}_S)\right\} \geq C_1 n^{-2\alpha}.$$

Theorem 1 shows that, before the true model is included in the selected model, we can append a variable which will increase the log-likelihood by at least $C_1 n^{-2\alpha}$ with probability tending to 1. This ensures that in the forward stage, our proposed STEPWISE approach will keep searching for signal variables until the true model is contained. To see this, suppose at the kth step of the forward stage that F_k satisfies $M \not\subseteq F_k$ and $|F_k| < q$, and let r be the index selected by STEPWISE. By Theorem 1, we obtain that, for any $\eta_1 > 0$, when n is sufficiently large,

$$\begin{aligned}\mathrm{EBIC}(F_{k,r}) - \mathrm{EBIC}(F_k) &= -2\ell_{F_{k,r}}(\hat{\boldsymbol{\beta}}_{F_{k,r}}) + (|F_k|+1)n^{-1}(\log n + 2\eta_1\log p)\\ &\quad - \left[-2\ell_{F_k}(\hat{\boldsymbol{\beta}}_{F_k}) + |F_k|n^{-1}(\log n + 2\eta_1\log p)\right]\\ &\leq -2C_1 n^{-2\alpha} + n^{-1}(\log n + 2\eta_1\log p) < 0,\end{aligned}$$

with probability at least $1 - 6\exp(-6q\log p)$, where the last inequality is due to Condition (6). Therefore, with high probability the forward stage of STEPWISE continues as long as $\mathcal{M} \not\subseteq F_k$ and $|F_k| < q$. We next establish an upper bound of the number of steps in the forward stage needed to include the true model.

Theorem 2. *Under the same conditions as in Theorem 1 and if*

$$\max_{S:|S|\leq q}\left\{\max_{j\in\mathcal{S}^c\cap\mathcal{M}^c}\left|E\left[\left\{Y-\mu(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S)\right\}X_j\right]\right|\right\}=o(n^{-\alpha}),$$

then there exists some constant $C_2>2$ such that $\mathcal{M}\subset F_k$, for some F_k in the forward stage of STEPWISE and $k\leq C_2|\mathcal{M}|$, with probability at least $1-18\exp(-4q\log p)$.

The "max" condition, as assumed in Section 5.3 of [27], relaxes the partial orthogonality assumption that $\mathbf{X}_{\mathcal{M}^c}$ are independent of $\mathbf{X}_{\mathcal{M}}$, and ensures that with probability tending to 1, appending a signal variable increases log-likelihood more than adding a noise variable does, uniformly over all possible models S satisfying $\mathcal{M}\not\subseteq S, |S|<q$. This entails that the proposed procedure is much more likely to select a signal variable, in lieu of a noise variable, at each step. Since EBIC is a consistent model selection criterion [28,29], the following theorem guarantees termination of the proposed procedure with $\mathcal{M}\subset F_k$ for some k.

Theorem 3. *Under the same conditions as in Theorem 2 and if $\mathcal{M}\not\subset F_{k-1}$ and $\mathcal{M}\subset F_k$, the forward stage stops at the kth step with probability going to $1-\exp(-3q\log p)$.*

Theorem 3 ensures that the forward stage of STEPWISE will stop within a finite number of steps and will cover the true model with probability at least $1-q\exp(-3q\log p)\geq 1-\exp(-2q\log p)$. We next consider the backward stage and provide a probability bound of removing a signal from a set in which the set of true signals $\mathcal{M}$ is contained.

Theorem 4. *Under the same conditions as in Theorem 2, $BIC(S\backslash\{r\})-BIC(S)>0$ uniformly over $r\in\mathcal{M}$ and S satisfying $\mathcal{M}\subset S$ and $|S|\leq q$, with probability at least $1-6\exp(-6q\log p)$.*

Theorem 4 indicates that with probability at $1-6\exp(-6q\log p)$, BIC would decrease when removing a signal variable from a model that contains the true model. That is, with high probability, back elimination is to reduce false positives.

Recall that F_{k^*} denotes the model selected at the end of the forward selection stage. By Theorem 2, $\mathcal{M}\subset F_{k^*}$ with probability at least $1-18\exp(-4q\log p)$. Then Theorem 4 implies that at each step of the backward stage, a signal variable will not be removed from the model with probability at least $1-6\exp(-6q\log p)$. By Theorem 2, $|F_{k^*}|\leq C_2|\mathcal{M}|$. Thus, the backward elimination will carry out at most $(C_2-1)|\mathcal{M}|$ steps. Combining results from Theorems 2 and 3 yields that $\mathcal{M}\subset\hat{\mathcal{M}}$ with probability at least $1-18\exp(-4q\log p)-6(C_2-1)|\mathcal{M}|\exp(-6q\log p)$. Let $\hat{\boldsymbol{\beta}}$ be the estimate of $\boldsymbol{\beta}_*$ in model (1) at the termination of STEPWISE. By convention, the estimates of the coefficients of the unselected covariates are 0.

Theorem 5. *Under the same conditions as in Theorem 2, we have that $\mathcal{M}\subseteq\hat{\mathcal{M}}$ and*

$$\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_*\|\to 0$$

in probability.

The theorem warrants that the proposed STEPWISE yields consistent estimates, a property not shared by many regularized methods, including LASSO. Our later simulations verified this. Proof of main theorems and lemmas are provided in Appendix A.

4. Simulation Studies

We compared the proposal with the other competing methods, including the penalized methods, such as least absolute shrinkage and selection operator (LASSO); the differential geometric least angle

regression (dgLARS) [11,12]; the forward regression (FR) approach [7]; the sequentially conditioning (SC) approach [8]; and the screening methods, such as sure independence screening (SIS) [5], which is popular in practice. As SIS does not directly generate a predictive model, we applied LASSO for the top $[n/\log(n)]$ variables chosen by SIS and denoted the procedure by SIS+LASSO. As the FR, SC and STEPWISE approaches involve forward searching and to make them comparable, we applied the same stopping rule, for example, Equation (3) with the same γ, to their forward steps. In particular, the STEPWISE approach, with $\eta_1 = \gamma$ and $\eta_2 = 0$, is equivalent to FR and asymptotically equivalent to SC. By varying γ in FR and SC between γ_L and γ_H, we explored the impact of γ on inducing false positives and negatives. In our numerical studies, we fixed $\gamma_H = 10$ and set $\gamma_L = \eta_1$. To choose η_1 and η_2 in (3) and (4) in STEPWISE, we performed 5-fold cross-validation to minimize the mean squared prediction error (MSPE), and reported the results in Table 1. Since the proposed STEPWISE algorithm uses the (E)BIC criterion, for a fair comparison we chose the tuning parameter in dgLARS by using the BIC criterion as well, and coined the corresponding approach as dgLARS(BIC). The regularization parameter in LASSO was chosen via the following two approaches: (1) giving the smallest BIC for the models on the LASSO path, denoted by LASSO(BIC); (2) using the one-standard-error rule, denoted by LASSO(1SE), which chooses the most parsimonious model whose error is no more than one standard error above the error of the best model in cross-validation [30].

Table 1. The values of η_1 and η_2 used in the simulation studies.

	Normal Model	**Binomial Model**	**Poisson Model**
Example 1	(0.5, 3)	(0.5, 3)	(1, 3)
Example 2	(0.5, 3)	(1, 3)	(1, 3)
Example 3	(1, 3)	(0.5, 3)	(0.5, 1)
Example 4	(1, 3.5)	(0, 1)	(1, 3)
Example 5	(0.5, 3)	(0.5, 2)	(0.5, 3)

Note: values for η_1 and η_2 were searched on the grid $\{0, 0.25, 0.5, 1\}$ and $\{1, 2, 3, 3.5, 4, 4.5, 5\}$, respectively.

Denote by $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^{\mathrm{T}}$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$, the covariate vector for subject i $(1, \ldots, n)$ and the true coefficient vector. The following five examples generated $\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}$, the inner product of the coefficient and covariate vectors for each individual, which were used to generate outcomes from the normal, binomial and Poisson models.

Example 1. *For each i,*

$$c\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta} = c \times \left(\sum_{j=1}^{p_0} \beta_j X_{ij} + \sum_{j=p_0+1}^{p} \beta_j X_{ij} \right), \; i = 1, \ldots, n,$$

where $\beta_j = (-1)^{B_j}(4\log n/\sqrt{n} + |Z_j|)$, *for* $j = 1, \ldots, p_0$ *and* $\beta_j = 0$ *otherwise* B_j *was a binary random variable with* $P(B_j = 1) = 0.4$ *and* Z_j *was generated by a standard normal distribution;* $p_0 = 8$; X_{ij}*s were independently generated from a standardized exponential distribution, that is,* $\exp(1) - 1$. *Here and also in the other examples, c (specified later) controls the signal strengths.*

Example 2. *This scenario is the same as* **Example 1** *except that* X_{ij} *was independently generated from a standard normal distribution.*

Example 3. *For each i,*

$$c\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta} = c \times \left(\sum_{j=1}^{p_0} \beta_j X_{ij} + \sum_{j=p_0+1}^{p} \beta_j X_{ij}^{*} \right), \; i = 1, \ldots, n,$$

where $\beta_j = 2j$ *for* $1 \leq j \leq p_0$ *and* $p_0 = 5$. *We simulated every component of* $\mathbf{Z}_i = (Z_{ij}) \in R^p$ *and* $\mathbf{W}_i = (W_{ij}) \in R^p$ *independently from a standard normal distribution. Next, we generated* $\mathbf{X}_i$ *according to* $X_{ij} = (Z_{ij} + W_{ij})/\sqrt{2}$ *for* $1 \leq j \leq p_0$ *and* $X^*_{ij} = (Z_{ij} + \sum_{j'=1}^{p_0} Z_{ij'})/2$ *for* $p_0 < j \leq p$.

Example 4. *For each* i,

$$c\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta} = c \times \left(\sum_{j=1}^{500} \beta_j X_{ij} + \sum_{j=501}^{p} \beta_j X_{ij} \right), \; i = 1, \ldots, n,$$

where the first 500 X_{ij}*s were generated from the multivariate normal distribution with mean* $\mathbf{0}$ *and a covariance matrix with all of the diagonal entries being 1 and* $cov(X_{ij}, X_{ij'}) = 0.5^{|j-j'|}$ *for* $1 \leq j, j' \leq p$. *The remaining* $p - 500$ X_{ij}*s were generated through the autoregressive processes with* $X_{i,501} \sim$ *Unif(-2, 2),* $X_{ij} = 0.5\, X_{i,j-1} + 0.5\, X^*_{ij'}$ *for* $j = 502, \ldots, p$, *where* $X^*_{ij} \sim$ *Unif(-2, 2) were generated independently. The coefficients* β_j *for* $j = 1, \ldots, 7, 501, \ldots, 507$ *were generated from* $(-1)^{B_j}(4\log n/\sqrt{n} + |Z_j|)$, *where* B_j *was a binary random variable with* $P(B_j = 1) = 0.4$ *and* Z_j *was from a standard normal distribution. The remaining* β_j *were zeros.*

Example 5. *For each* i,

$$c\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta} = c \times (-0.5X_{i1} + X_{i2} + 0.5X_{i,100}), \; i = 1, \ldots, n,$$

where $\mathbf{X}_i$ *were generated from a multivariate normal distribution with mean* $\mathbf{0}$ *and a covariance matrix with all of the diagonal entries being 1 and* $cov(X_{ij}, X_{ij'}) = 0.9^{|j-j'|}$ *for* $1 \leq j, j' \leq p$. *All of the coefficients were zero except for* X_{i1}, X_{i2} *and* $X_{i,100}$.

Examples 1 and 3 were adopted from [7], while **Examples 2 and 4** were borrowed from [5,15], respectively. We then generated the responses from the following three models.

Normal model: $Y_i = c\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta} + \epsilon_i$ with $\epsilon_i \sim N(0, 1)$.
Binomial model: $Y_i \sim$ Bernoulli($\exp(c\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})/\{1 + \exp(c\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})\}$).
Poisson model: $Y_i \sim$ Poisson($\exp(c\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})$).

We considered $n = 400$ and $p = 1000$ throughout all of the examples. We specified the magnitude of the coefficients in the GLMs with a constant multiplier, c. For Examples 1–5, this constant was set, respectively for the normal, binomial and Poisson models, to be: (1, 1, 0.3), (1, 1.5, 0.3), (1, 1, 0.1), (1, 1.5, 0.3) and (1, 3, 2). For each parameter configuration, we simulated 500 independent data sets. We evaluated the performances of the methods by the criteria of true positives (TP), false positives (FP), the estimated probability of including the true models (PIT), the mean squared error (MSE) of $\hat{\boldsymbol{\beta}}$ and the mean squared prediction error (MSPE). To compute the MSPE, we randomly partitioned the samples into the training (75%) and testing (25%) sets. The models obtained from the training datasets were used to predict the responses in the testing datasets. Tables 2–4 report the average TP, FP, PIT, MSE and MSPE over 500 datasets along with the standard deviations. The findings are summarized below.

First, the proposed STEPWISE method was able to detect all the true signals with nearly zero FPs. Specifically, in all of the Examples, STEPWISE outperformed the other methods by detecting more TPs with fewer FPs, whereas LASSO, SIS+LASSO and dgLARS included much more FPs.

Second, though a smaller γ in FR and SC led to the inclusion of all TPs with a PIT close to 1, it incurred more FPs. On the other hand, a larger γ may eliminate some TPs, resulting in a smaller PIT and a larger MSPE.

Third, for the normal model, the STEPWISE method yielded an MSE close to 0, the smallest among all the competing methods. The binary and Poisson data challenged all of the methods, and the MSEs for all the methods were non-negligible. However, the STEPWISE method still produced

the lowest MSE. The results seemed to verify the consistency of $\hat{\boldsymbol{\beta}}$, which distinguished the proposed STEPWISE method from the other regularized methods and highlighted its ability to provide a more accurate means to characterize the effects of high dimensional predictors.

Table 2. Normal model.

Example	Method	TP	FP	PIT	MSE ($\times 10^{-4}$)	MSPE
1 ($p_0 = 8$)	LASSO(1SE)	8.00 (0.00)	5.48 (6.61)	1.00 (0.00)	2.45	1.148
	LASSO(BIC)	8.00 (0.00)	2.55 (2.48)	1.00 (0.00)	2.58	1.172
	SIS+LASSO(1SE)	8.00 (0.00)	6.59 (4.22)	1.00 (0.00)	1.49	1.042
	SIS+LASSO(BIC)	8.00 (0.00)	6.04 (3.33)	1.00 (0.00)	1.37	1.025
	dgLARS(BIC)	8.00 (0.00)	3.52(2.53)	1.00 (0.00)	2.25	1.130
	SC (γ_L)	8.00 (0.00)	3.01 (1.85)	1.00 (0.00)	1.09	0.895
	SC (γ_H)	7.60 (1.59)	0.00 (0.00)	0.94 (0.24)	14.56	5.081
	FR (γ_L)	8.00 (0.00)	2.96 (2.04)	1.00 (0.00)	1.08	0.896
	FR (γ_H)	7.88 (0.84)	0.00 (0.00)	0.98 (0.14)	3.74	2.040
	STEPWISE	8.00 (0.00)	0.00 (0.00)	1.00 (0.00)	0.21	0.972
2 ($p_0 = 8$)	LASSO(1SE)	8.00 (0.00)	4.74 (4.24)	1.00 (0.00)	2.46	1.154
	LASSO(BIC)	8.00 (0.00)	2.12 (2.02)	1.00 (0.00)	2.62	1.182
	SIS+LASSO	7.99 (0.10)	6.84 (4.57)	0.99 (0.10)	1.65	1.058
	SIS+LASSO(BIC)	7.99 (0.10)	6.11 (3.85)	0.99 (0.10)	1.56	1.041
	dgLARS(BIC)	8.00 (0.00)	3.26(2.62)	1.00 (0.00)	2.28	1.138
	SC (γ_L)	8.00 (0.00)	2.73 (1.53)	1.00 (0.00)	0.98	0.901
	SC (γ_H)	7.30 (2.11)	0.00 (0.00)	0.90 (0.30)	23.70	6.397
	FR (γ_L)	8.00 (0.00)	2.45 (1.65)	1.00 (0.00)	0.92	0.907
	FR (γ_H)	7.94 (0.60)	0.00 (0.00)	0.99 (0.00)	2.69	2.062
	STEPWISE	8.00 (0.00)	0.01 (0.10)	1.00 (0.00)	0.21	0.972
3 ($p_0 = 5$)	LASSO(1SE)	5.00 (0.00)	8.24 (2.63)	1.00 (0.00)	3.07	1.084
	LASSO(BIC)	5.00 (0.00)	12.33 (3.28)	1.00 (0.00)	27.97	2.398
	SIS+LASSO(1SE)	0.97 (0.26)	15.94 (2.93)	0.00 (0.00)	1406.22	76.024
	SIS+LASSO(BIC)	0.97 (0.26)	16.20 (2.81)	0.00 (0.00)	1354.54	71.017
	dgLARS(BIC)	5.00 (0.00)	53.91 (14.44)	1.00 (0.00)	6.63	0.979
	SC (γ_L)	4.48 (0.50)	0.25 (0.44)	0.48 (0.50)	21.74	3.086
	SC (γ_H)	4.48 (0.50)	0.14 (0.35)	0.48 (0.50)	21.70	2.065
	FR (γ_L)	5.00 (0.00)	0.23 (0.66)	1.00 (0.00)	0.27	0.973
	FR (γ_H)	5.00 (0.00)	0.14 (0.35)	1.00 (0.00)	0.15	0.074
	STEPWISE	5.00 (0.00)	0.03 (0.22)	1.00 (0.00)	0.18	0.976
4 ($p_0 = 14$)	LASSO(1SE)	14.00 (0.00)	29.84 (15.25)	1.00 (0.00)	13.97	1.148
	LASSO(BIC)	13.94 (0.24)	4.92 (5.54)	0.94 (0.24)	38.69	1.995
	SIS+LASSO(1SE)	11.44 (1.45)	15.19 (7.29)	0.05 (0.21)	133.38	4.714
	SIS+LASSO(BIC)	11.35 (1.51)	10.98 (7.19)	0.05 (0.21)	137.06	4.940
	dgLARS(BIC)	14.00 (0.00)	13.93 (6.68)	1.00 (0.00)	18.08	1.329
	SC (γ_L)	13.68 (0.60)	0.86 (0.62)	0.75 (0.44)	11.80	1.148
	SC (γ_H)	4.20 (2.80)	0.03 (0.17)	0.03 (0.17)	407.86	6.567
	FR (γ_L)	14.00 (0.00)	0.50 (0.76)	1.00 (0.00)	1.23	0.940
	FR (γ_H)	4.99 (3.07)	0.00 (0.00)	0.03 (0.17)	360.65	6.640
	STEPWISE	14.00 (0.00)	0.00 (0.00)	1.00 (0.00)	0.91	0.958
5 ($p_0 = 3$)	LASSO(1SE)	3.00 (0.00)	22.76 (9.05)	1.00 (0.00)	1.01	0.044
	LASSO(BIC)	3.00 (0.00)	8.29 (3.23)	1.00 (0.00)	1.75	0.054
	SIS+LASSO(1SE)	3.00 (0.00)	8.40 (3.10)	1.00 (0.00)	0.44	0.041
	SIS+LASSO(BIC)	3.00 (0.00)	9.58 (3.36)	1.00 (0.00)	0.29	0.040
	dgLARS(BIC)	3.00 (0.00)	13.39 (4.94)	1.00 (0.00)	1.28	0.048
	SC (γ_L)	3.00 (0.00)	1.47 (0.67)	1.00 (0.00)	0.03	0.038
	SC (γ_H)	2.01 (0.10)	0.01 (0.10)	0.01 (0.10)	4.51	0.008
	FR (γ_L)	3.00 (0.00)	1.21 (1.01)	1.00 (0.00)	0.03	0.038
	FR (γ_H)	3.00 (0.00)	0.00 (0.00)	1.00 (0.00)	0.01	0.003
	STEPWISE	3.00 (0.00)	0.00 (0.00)	1.00 (0.00)	0.01	0.039

Note: TP, true positives; FP, false positives; PIT, probability of including all true predictors in the selected predictors; MSE, mean squared error of $\hat{\boldsymbol{\beta}}$; MSPE, mean squared prediction error; numbers in the parentheses are standard deviations; LASSO(BIC), LASSO with the tuning parameter chosen to give the smallest BIC for the models on the LASSO path; LASSO(1SE), LASSO with the tuning parameter chosen by the one-standard-error rule; SIS+LASSO(BIC), sure independence screening by [5] followed by LASSO(BIC); SIS+LASSO(1SE), sure independence screening followed by LASSO(1SE); dgLARS(BIC), differential geometric least angle regression by [11,12] with the tuning parameter chosen to give the smallest BIC on the dgLARS path; SC(γ), sequentially conditioning approach by [8]; FR(γ), forward regression by [7]; STEPWISE, the proposed method; in FR and SC, the smaller and large values of γ are presented as γ_L and γ_H, respectively; p_0 denotes the number of true signals; LASSO(1SE), LASSO(BIC), SIS and dgLARS were conducted via `R` packages `glmnet` [31], `ncvreg` [32], `screening` [33] and `dglars` [34], respectively

Table 3. Binomial model.

Example	Method	TP	FP	PIT	MSE	MSPE
1 ($p_0 = 8$)	LASSO(1SE)	7.99 (0.10)	4.77 (5.56)	0.99 (0.10)	0.021	0.104
	LASSO(BIC)	7.99 (0.10)	3.19 (2.34)	0.99 (0.10)	0.021	0.104
	SIS+LASSO(1SE)	7.94 (0.24)	35.42 (6.77)	0.94 (0.24)	0.119	0.048
	SIS+LASSO(BIC)	7.94 (0.24)	16.83 (21.60)	0.94 (0.24)	0.119	0.073
	dgLARS(BIC)	8.00 (0.00)	3.27 (2.29)	1.00 (0.00)	0.019	0.102
	SC (γ_L)	8.00 (0.00)	2.81 (1.47)	1.00 (0.00)	0.009	0.073
	SC (γ_H)	1.02 (0.14)	0.00 (0.00)	0.00 (0.00)	0.030	0.028
	FR (γ_L)	8.00 (0.00)	3.90 (2.36)	1.00 (0.00)	0.032	0.066
	FR (γ_H)	2.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.025	0.027
	STEPWISE	7.98 (0.14)	0.08 (0.53)	0.98 (0.14)	0.002	0.094
2 ($p_0 = 8$)	LASSO(1SE)	7.98 (0.14)	3.29 (2.76)	0.98 (0.14)	0.054	0.073
	LASSO(BIC)	7.99 (0.10)	3.84 (2.72)	0.99 (0.10)	0.052	0.067
	SIS+LASSO(1SE)	7.92 (0.27)	28.20 (7.31)	0.92 (0.27)	0.038	0.030
	SIS+LASSO(BIC)	7.92 (0.27)	9.60 (12.92)	0.92 (0.27)	0.051	0.058
	dgLARS(BIC)	7.99 (0.10)	3.94 (2.65)	0.99 (0.10)	0.050	0.067
	SC (γ_L)	7.72 (0.45)	0.39 (0.49)	0.72 (0.45)	0.005	0.063
	SC (γ_H)	1.13 (0.37)	0.00 (0.00)	0.00 (0.00)	0.069	0.044
	FR (γ_L)	7.99 (0.10)	0.66 (0.76)	0.99 (0.10)	0.014	0.051
	FR (γ_H)	2.10 (0.30)	0.00 (0.00)	0.00 (0.00)	0.061	0.033
	STEPWISE	7.99 (0.10)	0.02 (0.14)	0.99 (0.10)	0.004	0.056
3 ($p_0 = 5$)	LASSO(1SE)	4.51 (0.52)	7.36 (2.57)	0.52 (0.50)	0.155	0.051
	LASSO(BIC)	4.98 (0.14)	5.97 (2.25)	0.98 (0.14)	0.118	0.037
	SIS+LASSO(1SE)	0.85 (0.46)	10.66 (3.01)	0.00 (0.00)	0.206	0.186
	SIS+LASSO(BIC)	0.85 (0.46)	12.10 (3.13)	0.00 (0.00)	0.197	0.185
	dgLARS(BIC)	4.92 (0.27)	16.21 (6.21)	0.92 (0.27)	0.112	0.035
	SC (γ_L)	4.32 (0.49)	0.47 (0.50)	0.33 (0.47)	0.016	0.048
	SC (γ_H)	2.62 (1.34)	0.42 (0.50)	0.00 (0.00)	0.104	0.066
	FR (γ_L)	4.98 (0.14)	0.67 (0.79)	0.98 (0.14)	0.020	0.033
	FR (γ_H)	2.98 (0.95)	0.40 (0.49)	0.00 (0.00)	0.087	0.043
	STEPWISE	4.97 (0.17)	0.04 (0.28)	0.97 (0.17)	0.014	0.034
4 ($p_0 = 14$)	LASSO(1SE)	9.96 (1.89)	6.78 (7.92)	0.01 (0.01)	0.112	0.107
	LASSO(BIC)	9.33 (1.86)	2.79 (2.87)	0.00 (0.00)	0.112	0.118
	SIS+LASSO(1SE)	10.03 (1.62)	28.01 (9.54)	0.03 (0.17)	0.098	0.070
	SIS+LASSO(BIC)	8.90 (1.99)	5.42 (10.64)	0.01 (0.10)	0.114	0.120
	dgLARS(BIC)	9.31 (1.85)	2.84 (2.86)	0.00 (0.00)	0.110	0.117
	SC (γ_L)	9.48 (1.40)	2.35 (2.14)	0.00 (0.00)	0.043	0.070
	SC (γ_H)	1.17 (0.40)	0.00 (0.00)	0.00 (0.00)	0.125	0.049
	FR (γ_L)	11.83 (1.39)	1.58 (1.60)	0.09 (0.29)	0.026	0.048
	FR (γ_H)	2.06 (0.24)	0.00 (0.00)	0.00 (0.00)	0.119	0.032
	STEPWISE	11.81 (1.42)	1.52 (1.58)	0.09 (0.29)	0.026	0.048
5 ($p_0 = 3$)	LASSO(1SE)	2.00 (0.00)	1.55 (1.76)	0.00 (0.00)	0.008	0.215
	LASSO(BIC)	2.00 (0.00)	1.86 (1.57)	0.00 (0.00)	0.008	0.213
	SIS+LASSO(1SE)	2.23 (0.42)	10.81 (6.45)	0.23 (0.42)	0.007	0.192
	SIS+LASSO(BIC)	2.10 (0.30)	3.60 (4.65)	0.10 (0.30)	0.007	0.206
	dgLARS(BIC)	2.00 (0.00)	1.64 (1.49)	0.00 (0.00)	0.008	0.213
	SC (γ_L)	2.27 (0.49)	7.16 (3.20)	0.29 (0.46)	0.060	0.166
	SC (γ_H)	1.87 (0.34)	0.03 (0.17)	0.00 (0.00)	0.005	0.030
	FR (γ_L)	2.96 (0.20)	8.88 (5.39)	0.96 (0.20)	0.013	0.147
	FR (γ_H)	1.97 (0.17)	0.03 (0.17)	0.00 (0.00)	0.005	0.019
	STEPWISE	2.89 (0.31)	0.76 (1.70)	0.89 (0.31)	0.001	0.194

Note: abbreviations are explained in the footnote of Table 2.

Table 4. Poisson model.

Example	Method	TP	FP	PIT	MSE	MSPE
1 ($p_0 = 8$)	LASSO(1SE)	7.93 (0.43)	4.64 (4.82)	0.96 (0.19)	0.001	4.236
	LASSO(BIC)	7.99 (0.10)	14.37 (14.54)	0.99 (0.10)	0.001	3.133
	SIS+LASSO(1SE)	7.89 (0.37)	25.37 (8.39)	0.91 (0.29)	0.001	3.247
	SIS+LASSO(BIC)	7.89 (0.37)	17.77 (11.70)	0.91 (0.29)	0.001	3.078
	dgLARS(BIC)	8.00 (0.00)	13.28 (14.31)	1.00 (0.00)	0.001	3.183
	SC (γ_L)	7.96 (0.20)	4.94 (3.46)	0.96 (0.20)	0.001	2.874
	SC (γ_H)	5.05 (1.70)	0.04 (0.24)	0.07 (0.26)	0.001	3.902
	FR (γ_L)	7.93 (0.26)	4.86 (3.73)	0.93 (0.26)	0.001	2.837
	FR (γ_H)	5.13 (1.61)	0.06 (0.31)	0.07 (0.26)	0.001	3.833
	STEPWISE	7.91 (0.29)	2.77 (2.91)	0.91 (0.29)	0.001	3.410
2 ($p_0 = 8$)	LASSO(1SE)	8.00 (0.00)	2.23 (3.52)	1.00 (0.00)	0.001	3.981
	LASSO(BIC)	8.00 (0.00)	8.98 (8.92)	1.00 (0.00)	0.001	3.107
	SIS+LASSO(1SE)	7.98 (0.14)	22.85 (7.08)	0.98 (0.14)	0.001	2.824
	SIS+LASSO(BIC)	7.98 (0.14)	13.55 (8.24)	0.98 (0.14)	0.001	2.937
	dgLARS(BIC)	8.00 (0.00)	8.91 (9.10)	1.00 (0.00)	0.001	3.099
	SC (γ_L)	8.00 (0.00)	3.89 (2.89)	1.00 (0.00)	0.000	2.979
	SC (γ_H)	5.68 (1.45)	0.00 (0.00)	0.12 (0.33)	0.001	3.971
	FR (γ_L)	8.00 (0.00)	3.60 (2.80)	1.00 (0.00)	0.000	3.032
	FR (γ_H)	5.71 (1.42)	0.00 (0.00)	0.10 (0.30)	0.001	3.911
	STEPWISE	7.98 (0.14)	2.00 (2.23)	0.98 (0.14)	0.000	3.589
3 ($p_0 = 5$)	LASSO(1SE)	4.37 (0.51)	6.88 (2.61)	0.38(0.48)	0.001	1.959
	LASSO(BIC)	4.79 (0.41)	5.62 (2.17)	0.79 (0.41)	0.000	2.044
	SIS+LASSO(1SE)	0.86 (0.47)	10.11 (2.55)	0.00 (0.00)	0.002	3.266
	SIS+LASSO(BIC)	0.86 (0.47)	11.86 (2.99)	0.00 (0.00)	0.002	3.160
	dgLARS(BIC)	4.55 (0.51)	18.29 (6.13)	0.56 (0.49)	0.001	1.877
	SC (γ_L)	4.73 (0.45)	0.53 (0.66)	0.73 (0.45)	0.000	2.479
	SC (γ_H)	2.84 (0.63)	0.40 (0.49)	0.00 (0.00)	0.001	0.664
	FR (γ_L)	4.54 (0.52)	1.98 (2.19)	0.55 (0.50)	0.000	2.128
	FR (γ_H)	2.71 (0.70)	0.43 (0.50)	0.00 (0.00)	0.001	0.605
	STEPWISE	4.54 (0.52)	1.77 (2.01)	0.55 (0.50)	0.000	2.132
4 ($p_0 = 14$)	LASSO(1SE)	10.01 (1.73)	3.91 (6.03)	0.01 (0.10)	0.003	15.582
	LASSO(BIC)	12.11 (1.46)	36.56 (22.43)	0.19 (0.39)	0.002	5.688
	SIS+LASSO(1SE)	10.42 (1.66)	21.41 (8.87)	0.03 (0.17)	0.003	11.316
	SIS+LASSO(BIC)	10.73 (1.66)	32.67 (8.92)	0.03 (0.17)	0.003	8.545
	dgLARS(BIC)	12.05 (1.52)	38.70 (28.97)	0.18 (0.38)	0.002	5.111
	SC (γ_L)	10.33 (1.63)	10.48 (6.66)	0.02 (0.14)	0.002	4.499
	SC (γ_H)	5.32 (1.92)	0.52 (1.37)	0.00 (0.00)	0.003	14.005
	FR (γ_L)	12.00 (1.71)	8.93 (6.36)	0.23 (0.42)	0.001	4.503
	FR (γ_H)	5.65 (2.13)	0.38 (1.15)	0.00 (0.00)	0.003	13.802
	STEPWISE	11.80 (1.72)	5.97 (5.37)	0.19 (0.39)	0.001	5.809
5 ($p_0 = 3$)	LASSO(1SE)	2.00 (0.00)	1.13 (2.85)	0.00 (0.00)	0.003	2.674
	LASSO(BIC)	2.01 (0.10)	2.82 (2.52)	0.01 (0.10)	0.003	2.583
	SIS+LASSO(1SE)	2.87 (0.34)	9.28 (3.85)	0.87 (0.34)	0.002	2.455
	SIS+LASSO(BIC)	2.87 (0.34)	9.88 (4.29)	0.87 (0.34)	0.002	2.355
	dgLARS(BIC)	2.00 (0.00)	2.88 (2.38)	0.00 (0.00)	0.003	2.562
	SC (γ_L)	2.75 (0.44)	3.27 (1.75)	0.75 (0.44)	0.001	2.339
	SC (γ_H)	2.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.003	1.086
	FR (γ_L)	3.00 (0.00)	2.80 (1.73)	1.00 (0.00)	0.001	2.326
	FR (γ_H)	2.40 (0.49)	0.00 (0.00)	0.40 (0.49)	0.002	0.981
	STEPWISE	3.00 (0.00)	0.35 (0.59)	1.00 (0.00)	0.001	2.977

Note: abbreviations are explained in the footnote of Table 2.

5. Real Data Analysis

5.1. A Study of Gene Regulation in the Mammalian Eye

To demonstrate the utility of our proposed method, we analyzed a microarray dataset from [35] with 120 twelve-week male rats selected for eye tissue harvesting. The dataset contained more than 31,042 different probe sets (Affymetric GeneChip Rat Genome 230 2.0 Array); see [35] for a more detailed description of the data.

Although our method was applicable to the original 31,042 probe sets, many probes turned out to have very small variances and were unlikely to be informative for correlative analyses. Therefore, using variance as the screening criterion, we selected 5000 genes with the largest variances in expressions and

correlated them with gene *TRIM32* that has been found to cause Bardet–Biedl syndrome, a genetically heterogeneous disease of multiple organ systems including the retina [36].

We applied the proposed STEPWISE method to the dataset with $n = 120$ and $p = 5000$, and treated the *TRIM32* gene expression as the response variable and the expressions of 5000 genes as the predictors. With no prior biological information available, we started with the empty set. To choose η_1 and η_2, we carried out 5-fold cross-validation to minimize the mean squared prediction error (MSPE) by using the following grid search: $\eta_1 = \{0, 0.25, 0.5, 1\}$ and $\eta_2 = \{1, 2, 3, 4, 5\}$, and set $\eta_1 = 1$ and $\eta_2 = 4$. We also performed the same procedure to choose the γ for FR and SC. The regularization parameters in LASSO and dgLARS were selected to minimize BIC values.

In the forward step, STEPWISE selected the probes of *1376747_at*, *1381902_at*, *1382673_at* and *1375577_at*, and the backward step eliminated probe *1375577_at*. The STEPWISE procedure produced the following final predictive model:

TRIM32 = 4.6208 + 0.2310 × (*1376747_at*) + 0.1914 × (*1381902_at*) + 0.1263 × (*1382673_at*). Table A1 in Appendix B presents the numbers of overlapping genes among competing methods. It shows that the two out of three probes, *1381902_at* and *1376747_at*, selected from our method are also discovered by the other methods, except for dgLARS.

Next, we performed Leave-One-Out Cross-Validation (LOOCV) to obtain the distribution of the model size (MS) and MSPE for the competing methods.

As reported in Table 5 and Figure 1, LASSO, SIS+LASSO and dgLARS tended to select more variables than the forward approaches and STEPWISE. Among all of the methods, STEPWISE selected the fewest variables but with almost the same MSPE as the other methods.

Table 5. Comparisons of MSPE among competing methods using the mammalian eye data set.

	STEPWISE	FR	LASSO	SIS+LASSO	SC	dgLARS
Training set	0.005	0.005	0.005	0.006	0.005	0.014
Testing set	0.011	0.012	0.010	0.009	0.014	0.020

Note: The mean squared prediction error (MSPE) was averaged over 120 splits. LASSO, least absolute shrinkage and selection operator with regularization parameter that gives the smallest BIC; SIS+LASSO, sure independence screening by [5] followed by LASSO; dgLARS, differential geometric least angle regression by [11,12] that gives the smallest BIC; SC(γ), sequentially conditioning approach by [8]; FR(γ), forward regression by [7]; STEPWISE, the proposed method. STEPWISE was performed with $\eta_1 = 1$ and $\eta_2 = 4$; FR and SC were performed with $\gamma = 1$.

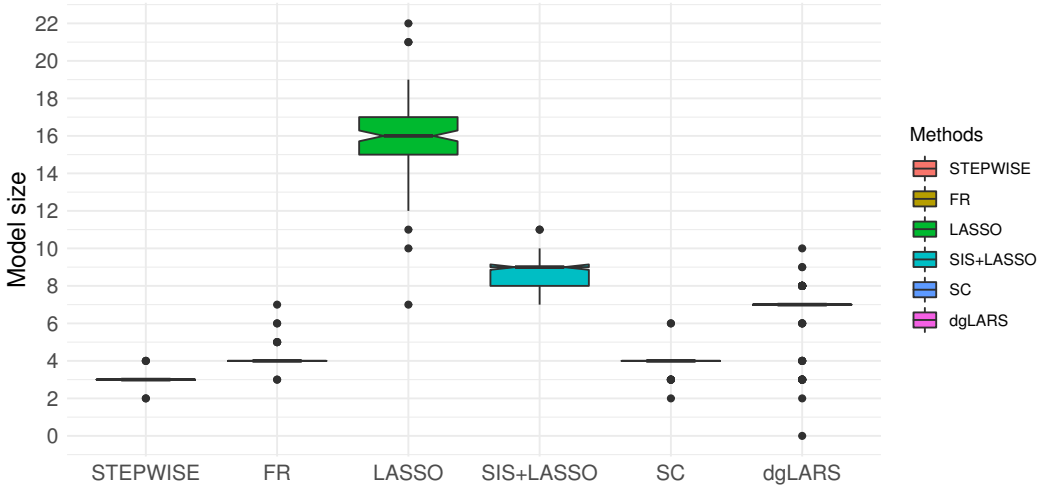


Figure 1. Box plot of model sizes for each method over 120 different training samples from the mammalian eye data set. STEPWISE was performed with $\eta_1 = 1$ and $\eta_2 = 4$, and FR and SC were conducted with $\gamma = 1$.

5.2. An Esophageal Squamous Cell Carcinoma Study

Esophageal squamous cell carcinoma (ESCC), the most common histological type of esophageal cancer, is known to be associated with poor overall survival, making early diagnosis crucial for treatment and disease management [37]. Several studies have investigated the roles of circulating microRNAs (miRNAs) in diagnosis of ESCC [38].

Using a clinical study that investigated the roles of miRNAs on the ESCC [39], we aimed to use miRNAs to predict ESCC risks and estimate their impacts on the development of ESCC. Specifically, with a dataset of serum profiling of 2565 miRNAs from 566 ESCC patients and 4965 controls without cancer, we demonstrated the utility of the proposed STEPWISE method in predicting ESCC with miRNAs.

To proceed, we used a balance sampling scheme (283 cases and 283 controls) in the training dataset. The design of yielding an equal number of cases and controls in the training set has proved to be useful [39] for handling imbalanced outcomes as we encountered here. To validate our findings, samples were randomly divided into a training ($n_1 = 566$, $p = 2565$) and testing set ($n_2 = 4965$, $p = 2565$).

The training set consisted of 283 patients with ESCC (median age of 65 years, 79% male) and 283 control patients (median age of 68 years, 46.3% male), and the testing set consisted of 283 patients with ESCC (median age of 67 years, 85.7% male) and 4682 control patients (median age of 67.5 years, 44.5% male). Control patients without ESCC came from three sources: 323 individuals from National Cancer Center Biobank (NCCB); 2670 individuals from the Biobank of the National Center for Geriatrics and Gerontology (NCGG); and 1972 individuals from Minoru Clinic (MC). More detailed characteristics of cases and controls in the training and testing sets are given in Table 6.

Table 6. Clinicopathological characteristics of study participants of the ESCC data set.

Covariates	Training Set n_1 (%)	Testing set n_2 (%)
Esophageal squamous cell carcinoma (ESCC) patients		
Total number of patients	283	283
Age, median (range)	65 [40, 86]	67 [37, 90]
Gender:		
Male	224 (79.0%)	247 (87.3%)
Female	59 (21.0%)	36 (12.7%)
Stage:		
0	24 (8.5%)	27 (9.5%)
1	127 (44.9%)	128 (45.2%)
2	58 (20.5%)	57 (20.1%)
3	67 (23.7%)	61 (21.6%)
4	7 (2.4%)	10 (3.6%)
Non-ESCC Controls		
Total number of patients	283	4,682
Age, median (range)	68 [27, 92]	67.5 [20, 100]
Gender:		
Male	131 (46.3%)	2,086 (44.5%)
Female	152 (53.7%)	2,596 (55.5%)
Data sources of the controls:		
National Cancer Center Biobank (NCCB)	17 (6.0%)	306 (6.5%)
National Center for Geriatrics and Gerontology (NCGG)	158 (55.8%)	2,512 (53.7%)
Minoru clinic (MC)	108 (38.2%)	1,864 (39.8%)

We defined the binary outcome variable to be 1 if the subject was a case and 0 otherwise. As age and gender (0 = female, 1 = male) are important risk factors for ESCC [40,41] and it is common to adjust for them in clinical models, we set the initial set in STEPWISE to be $F_0 = \{\text{age, gender}\}$. With $\eta_1 = 0$ and $\eta_2 = 3.5$ that were also chosen from 5-fold CV, our procedure recruited three miRNAs. More

specifically, *miR-4783-3p*, *miR-320b*, *miR-1225-3p* and *miR-6789-5p* were selected among 2565 miRNAs by the forward stage from the training set, and then the backward stage eliminated *miR-6789-5p*.

In comparison, with $\gamma = 0$, both FR and SC selected four miRNAs, *miR-4783-3p*, *miR-320b*, *miR-1225-3p* and *miR-6789-5p*. The list of selected miRNAs by different methods are given in Table A2 in Appendix B.

Our findings were biologically meaningful, as the selected miRNAs had been identified by other cancer studies as well. Specifically, *miR-320b* was found to promote colorectal cancer proliferation and invasion by competing with its homologous *miR-320a* [42]. In addition, serum levels of *miR-320* family members were associated with clinical parameters and diagnosis in prostate cancer patients [43]. Reference [44] showed that *miR-4783-3p* was one of the miRNAs that could increase the risk of colorectal cancer death among rectal cancer cases. Finally, *miR-1225-5p* inhibited proliferation and metastasis of gastric carcinoma through repressing insulin receptor substrate-1 and activation of β-catenin signaling [45].

Aiming to identify a final model without resorting to a pre-screening procedure that may miss out on important biomarkers, we applied STEPWISE to reach the following predictive model for ESCC based on patients' demographics and miRNAs:

$\text{logit}^{-1}(-35.70 + 1.41 \times \textit{miR-4783-3p} + 0.98 \times \textit{miR-320b} + 1.91 \times \textit{miR-1225-3p} + 0.10 \times \textit{Age} - 2.02 \times \textit{Gender})$, where $\text{logit}^{-1}(x) = \exp(x)/(1 + \exp(x))$.

In the testing dataset, the model had an area under the receiver operating curve (AUC) of 0.99 and achieved a high accuracy of 0.96, with a sensitivity and specificity of 0.97 and 0.95, respectively. Additionally, using the testing cohort, we evaluated the performances of the models sequentially selected by STEPWISE. Starting with a model containing age and gender, STEPWISE selected *miR-4783-3p*, *miR-320b* and *miR-1225-3p* in turn. Figure 2, showing the corresponding receiver operating curves (ROC) for these sequential models, revealed the improvement by sequentially adding predictors to the model and justified the importance of these variables in the final model. In addition, Figure 2e illustrated that adding an extra miRNA selected by FR and SC made little improvement of the model's predictive power.

Furthermore, we conducted subgroup analysis within the testing cohort to study how the sensitivity of the final model differed by cancer stage, one of the most important risk factors. The sensitivities for stages 0, i.e., non-invasive cancer, 9 ($n = 27$), 1 ($n = 128$), 2 ($n = 57$), 3 ($n = 61$) and 4 ($n = 10$) were 1.00, 0.98, 0.97, 0.97 and 1.00, respectively. We next evaluated how the specificity varied across controls coming from different data sources. The specificities for the various control groups, namely, NCCB ($n = 306$), NCGG ($n = 2512$) and MC ($n = 1864$), were 0.99, 0.99 and 0.98, respectively. The results indicated the robust performance of the miRNA-based model toward cancer stages and data sources.

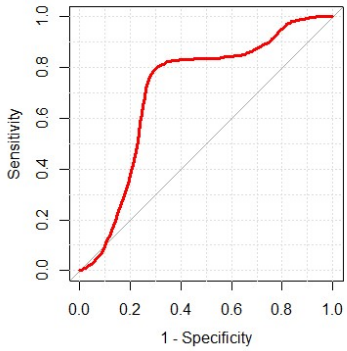
Finally, to compare STEPWISE with the other competing methods, we repeatedly applied the aforementioned balance sampling procedure and split the ESCC data into the training and testing sets 100 times. We obtained MSPE and the average of accuracy, sensitivity, specificity, and AUC. Figure 3 reported the model size of each method. Though STEPWISE selected fewer variables compared to the other variable selection methods (for example, LASSO selected 11-31 variables and dgLARS selected 12–51 variables), it achieved comparable prediction accuracy, specificity, sensitivity and AUC (see Table 7), evidencing the utility of STEPWISE for generating parsimonious models while maintaining competitive predictability.

Table 7. Comparisons of competing methods over 100 independent splits of the ESCC data into training and testing sets.

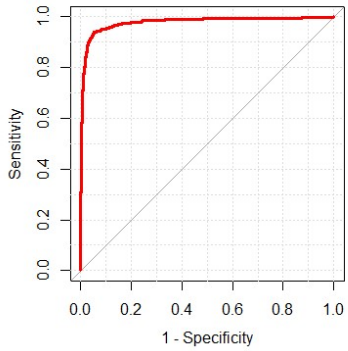
Training Set	MSPE	Accuracy	Sensitivity	Specificity	AUC
STEPWISE	0.02	0.97	0.98	0.97	1.00
SC	0.01	0.99	0.98	0.98	1.00
FR	0.02	0.99	0.97	0.97	1.00
LASSO	0.01	0.98	1.00	0.97	1.00
SIS+LASSO	0.01	0.99	1.00	0.99	1.00
dgLARS	0.04	0.96	0.99	0.94	1.00
Training Set	**MSPE**	**Accuracy**	**Sensitivity**	**Specificity**	**AUC**
STEPWISE	0.04	0.96	0.97	0.95	0.99
SC	0.03	0.96	0.97	0.96	0.99
FR	0.04	0.96	0.97	0.95	0.99
LASSO	0.03	0.96	0.99	0.95	1.00
SIS+LASSO	0.02	0.97	0.99	0.96	1.00
dgLARS	0.05	0.94	0.98	0.94	1.00

Note: Values were averaged over 100 splits. STEPWISE was performed with $\eta_1 = 0$ and $\eta_2 = 1$. SC and FR were performed with $\gamma = 1$. The regularization parameters in LASSO and dgLARS were selected to minimize the BIC.

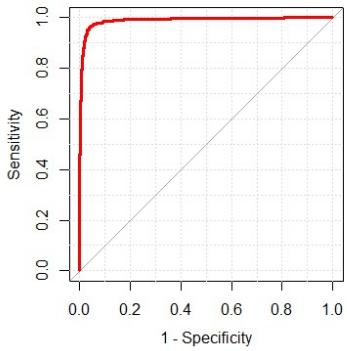
We used `R` software [46] to obtain the numerical results in Sections 4 and 5 with following packages: `ggplot2` [47], `ncvreg` [32], `glmnet` [31], `dglars` [34] and `screening` [33].



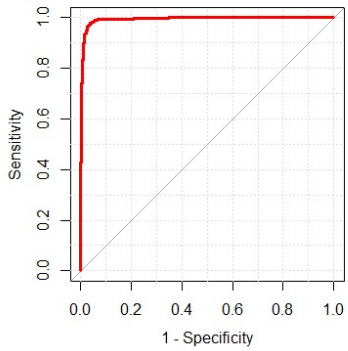
(a) Model 1, AUC = 0.71



(b) Model 2, AUC = 0.97

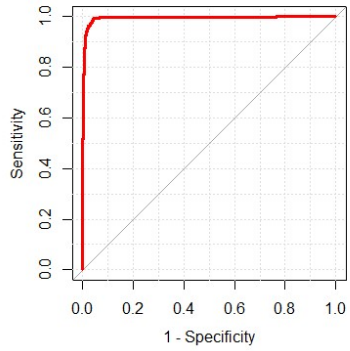


(c) Model 3, AUC = 0.98



(d) Model 4, AUC = 0.99

Figure 2. *Cont.*



(e) Model 5, AUC = 0.99

Figure 2. Comparisons of ROC curves for the selected models in the ESCC data set by the sequentially selected order: Model 1: $-2.52 + 0.02 \times$ *Age* $- 1.86 \times$ *Gender*; Model 2: $-20.64 + 0.08 \times$ *Age* $- 2.12 \times$ *Gender* $+ 2.02 \times$ *miR-4783-3p*; Model 3: $-24.21 + 0.09 \times$ *Age* $- 2.16 \times$ *Gender* $+ 1.44 \times$ *miR-4783-3p* $-1.31 \times$ *miR-320b*; Model 4: $-35.70 + 0.10 \times$ *Age* $- 2.02 \times$ *Gender* $+ 1.40 \times$ *miR-4783-3p* $-0.98 \times$ *miR-320b* $+1.91 \times$ *miR-1225-3p*; Model 5: $-53.10 + 0.10 \times$ *Age* $- 1.85 \times$ *Gender* $+ 1.43 \times$ *miR-4783-3p* $-0.92 \times$ *miR-320b* $+1.43 \times$ *miR-1225-3p* $+2.10 \times$ *miR-6789-5p*.

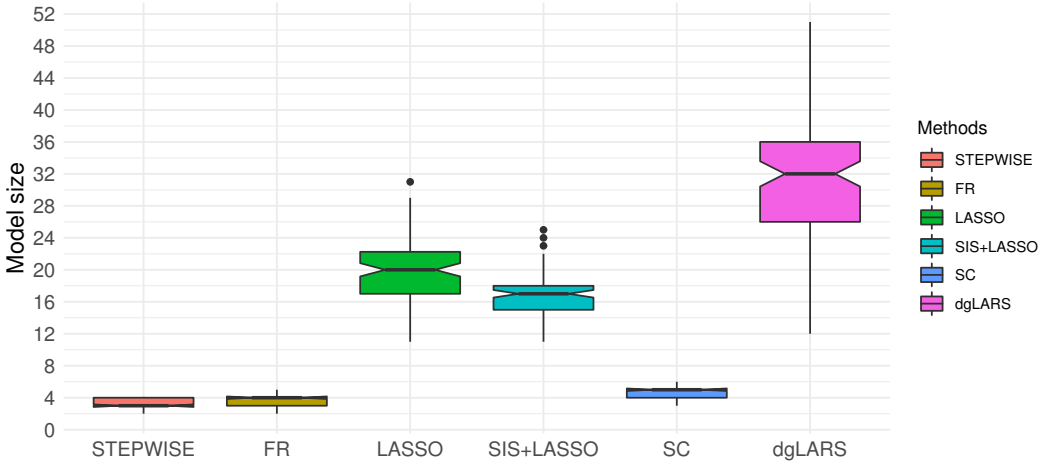


Figure 3. Box plot of model sizes for each method based on 100 ESCC training datasets. Performance of STEPWISE is reported with $\eta_1 = 0$ and $\eta_2 = 3.5$. Performances of SC and FR are reported with $\gamma = 0$.

6. Discussion

We have proposed to apply STEPWISE to produce final models in ultrahigh dimensional settings, without resorting to a pre-screening step. We have shown that the method identifies or includes the true model with probability going to 1, and produces consistent coefficient estimates, which are useful for properly interpreting the actual impacts of risk factors. The theoretical properties of STEPWISE were established under mild conditions, which are worth discussing. As in practice covariates are often standardized for various reasons, Condition (2) is assumed without loss of generality. Conditions (3) and (4) are generally satisfied under common GLM models, including Gaussian, binomial, Poisson and gamma distributions. Condition (5) is also often satisfied in practice. Proposition 2 in [26] may be used as a tool to verify Condition (5) as well. Conditions (1) and (6) are in good faith with the unknown true model size $|\mathcal{M}|$ and minimum signal strength $n^{-\alpha}$ in practice. The "irrepresentable" condition (6) is strong and may not hold in some real datasets, see, e.g., [48,49]. However, the condition holds under some commonly used covariance structures, including AR(1) and compound symmetry structure [48].

As shown in simulation studies and real data analyses, STEPWISE tends to generate models as predictive as the other well-known methods, with fewer variables (Figure 3). Parsimonious models

are useful for biomedical studies as they explain data with a small number of important predictors, and offer practitioners a realistic list of biomarkers to investigate. With categorical outcome data frequently observed in biomedical studies (e.g., histology types of cancer), STEPWISE can be extended to accommodate multinomial classification, with more involved notation and computation. We will pursue this elsewhere.

There are several open questions. First, our final model was determined by using (E)BIC, which involves two extra parameters η_1 and η_2. In our numerical experiments, we used cross-validation to choose them, which seemed to work well. However, more in-depth research is needed to find their optimal values to strike a balance between false positives and false negatives. Second, despite our consistent estimates, drawing inferences based on them remains challenging. Statistical inference, which accounts for uncertainty in estimation, is key for properly interpreting analysis results and drawing appropriate conclusions. Our asymptotic results, nevertheless, are a stepping stone toward this important problem.

Supplementary Materials: An R package, STEPWISE, was developed and is available at https://github.com/AlexPijyan/STEPWISE, along with the examples shown in the paper.

Author Contributions: Conceptualization, Q.Z., H.H. and Y.L.; Formal analysis, A.P.; Methodology, A.P., Q.Z., H.H. and Y.L.; Project administration, H.H.; Software, A.P.; Supervision, H.H.; Writing – original draft, Q.Z., H.H. and Y.L.; Writing – review & editing, H.H. and Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded partially by grants from NSF (DMS1915099, DMS1952486) and NIH (R01AG056764, U01CA209414, R03AG067611).

Acknowledgments: We are thankful to the Editor, the AE and two referees for insightful suggestions that helped improve the manuscript.

Conflicts of Interest: 'The authors declare no conflict of interest.

Appendix A. Proofs of Main Theorems

Since $b(\cdot)$ is twice continuously differentiable with a nonnegative second derivative $b''(\cdot)$, $b_{\max} := \max_{|t|\le K^3} |b(t)|$, $\mu_{\max} := \max_{|t|\le K^3} |b'(t)|$ and $\sigma_{\max} := \sup_{|t|\le K^3} |b''(t)|$ are bounded above, where L and K are some constants from Conditions (1) and (2), respectively. Let $\mathbb{G}_n\{f(\xi)\} = n^{-1/2}\sum_{i=1}^n (f(\xi_i) - E[f(\xi_i)])$ for a sequence of i.i.d. random variables ξ_i $(i = 1, \ldots, n)$ and a non-random function $f(\cdot)$.

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

$$\ell_{S\cup\{r\}}(\boldsymbol{\beta}_{S+r}) := \mathbb{E}_n\left\{L\left(\boldsymbol{\beta}_S^{\mathrm{T}}\mathbf{X}_S + \beta_r X_r, Y\right)\right\}. \tag{A1}$$

We use $\hat{\boldsymbol{\beta}}_{S+r}$ to denote the maximizer of (A1). Thus, $\hat{\boldsymbol{\beta}}_{S+r} = \hat{\boldsymbol{\beta}}_{S\cup\{r\}}$. In addition, denote the maximizer of $E[\ell_{S\cup\{r\}}(\boldsymbol{\beta}_{S+r})]$ by $\boldsymbol{\beta}^*_{S+r}$. Due to the concavity of the log-likelihood in GLMs with the canonical link, $\boldsymbol{\beta}^*_{S+r}$ is unique.

Proof of Theorem 1. Given an index set S and $r \in S^c$, let $\mathcal{B}_S^0(d) = \{\boldsymbol{\beta}_S : \|\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*\| \le d/(K\sqrt{|S|})\}$ where $d = A_2\sqrt{q^3 \log p/n}$ with A_2 defined in Lemma A6.

Let Ω be the event that

$$\Big\{\sup_{|S|\le q, \boldsymbol{\beta}_S\in\mathcal{B}_S^0(d)} \left|\mathbb{G}_n\left[L\left(\boldsymbol{\beta}_S^{\mathrm{T}}\mathbf{X}_S, Y\right) - L\left(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S, Y\right)\right]\right| \le 20A_1 d\sqrt{q\log p} \quad \text{and}$$
$$\max_{|S|\le q} |\mathbb{G}_n\left[L(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S, Y)\right]| \le 10(A_1K^2 + b_{\max})\sqrt{q\log p}\Big\},$$

where A_1 is some constant defined in Lemma A4. By Lemma A4, $P(\Omega) \ge 1 - 6\exp(-6q\log p)$. Thus in the rest of the proof, we only consider the sample points in Ω.

In the proof of Lemma A6, we show that $\max_{|S|\le q}\|\hat{\boldsymbol{\beta}}_S-\boldsymbol{\beta}^*_S\|\le A_2K^{-1}(q^2\log p/n)^{1/2}$ under Ω. Then given an index set S and $\boldsymbol{\beta}_S$ such that $|S|<q$, $\|\boldsymbol{\beta}_S-\boldsymbol{\beta}^*_S\|\le A_2K^{-1}(q^2\log p/n)^{1/2}$, and for any $j\in S^c$,

$$\begin{aligned}
&\ell_{S\cup\{j\}}(\boldsymbol{\beta}^*_{S+j})-\ell_S(\hat{\boldsymbol{\beta}}_S)\ge \inf_{\|\boldsymbol{\beta}_S-\boldsymbol{\beta}^*_S\|\le A_2K^{-1}(q^2\log p/n)^{1/2}}\ell_{S\cup\{j\}}(\boldsymbol{\beta}^*_{S+j})-\ell_S(\boldsymbol{\beta}_S)\\
&=n^{-1/2}\mathbb{G}_n\left[L(\boldsymbol{\beta}^{*\mathrm{T}}_{S+j}\mathbf{X}_{S\cup\{j\}},Y)\right]-n^{-1/2}\mathbb{G}_n\left[L(\boldsymbol{\beta}^{*\mathrm{T}}_S\mathbf{X}_S,Y)\right]\\
&\quad-\sup_{\|\boldsymbol{\beta}_S-\boldsymbol{\beta}^*_S\|\le A_2K^{-1}(q^2\log p/n)^{1/2}}\left|n^{-1/2}\mathbb{G}_n\left[L(\boldsymbol{\beta}^{\mathrm{T}}_S\mathbf{X}_S,Y)-L(\boldsymbol{\beta}^{*\mathrm{T}}_S\mathbf{X}_S,Y)\right]\right|\\
&\quad+E\left[L(\boldsymbol{\beta}^{*\mathrm{T}}_{S+j}\mathbf{X}_{S\cup\{j\}},Y)\right]-E\left[L(\boldsymbol{\beta}^{*\mathrm{T}}_S\mathbf{X}_S,Y)\right]\\
&\ge-20(A_1K^2+b_{\max})\sqrt{q\log p/n}-20A_1A_2q^2\log p/n+\frac{\sigma_{\min}\kappa_{\min}}{2}\|\boldsymbol{\beta}^*_{S+j}-(\boldsymbol{\beta}^{*\mathrm{T}}_S,0)^{\mathrm{T}}\|^2,
\end{aligned}$$

where the second inequality follows from the event Ω and Lemma A5.

By Lemma A1, if $\mathcal{M}\not\subseteq S$, there exists $r\in S^c\cap\mathcal{M}$, such that $\|\boldsymbol{\beta}^{*\mathrm{T}}_{S+r}-(\boldsymbol{\beta}^{*\mathrm{T}}_S,0)\|\ge C\sigma^{-1}_{\max}\kappa^{-1}_{\max}n^{-\alpha}$. Thus, there exists some constant C_1 that does not depend on n such that

$$\begin{aligned}
&\max_{j\in S^c}\ell_{S\cup\{j\}}(\hat{\boldsymbol{\beta}}_{S+j})-\ell_S(\hat{\boldsymbol{\beta}}_S)\ge\max_{j\in S^c}\ell_{S\cup\{j\}}(\boldsymbol{\beta}^*_{S+j})-\ell_S(\hat{\boldsymbol{\beta}}_S)\ge\ell_{S\cup\{r\}}(\boldsymbol{\beta}^*_{S+r})-\ell_S(\hat{\boldsymbol{\beta}}_S)\\
&\ge-20(A_1K^2+b_{\max})\sqrt{q\log p/n}-20A_1A_2q^2\log p/n+\frac{C^2\sigma_{\min}\kappa_{\min}n^{-2\alpha}}{2\sigma^2_{\max}\kappa^2_{\max}}\ge C_1n^{-2\alpha},\quad\text{(A2)}
\end{aligned}$$

where the first inequality follows from $\hat{\boldsymbol{\beta}}_{S+j}$ being the maximizer of (A1) and the second inequality follows from Conditions (1) and (6).

Withdrawing the restriction to Ω, we obtain that

$$P\left(\min_{|S|<q,\mathcal{M}\not\subseteq S}\max_{j\in S^c}\ell_{S\cup\{j\}}(\hat{\boldsymbol{\beta}}_{S\cup\{j\}})-\ell_S(\hat{\boldsymbol{\beta}}_S)\ge C_1n^{-2\alpha}\right)\ge1-6\exp(-6q\log p).$$

□

Proof of Theorem 2. We have shown that our forward stage will not stop when $\mathcal{M}\not\subseteq S$ and $|S|<q$ with probability converging to 1.

For any $r\in S^c\cap\mathcal{M}^c$, $\boldsymbol{\beta}^*_{S+r}$ is the unique solution to the equation $E\left[\{Y-\mu(\boldsymbol{\beta}^{\mathrm{T}}_{S+r}\mathbf{X}_{S\cup\{r\}})\}\mathbf{X}_{S\cup\{r\}}\right]=\mathbf{0}$. By the mean value theorem,

$$\begin{aligned}
&E\left[\{Y-\mu(\boldsymbol{\beta}^{*\mathrm{T}}_S\mathbf{X}_S)\}X_r\right]=E\left[\{\mu(\boldsymbol{\beta}^{\mathrm{T}}_*\mathbf{X})-\mu(\boldsymbol{\beta}^{*\mathrm{T}}_S\mathbf{X}_S)\}X_r\right]\\
&=E\left[\{\mu(\boldsymbol{\beta}^{\mathrm{T}}_*\mathbf{X})-\mu(\boldsymbol{\beta}^{*\mathrm{T}}_S\mathbf{X}_S)\}X_r\right]-E\left[\{\mu(\boldsymbol{\beta}^{\mathrm{T}}_*\mathbf{X})-\mu(\boldsymbol{\beta}^{*\mathrm{T}}_{S+r}\mathbf{X}_{S\cup\{r\}})\}X_r\right]\\
&=(\boldsymbol{\beta}^{*\mathrm{T}}_{S+r}-(\boldsymbol{\beta}^{*\mathrm{T}}_S,0))E\left[\sigma(\tilde{\boldsymbol{\beta}}^{\mathrm{T}}_{S+r}\mathbf{X}_{S\cup\{r\}})\mathbf{X}^{\otimes2}_{S\cup\{r\}}\right]\mathbf{e}_r,
\end{aligned}$$

where $\tilde{\boldsymbol{\beta}}_{S+r}$ is some point between $\boldsymbol{\beta}_{S+r}$ and $(\boldsymbol{\beta}^{*\mathrm{T}}_S,0)^{\mathrm{T}}$ and $\mathbf{e}_r$ is a vector of length $(|S|+1)$ with the rth element being 1.

Since $|\tilde{\boldsymbol{\beta}}^{\mathrm{T}}_{S+r}\mathbf{X}_{S\cup\{r\}}|\le|\boldsymbol{\beta}^{*\mathrm{T}}_{S+r}\mathbf{X}_{S\cup\{r\}}|+|(\boldsymbol{\beta}^{*\mathrm{T}}_S,0)\mathbf{X}_{S\cup\{r\}}|\le2K^2$ by Conditions (1) and (2), $|\sigma(\tilde{\boldsymbol{\beta}}^{\mathrm{T}}_{S+r}\mathbf{X}_{S\cup\{r\}})|\ge\sigma_{\min}$ and

$$o(n^{-\alpha})=\left|E\left[\left\{Y-\mu(\boldsymbol{\beta}^{*\mathrm{T}}_S\mathbf{X}_S)\right\}X_r\right]\right|\ge\sigma_{\min}\kappa_{\min}\|\boldsymbol{\beta}^{*\mathrm{T}}_{S+r}-(\boldsymbol{\beta}^{*\mathrm{T}}_S,0)\|.$$

Therefore, $\max_{S:|S|\le q,r\in S^c\cap\mathcal{M}^c}\|\boldsymbol{\beta}^{*\mathrm{T}}_{S+r}-(\boldsymbol{\beta}^{*\mathrm{T}}_S,0)\|=o(n^{-\alpha})$.

Under Ω that is defined in Theorem 1, $\max_{|S|\le q}\|\hat{\boldsymbol{\beta}}_S-\boldsymbol{\beta}^*_S\|\le A_2K^{-1}(q^2\log p/n)^{1/2}$. For any $j\in S^c$,

$$\begin{aligned}
&\ell_{S\cup\{j\}}(\boldsymbol{\beta}^*_{S+j})-\ell_S(\hat{\boldsymbol{\beta}}_S)\le \sup_{\|\boldsymbol{\beta}_S-\boldsymbol{\beta}^*_S\|\le A_2K^{-1}(q^2\log p/n)^{1/2}}\ell_{S\cup\{j\}}(\boldsymbol{\beta}^*_{S+j})-\ell_S(\boldsymbol{\beta}_S)\\
&\le\left|n^{-1/2}\mathbb{G}_n\left[L(\boldsymbol{\beta}^{*\mathrm{T}}_{S+j}\mathbf{X}_{S\cup\{j\}},Y)\right]\right|+\left|n^{-1/2}\mathbb{G}_n\left[L(\boldsymbol{\beta}^{*\mathrm{T}}_S\mathbf{X}_S,Y)\right]\right|\\
&\quad+\sup_{\|\boldsymbol{\beta}_S-\boldsymbol{\beta}^*_S\|\le A_2K^{-1}(q^2\log p/n)^{1/2}}\left|n^{-1/2}\mathbb{G}_n\left[L(\boldsymbol{\beta}^{\mathrm{T}}_S\mathbf{X}_S,Y)-L(\boldsymbol{\beta}^{*\mathrm{T}}_S\mathbf{X}_S,Y)\right]\right|\\
&\quad+\left|E\left[L(\boldsymbol{\beta}^{*\mathrm{T}}_{S+j}\mathbf{X}_{S\cup\{j\}},Y)\right]-E\left[L(\boldsymbol{\beta}^{*\mathrm{T}}_S\mathbf{X}_S,Y)\right]\right|\\
&\le 20(A_1K^2+b_{\max})\sqrt{qn^{-1}\log p}+20A_1A_2q^2n^{-1}\log p+\sigma_{\max}\kappa_{\max}\|\boldsymbol{\beta}^*_{S+j}-(\boldsymbol{\beta}^{*\mathrm{T}}_S,0)^{\mathrm{T}}\|^2/2,
\end{aligned}$$

where the second inequality follows from the event Ω and Lemma A5. Since $\max_{S:|S|<q,r\in S^c\cap\mathcal{M}^c}\|\boldsymbol{\beta}^*_{S+r}-(\boldsymbol{\beta}^{*\mathrm{T}}_S,0)^{\mathrm{T}}\|=o(n^{-\alpha})$ and $qn^{-1+4\alpha}\log p\to 0$,

$$\begin{aligned}
\max_{S:|S|<q,r\in S^c\cap\mathcal{M}^c}\ell_{S\cup\{r\}}(\boldsymbol{\beta}^*_{S+r})-\ell_S(\hat{\boldsymbol{\beta}}_S)&\le 20(A_1K^2+b_{\max})\sqrt{qn^{-1}\log p}+20A_1A_2q^2n^{-1}\log p\\
&\quad+\sigma_{\max}\kappa_{\max}\|\boldsymbol{\beta}^*_{S+j}-(\boldsymbol{\beta}^{*\mathrm{T}}_S,0)^{\mathrm{T}}\|^2/2=o(n^{-2\alpha}),
\end{aligned}$$

with probability at least $1-6\exp(-6q\log p)$. Then by Lemma A6,

$$\begin{aligned}
&\max_{S:|S|<q,r\in S^c\cap\mathcal{M}^c}\ell_{S\cup\{r\}}(\hat{\boldsymbol{\beta}}_{S+r})-\ell_S(\hat{\boldsymbol{\beta}}_S)\\
&\le\max_{S:|S|<q,r\in S^c\cap\mathcal{M}^c}|\ell_{S\cup\{r\}}(\hat{\boldsymbol{\beta}}_{S+r})-\ell_{S\cup\{r\}}(\boldsymbol{\beta}^*_{S+r})|+\max_{S:|S|<q,r\in S^c\cap\mathcal{M}^c}\left|\ell_{S\cup\{r\}}(\boldsymbol{\beta}^*_{S+r})-\ell_S(\hat{\boldsymbol{\beta}}_S)\right|\\
&\le A_3q^2n^{-1}\log p+o(n^{-2\alpha})=o(n^{-2\alpha}),
\end{aligned}\tag{A3}$$

with probability at least $1-12\exp(-6q\log p)$.

By Theorem 1, if $\mathcal{M}\not\subseteq S$, the forward stage would select a noise variable with probability less than $18\exp(-6q\log p)$.

For $k>|\mathcal{M}|$, $\mathcal{M}\not\subseteq S_k$ implies that at least $k-|\mathcal{M}|$ noise variables are selected within the k steps. Then for $k=C_2|\mathcal{M}|$ with $C_2>2$,

$$\begin{aligned}
&P(\mathcal{M}\not\subseteq S_k)\le\sum_{j=k-|\mathcal{M}|}^{k}\binom{k}{j}\{18\exp(-6q\log p)\}^j\le|\mathcal{M}|k^{|\mathcal{M}|}\{18\exp(-6q\log p)\}^{k-|\mathcal{M}|}\\
&\le 18\exp(-6q\log p+\log|\mathcal{M}|+|\mathcal{M}|\log k)\le 18\exp(-4q\log p).
\end{aligned}$$

Therefore, $\mathcal{M}\subset S_{C_2|\mathcal{M}|}$ with probability at least $1-18\exp(-4q\log p)$. □

Proof of Theorem 3. By Theorem 2, $\mathcal{M}$ will be included in F_k for some $k<q$ with probability going to 1. Therefore, the forward stage stops at the kth step if $\mathrm{EBIC}(F_{k+1})>\mathrm{EBIC}(F_k)$.

On the other hand, that $\mathrm{EBIC}(F_{k+1})<\mathrm{EBIC}(F_k)$ if and only if $2\ell_{F_{k+1}}(\hat{\boldsymbol{\beta}}_{F_{k+1}})-2\ell_{F_k}(\hat{\boldsymbol{\beta}}_{F_k})\ge(\log n+2\eta_1\log p)/n$. Thus, to show the forward stage stops at the kth step, we only need to show that with probability tending to 1,

$$2\ell_{F_{k+1}}(\hat{\boldsymbol{\beta}}_{F_{k+1}})-2\ell_{F_k}(\hat{\boldsymbol{\beta}}_{F_k})<(\log n+2\eta_1\log p)/n,\tag{A4}$$

for all $\eta_1>0$.

To prove (A4), we first verify the conditions (A4) and (A5) in [17]. Given any index S such that $\mathcal{M} \subseteq S$ and $|S| \leq q$, let $\boldsymbol{\beta}_{*S}$ be the subvector of $\boldsymbol{\beta}_*$ corresponding to S. We obtain that

$$E\left[(Y-\mu(\boldsymbol{\beta}_{*S}^{\mathrm{T}}\mathbf{X}_S))\mathbf{X}_S\right] = E\left[E\left[(Y-\mu(\boldsymbol{\beta}_{*\mathcal{M}}^{T}\mathbf{X}_{\mathcal{M}}))|\mathbf{X}_S\right]\mathbf{X}_S\right] = 0.$$

This implies $\boldsymbol{\beta}_S^* = \boldsymbol{\beta}_{*S}$.

Given any $\boldsymbol{\pi} \in \mathbb{R}^{|S|}$, let $\mathcal{H}_S := \left\{h(\boldsymbol{\pi},\boldsymbol{\beta}_S) = (\sigma_{\max}K^2|S|)^{-1}\sigma\left(\boldsymbol{\beta}_S^{\mathrm{T}}\mathbf{X}_S\right)\left(\boldsymbol{\pi}^{\mathrm{T}}\mathbf{X}_S\right)^2, \|\boldsymbol{\pi}\| = 1, \boldsymbol{\beta}_S \in \mathcal{B}_S^0(d)\right\}$. By Conditions (1) and (2), $h(\boldsymbol{\pi},\boldsymbol{\beta}_S)$ is bounded between -1 and 1 uniformly over $\|\boldsymbol{\pi}\| = 1$ and $\boldsymbol{\beta}_S \in \mathcal{B}_S^0(d)$.

By Lemma 2.6.15 in [50], the VC indices of $\mathcal{W} := \{(K\sqrt{|S|})^{-1}\boldsymbol{\pi}^{\mathrm{T}}\mathbf{X}_S, \|\boldsymbol{\pi}\| = 1\}$ and $\mathcal{V} := \{\boldsymbol{\beta}_S^{\mathrm{T}}\mathbf{X}_S, \boldsymbol{\beta}_S \in \mathcal{B}_S^0(d)\}$ are bounded by $|S|+2$. For the definitions of the VC index and covering numbers, we refer to pages 83 and 85 in [50]. The VC index of the class $\mathcal{U} := \{(K^2|S|)^{-1}(\boldsymbol{\pi}^{\mathrm{T}}\mathbf{X}_S)^2, \|\boldsymbol{\pi}\| = 1\}$ is the VC index of the class of sets $\{(\mathbf{X}_S,t) : (K^2|S|)^{-1}(\boldsymbol{\pi}^{\mathrm{T}}\mathbf{X}_S)^2 \leq t, \|\boldsymbol{\pi}\| = 1, t \in \mathbb{R}\}$. Since $\{(\mathbf{X}_S,t) : (K^2|S|)^{-1}(\boldsymbol{\pi}^{\mathrm{T}}\mathbf{X}_S)^2 \leq t\} = \{(\mathbf{X}_S,t) : 0 < (K\sqrt{|S|})^{-1}\boldsymbol{\pi}^{\mathrm{T}}\mathbf{X}_S \leq \sqrt{t}\} \cup \{(\mathbf{X}_S,t) : -\sqrt{t} < (K\sqrt{|S|})^{-1}\boldsymbol{\pi}^{\mathrm{T}}\mathbf{X}_S \leq 0\}$, each set of $\{(\mathbf{X}_S,t) : (K^2|S|)^{-1}(\boldsymbol{\pi}^{\mathrm{T}}\mathbf{X}_S)^2 \leq t, \|\boldsymbol{\pi}\| = 1, t \in \mathbb{R}\}$ is created by taking finite unions, intersections and complements of the basic sets $\{(\mathbf{X}_S,t) : (K\sqrt{|S|})^{-1}\boldsymbol{\pi}^{\mathrm{T}}\mathbf{X}_S < t\}$. Therefore, the VC index of $\{(\mathbf{X}_S,t) : (K^2|S|)^{-1}(\boldsymbol{\pi}^{\mathrm{T}}\mathbf{X}_S)^2 \leq t, \|\boldsymbol{\pi}\| = 1, t \in \mathbb{R}\}$ is of the same order as the VC index of $\{(\mathbf{X}_S,t) : (K\sqrt{|S|})^{-1}\boldsymbol{\pi}^{\mathrm{T}}\mathbf{X}_S < t\}$, by Lemma 2.6.17 in [50].

Then by Theorem 2.6.7 in [50], for any probability measure Q, there exists some universal constant C_3 such that $N(\epsilon,\mathcal{U},L_2(Q)) \leq (C_3/\epsilon)^{2(|S|+1)}$. Likewise, $N(d\epsilon,\mathcal{V},L_2(Q)) \leq (C_3/\epsilon)^{2(|S|+1)}$. Given a $\boldsymbol{\beta}_{S,0} \in \mathcal{B}_S^0(d)$, for any $\boldsymbol{\beta}_S$ in the ball $\{\boldsymbol{\beta}_S : \sup_{\mathbf{x}}|\boldsymbol{\beta}_S^{\mathrm{T}}\mathbf{x} - \boldsymbol{\beta}_{S,0}^{\mathrm{T}}\mathbf{x}| < d\epsilon\}$, we have $\sup_{\mathbf{x}}|\sigma(\boldsymbol{\beta}_S^{\mathrm{T}}\mathbf{x}) - \sigma(\boldsymbol{\beta}_{S,0}^{\mathrm{T}}\mathbf{x})| < Kd\epsilon$ by Condition (4). Let $\mathcal{V}' := \{\sigma_{\max}^{-1}\sigma(\boldsymbol{\beta}_S^{\mathrm{T}}\mathbf{X}_S), \boldsymbol{\beta}_S \in \mathcal{B}_S^0(d)\}$. By the definition of covering number, $N(Kd\epsilon,\mathcal{V}',L_2(Q)) \leq (C_3/\epsilon)^{2(|S|+1)}$Given a $\sigma(\boldsymbol{\beta}_{S,0}^{\mathrm{T}}\mathbf{x})$ and $\boldsymbol{\pi}_0^{\mathrm{T}}\mathbf{x}$, for any $\sigma(\boldsymbol{\beta}_S^{\mathrm{T}}\mathbf{x})$ in the ball $\{\sigma(\boldsymbol{\beta}_S^{\mathrm{T}}\mathbf{x}) : \sup_{\mathbf{x}}|\sigma(\boldsymbol{\beta}_S^{\mathrm{T}}\mathbf{x}) - \sigma(\boldsymbol{\beta}_{S,0}^{\mathrm{T}}\mathbf{x})| \leq Kd\epsilon\}$ and $\boldsymbol{\pi}$ in the ball $\{\boldsymbol{\pi} : \sup_{\mathbf{x}}|(\boldsymbol{\pi}^{\mathrm{T}}\mathbf{x})^2 - (\boldsymbol{\pi}_0^{\mathrm{T}}\mathbf{x})^2| < \epsilon\}$, $(\sigma_{\max}K^2|S|)^{-1}\sup_{\mathbf{x}}|\sigma(\boldsymbol{\beta}_S^{\mathrm{T}}\mathbf{x})(\boldsymbol{\pi}^{\mathrm{T}}\mathbf{x})^2 - \sigma(\boldsymbol{\beta}_{S,0}^{\mathrm{T}}\mathbf{x})(\boldsymbol{\pi}_0^{\mathrm{T}}\mathbf{x})^2| \leq (\sigma_{\max}^{-1}Kd + (K^2|S|)^{-1})\epsilon$. Thus, $N((\sigma_{\max}^{-1}Kd + (K^2|S|)^{-1})\epsilon,\mathcal{H}_S,L_2(Q)) \leq (C_3/\epsilon)^{4(|S|+1)}$, and consequently $N(\epsilon,\mathcal{H}_S,L_2(Q)) \leq (C_4/\epsilon)^{4(|S|+1)}$ for some constant C_4.

By Theorem 1.1 in [51] and $|S| \leq q$, we can find some constant C_5 such that

$$\begin{aligned}
&P\left(\sup_{\|\boldsymbol{\pi}\|=1,\boldsymbol{\beta}_S\in\mathcal{B}_S^0(d)}|\mathbb{G}_n\left[h(\boldsymbol{\pi},\boldsymbol{\beta}_S)\right]| \geq C_5\sqrt{q\log p}\right)\\
\leq&\frac{C_4'}{C_5\sqrt{q\log p}}\left(\frac{C_4'C_5^2q\log p}{4(|S|+1)}\right)^{4(|S|+1)}\exp(-2C_5^2q\log p)\\
\leq&\exp\left(4(|S|+1)\log(C_4'C_5^2q\log p) - 2C_5^2q\log p\right) \leq \exp\left(-5q\log p\right),
\end{aligned}$$

where C_4' is some constant that depends on C_4 only. Thus,

$$\begin{aligned}
&P\Big(\sup_{|S|\leq q,\|\boldsymbol{\pi}\|=1,\boldsymbol{\beta}_S\in\mathcal{B}_S^0(d)}\left|\mathbb{E}_n\left\{\sigma\left(\mathbf{X}_S^{\mathrm{T}}\boldsymbol{\beta}_S\right)\left(\boldsymbol{\pi}^{\mathrm{T}}\mathbf{X}_S\right)^2\right\} - E\left[\sigma\left(\mathbf{X}_S^{\mathrm{T}}\boldsymbol{\beta}_S\right)\left(\boldsymbol{\pi}^{\mathrm{T}}\mathbf{X}_S\right)^2\right]\right| \geq C_5K^2\sqrt{q^3\log p/n}\Big)\\
\leq&\sum_{s=|\mathcal{M}|}^{q}\left(\frac{ep}{s}\right)^s\exp\left(-5q\log p\right) \leq \exp(-3q\log p). \qquad \text{(A5)}
\end{aligned}$$

By Condition (5), $\sigma_{\min}\kappa_{\min} \leq \Lambda\left(E\left[\sigma\left(\mathbf{X}_S^{\mathrm{T}}\boldsymbol{\beta}_S\right)\mathbf{X}_S^{\otimes 2}\right]\right) \leq \sigma_{\max}\kappa_{\max}$, for all $\boldsymbol{\beta}_S \in \mathcal{B}_S^0(d)$ and $S : \mathcal{M} \subseteq S, |S| < q$. Then, by (A5),

$$\sigma_{\min}\kappa_{\min}/2 \leq \Lambda\left(\mathbb{E}_n\left\{\sigma\left(\mathbf{X}_S^{\mathrm{T}}\boldsymbol{\beta}_{*S}\right)\mathbf{X}_S^{\otimes 2}\right\}\right) \leq 2\sigma_{\max}\kappa_{\max}$$

uniformly over all S satisfying $\mathcal{M} \subseteq S$ and $|S| \leq q$, with probability at least $1 - \exp(-3q \log p)$. Hence, the condition (A4) in [17] is satisfied with probability at least $1 - \exp(-3q \log p)$.

Additionally, for any $\boldsymbol{\beta}_S \in \mathcal{B}_S^0(d)$,

$$\begin{aligned}
&\left|\mathbb{E}_n\left\{\sigma\left(\mathbf{X}_S^{\mathrm{T}}\boldsymbol{\beta}_S\right)\left(\boldsymbol{\pi}^{\mathrm{T}}\mathbf{X}_S\right)^2\right\} - \mathbb{E}_n\left\{\sigma\left(\mathbf{X}_S^{\mathrm{T}}\boldsymbol{\beta}_{*S}\right)\left(\boldsymbol{\pi}^{\mathrm{T}}\mathbf{X}_S\right)^2\right\}\right| \\
\leq &\left|n^{-1/2}\mathbb{G}_n\left\{\sigma\left(\mathbf{X}_S^{\mathrm{T}}\boldsymbol{\beta}_S\right)\left(\boldsymbol{\pi}^{\mathrm{T}}\mathbf{X}_S\right)^2\right\}\right| + \left|n^{-1/2}\mathbb{G}_n\left\{\sigma\left(\mathbf{X}_S^{\mathrm{T}}\boldsymbol{\beta}_{*S}\right)\left(\boldsymbol{\pi}^{\mathrm{T}}\mathbf{X}_S\right)^2\right\}\right| \\
&+ \left|E\left[\sigma\left(\mathbf{X}_S^{\mathrm{T}}\boldsymbol{\beta}_S\right)\left(\boldsymbol{\pi}^{\mathrm{T}}\mathbf{X}_S\right)^2\right] - E\left[\sigma\left(\mathbf{X}_S^{\mathrm{T}}\boldsymbol{\beta}_{*S}\right)\left(\boldsymbol{\pi}^{\mathrm{T}}\mathbf{X}_S\right)^2\right]\right| \\
\leq &2C_5K^2\sqrt{q^3\log p/n} + \mu_{\max}\|\boldsymbol{\beta}_S - \boldsymbol{\beta}_{*S}\|\sqrt{|S|}K\lambda_{\max}.
\end{aligned}$$

Hence, the condition (A5) in [17] is satisfied uniformly over all S such that $\mathcal{M} \subseteq S$ and $|S| \leq q$, with probability at least $1 - \exp(-3q \log p)$.

Then (A4) can be shown by following the proof of Equation (3.2) in [17]. Thus, our forward stage stops at the kth step with probability at least $1 - \exp(-3q \log p)$. □

Proof of Theorem 4. Suppose that a covariate X_r is removed from S. For any $r \in \mathcal{M}$, since $\mathcal{M} \not\subseteq S\backslash\{r\}$ and r is the only element that is in $(S\backslash\{r\})^c \cap \mathcal{M}$, by Lemma A1 and (A2)

$$\begin{aligned}
&\ell_S(\hat{\boldsymbol{\beta}}_S) - \ell_{S\backslash\{r\}}(\hat{\boldsymbol{\beta}}_{S\backslash\{r\}}) \geq \ell_S(\boldsymbol{\beta}_S^*) - \ell_{S\backslash\{r\}}(\hat{\boldsymbol{\beta}}_{S\backslash\{r\}}) \\
= &\ell_{S\backslash\{r\}\cup\{r\}}(\boldsymbol{\beta}^*_{S\backslash\{r\}+r}) - \ell_{S\backslash\{r\}}(\hat{\boldsymbol{\beta}}_{S\backslash\{r\}}) \geq C_1 n^{-2\alpha},
\end{aligned}$$

with probability at least $1 - 6\exp(-6q \log p)$. From the proof of Theorem 1, we have for any $\eta_2 > 0$, $\mathrm{BIC}(S) - \mathrm{BIC}(S\backslash\{r\}) \leq -2C_1 n^{-2\alpha} + \eta_2 n^{-1}\log n < 0$, uniformly over $r \in \mathcal{M}$ and S satisfying $\mathcal{M} \subset S$ and $|S| \leq q$, with probability at least $1 - 6\exp(-6q \log p)$. □

Proof of Theorem 5. By Theorems 1–3, we have that the event $\Omega_1 := \{|\hat{\mathcal{M}}| \leq q \text{ and } \mathcal{M} \subseteq \hat{\mathcal{M}}\}$ holds with probability at least $1 - 25\exp(-2q \log p)$. Thus, in the rest of the proof, we restrict our attention on Ω_1.

As shown in the proof of Theorem 3, we obtain that $\boldsymbol{\beta}^*_{\hat{\mathcal{M}}} = \boldsymbol{\beta}_{*\hat{\mathcal{M}}}$. Then by Lemma A6, we have $\|\hat{\boldsymbol{\beta}}_{\hat{\mathcal{M}}} - \boldsymbol{\beta}^*_{\hat{\mathcal{M}}}\| \leq A_2K^{-1}\sqrt{q^2 \log p/n}$ with probability at least $1 - 6\exp(-6q \log p)$. Withdrawing the attention on Ω_1, we obtain that

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_*\| = \|\hat{\boldsymbol{\beta}}_{\hat{\mathcal{M}}} - \boldsymbol{\beta}_{*\hat{\mathcal{M}}}\| = \|\hat{\boldsymbol{\beta}}_{\hat{\mathcal{M}}} - \boldsymbol{\beta}^*_{\hat{\mathcal{M}}}\| \leq A_2K^{-1}\sqrt{q^2 \log p/n},$$

with probability at least $1 - 31\exp(-2q \log p)$. □

Additional Lemmas and Proofs

Lemma A1. *Given a model S such that $|S| < q$, $\mathcal{M} \not\subseteq S$, under Condition (6),*
(i): $\exists r \in S^c \cap \mathcal{M}$, such that $\boldsymbol{\beta}^*_{S+r} \neq (\boldsymbol{\beta}_S^{*\mathrm{T}}, 0)^{\mathrm{T}}$.
(ii): Suppose Conditions (1), (2) and (6′) hold. $\exists r \in S^c \cap \mathcal{M}$, such that $\|\boldsymbol{\beta}^{*\mathrm{T}}_{S+r} - (\boldsymbol{\beta}_S^{*\mathrm{T}}, 0)\| \geq C\sigma_{\max}^{-1}\kappa_{\max}^{-1}n^{-\alpha}$.

Proof. As $\boldsymbol{\beta}^*_{S+j}$ is the maximizer of $E\left[\ell_{S\cup\{j\}}(\boldsymbol{\beta}_{S+j})\right]$, by the concavity of $E\left[\ell_{S\cup\{j\}}(\boldsymbol{\beta}_{S+j})\right]$, $\boldsymbol{\beta}^*_{S+j}$ is the solution to the equation $E\left[\left(Y - \mu(\boldsymbol{\beta}_S^{*T}\mathbf{X}_S + \beta_j X_j)\right)\mathbf{X}_{S\cup\{j\}}\right] = \mathbf{0}$,
(i): Suppose that $\boldsymbol{\beta}^*_{S+j} = (\boldsymbol{\beta}_S^{*\mathrm{T}}, 0)^{\mathrm{T}}, \forall j \in S^c \cap \mathcal{M}$. Then,

$$\begin{aligned}
&0 = E\left[\left(Y - \mu(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S)\right)X_j\right] = E\left[\left(\mu(\boldsymbol{\beta}_*^{\mathrm{T}}\mathbf{X}) - \mu(\boldsymbol{\beta}_S^{*T}\mathbf{X}_S)\right)X_j\right] \\
&\Rightarrow \max_{j\in S^c\cap\mathcal{M}}\left|E\left[\left(\mu(\boldsymbol{\beta}_*^{\mathrm{T}}\mathbf{X}) - \mu(\boldsymbol{\beta}_S^{*T}\mathbf{X}_S)\right)X_j\right]\right| = 0,
\end{aligned}$$

which violates the Condition (6). Therefore, we can find a $r \in S^c \cap \mathcal{M}$, such that $\boldsymbol{\beta}^*_{S+r} \neq (\boldsymbol{\beta}^{*\mathrm{T}}_S, 0)^{\mathrm{T}}$.
(ii): Let $r \in S^c \cap \mathcal{M}$ satisfy that $\left|E\left[\left(\mu(\boldsymbol{\beta}^{\mathrm{T}}_* \mathbf{X}) - \mu(\boldsymbol{\beta}^{*\mathrm{T}}_S \mathbf{X}_S)\right) X_r\right]\right| > Cn^{-\alpha}$. Without loss of generality, we assume that X_r is the last element of $\mathbf{X}_{S\cup\{r\}}$. By the mean value theorem,

$$\begin{aligned}
&E\left[\left(\mu(\boldsymbol{\beta}^{\mathrm{T}}_* \mathbf{X}) - \mu(\boldsymbol{\beta}^{*\mathrm{T}}_S \mathbf{X}_S)\right) X_r\right] \\
&= E\left[\left(\mu(\boldsymbol{\beta}^{\mathrm{T}}_* \mathbf{X}) - \mu(\boldsymbol{\beta}^{*\mathrm{T}}_S \mathbf{X}_S)\right) X_r\right] - E\left[\left(\mu(\boldsymbol{\beta}^{\mathrm{T}}_* \mathbf{X}) - \mu(\boldsymbol{\beta}^{*\mathrm{T}}_{S+r} \mathbf{X}_{S\cup\{r\}})\right) X_r\right] \\
&= E\left[\left(\mu(\boldsymbol{\beta}^{*\mathrm{T}}_{S+r} \mathbf{X}_{S\cup\{r\}}) - \mu((\boldsymbol{\beta}^{*\mathrm{T}}_S, 0) \mathbf{X}_{S\cup\{r\}})\right) X_r\right] \\
&= (\boldsymbol{\beta}^{*\mathrm{T}}_{S+r} - (\boldsymbol{\beta}^{*\mathrm{T}}_S, 0)) E\left[\sigma(\tilde{\boldsymbol{\beta}}^{\mathrm{T}}_{S+r} \mathbf{X}_{S\cup\{r\}}) \mathbf{X}^{\otimes 2}_{S\cup\{r\}}\right] \mathbf{e}_r, \qquad \text{(A6)}
\end{aligned}$$

where $\tilde{\boldsymbol{\beta}}_{S+r}$ is some point between $\boldsymbol{\beta}^*_{S+r}$ and $(\boldsymbol{\beta}^{*\mathrm{T}}_S, 0)^{\mathrm{T}}$ and $\mathbf{e}_r$ is a vector of length $(|S|+1)$ with the rth element being 1.

As $\tilde{\boldsymbol{\beta}}_{S+r}$ is some point between $\boldsymbol{\beta}^*_{S+r}$ and $(\boldsymbol{\beta}^{*\mathrm{T}}_S, 0)^{\mathrm{T}}$, $|\tilde{\boldsymbol{\beta}}^{\mathrm{T}}_{S+r} \mathbf{X}_{S\cup\{r\}}| \leq |\boldsymbol{\beta}^{*\mathrm{T}}_{S+r} \mathbf{X}_{S\cup\{r\}}| + |(\boldsymbol{\beta}^{*\mathrm{T}}_S, 0) \mathbf{X}_{S\cup\{r\}}| \leq 2K^2$, by Conditions (1) and (2). Thus, $|\sigma(\tilde{\boldsymbol{\beta}}^{\mathrm{T}}_{S+r} \mathbf{X}_{S\cup\{r\}})| \leq \sigma_{\max}$. By (A6) and Condition (5),

$$\begin{aligned}
Cn^{-\alpha} &\leq \left|E\left[\left(\mu(\boldsymbol{\beta}^{\mathrm{T}}_* \mathbf{X}) - \mu(\boldsymbol{\beta}^{*\mathrm{T}}_S \mathbf{X}_S)\right) X_r\right]\right| \\
&\leq \|\boldsymbol{\beta}^{*\mathrm{T}}_{S+r} - (\boldsymbol{\beta}^{*\mathrm{T}}_S, 0)\| \sigma_{\max} \lambda_{\max}\left(E\left[\mathbf{X}^{\otimes 2}_{S\cup\{r\}}\right]\right) \|\mathbf{e}_r\| \leq \sigma_{\max} \kappa_{\max} \|\boldsymbol{\beta}^{*\mathrm{T}}_{S+r} - (\boldsymbol{\beta}^{*\mathrm{T}}_S, 0)\|.
\end{aligned}$$

Therefore, $\|\boldsymbol{\beta}^{*\mathrm{T}}_{S+r} - (\boldsymbol{\beta}^{*\mathrm{T}}_S, 0)\| \geq C\sigma^{-1}_{\max}\kappa^{-1}_{\max} n^{-\alpha}$. □

Lemma A2. *Let $\xi_i, i = 1, \ldots, n$ be n i.i.d random variables such that $|\xi_i| \leq B$ for a constant $B > 0$. Under Conditions (1)–(3), we have $E\left[|Y_i\xi_i - E\left[Y_i\xi_i\right]|^m\right] \leq m!(2B(\sqrt{2}M + \mu_{\max}))^m$, for every $m \geq 1$.*

Proof. By Conditions (1) and (2), $|\boldsymbol{\beta}^{\mathrm{T}}_* \mathbf{X}_i| \leq KL$, $\forall i \geq 1$ and consequently $\left|\mu(\boldsymbol{\beta}^{\mathrm{T}}_* \mathbf{X}_i)\right| \leq \mu_{\max}$. Then by Condition (3),

$$\begin{aligned}
E[|Y_i|^m] &= E[|\epsilon_i + \mu(\boldsymbol{\beta}^{\mathrm{T}}_* \mathbf{X}_i)|^m] \leq \sum_{t=0}^{m} \binom{m}{t} E\left[|\epsilon_i|^t\right] \mu^{m-t}_{\max} \\
&\leq \sum_{t=0}^{m} t! \binom{m}{t} M^t \mu^{m-t}_{\max} \leq m!(M + \mu_{\max})^m,
\end{aligned}$$

for every $m \geq 1$. By the same arguments, it can be shown that, for every $m \geq 1$, $E\left[|Y_i\xi_i - E\left[Y_i\xi_i\right]|^m\right] \leq E\left[\left(|Y_i\xi_i| + |E\left[Y_i\xi_i\right]|\right)^m\right] \leq m!(2B(M + \mu_{\max}))^m$. □

Lemma A3. *Under Conditions (1)–(3), when n is sufficiently large such that $28\sqrt{q \log p / n} < 1$, we have $\sup_{\boldsymbol{\beta}\in\mathbb{B}} \left|\mathbb{E}_n\left\{L(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, Y)\right\}\right| \leq 2(M + \mu_{\max})K^3 + b_{\max}$, with probability $1 - 2\exp(-10q \log p)$.*

Proof. By Conditions (2), $\sup_{\boldsymbol{\beta}\in\mathbb{B}} \left|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\right| \leq K^3$. Thus,

$$\begin{aligned}
&\sup_{\boldsymbol{\beta}\in\mathbb{B}} \left|\mathbb{E}_n\left\{L(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, Y)\right\}\right| \leq \sup_{\boldsymbol{\beta}\in\mathbb{B}} \left|\mathbb{E}_n\left\{\left|Y\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\right|\right\}\right| + b_{\max} \\
&\leq \left(\left|\mathbb{E}_n\left\{|Y| - E\left[|Y|\right]\right\}\right| + E\left[|Y|\right]\right) K^3 + b_{\max} \\
&\leq \left(\left|\mathbb{E}_n\left\{|Y| - E\left[|Y|\right]\right\}\right|\right) K^3 + (M + \mu_{\max})K^3 + b_{\max},
\end{aligned}$$

where the last inequality follows from that $E[|Y|] \leq M + \mu_{\max}$ as shown in the proof of Lemma A2.

Let $\xi_i = 1\{Y_i > 0\} - 1\{Y_i < 0\}$. Thus $|\xi_i| \leq 1$. By Lemma A2, we have $E\left[\left||Y_i| - E\left[|Y_i|\right]\right|^m\right] \leq m!(2(M+\mu_{\max}))^m$. Applying Bernstein's inequality (e.g., Lemma 2.2.11 in [50]) yields that

$$\begin{aligned} &P\left(\left|\mathbb{E}_n\left\{|Y| - E\left[|Y|\right]\right\}\right| > 10(M+\mu_{\max})\sqrt{q\log p/n}\right) \\ &\leq 2\exp\left(-\frac{1}{2}\frac{196q\log p}{4+20\sqrt{q\log p/n}}\right) \leq 2\exp(-10q\log p), \end{aligned} \tag{A7}$$

when n is sufficiently large such that $20\sqrt{q\log p/n} < 1$. Since $10(M+\mu_{\max})\sqrt{q\log p/n} = o(1)$, then

$$P\left(\sup_{\boldsymbol{\beta}\in\mathbb{B}}\left|\mathbb{E}_n\left\{L(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, Y)\right\}\right| \geq 2(M+\mu_{\max})K^3 + b_{\max}\right) \leq 2\exp(-10q\log p).$$

□

Lemma A4. *Given an index set S and $r \in S^c$, let $\mathcal{B}_S^0(d) = \{\boldsymbol{\beta}_S : \|\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*\| \leq d/(K\sqrt{|S|})\}$ and $A_1 := (M + 2\mu_{\max})$. Under Conditions (1)–(3), when n is sufficiently large such that $10\sqrt{q\log p/n} < 1$, we have*

1. $|\mathbb{G}_n\left[L\left(\boldsymbol{\beta}_S^{\mathrm{T}}\mathbf{X}_S, Y\right) - L\left(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S, Y\right)\right]| \leq 20A_1 d\sqrt{q\log p}$, *uniformly over $\boldsymbol{\beta}_S \in \mathcal{B}_S^0(d)$ and $|S| \leq q$, with probability at least $1 - 4\exp(-6q\log p)$.*
2. $|\mathbb{G}_n\left[L(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S, Y)\right]| \leq 10(A_1K^2 + b_{\max})\sqrt{q\log p}$, *uniformly over $|S| \leq q$, with probability at least $1 - 2\exp(-8q\log p)$.*

Proof. : (1): Let $\mathcal{R}_{|S|}(d)$ be a $|S|$-dimensional ball with center at 0 and radius $d/(K\sqrt{|S|})$. Then $\mathcal{B}_S^0(d) = \mathcal{R}_{|S|}(d) + \boldsymbol{\beta}_S^*$. Let $\mathcal{C}_{|S|} := \{\mathcal{C}(\boldsymbol{\xi}_k)\}$ be a collection of cubes that cover the ball $R_{|S|}(d)$, where $\mathcal{C}(\boldsymbol{\xi}_k)$ is a cube containing $\boldsymbol{\xi}_k$ with sides of length $d/(K\sqrt{|S|}n^2)$ and $\boldsymbol{\xi}_k$ is some point in $\mathcal{R}_{|S|}(d)$. As the volume of $\mathcal{C}(\boldsymbol{\xi}_k)$ is $\left(d/(K\sqrt{|S|}n^2)\right)^{|S|}$ and the volume of $\mathcal{R}_{|S|}(d)$ is less than $(2d/(K\sqrt{|S|}))^{|S|}$, we can select $\boldsymbol{\xi}_k$s so that no more than $(4n^2)^{|S|}$ cubes are needed to cover $\mathcal{R}_{|S|}(d)$. We thus assume $|\mathcal{C}_{|S|}| \leq (4n^2)^{|S|}$. For any $\boldsymbol{\xi} \in \mathcal{C}(\boldsymbol{\xi}_k)$, $\|\boldsymbol{\xi} - \boldsymbol{\xi}_k\| \leq d/(Kn^2)$. In addition, let $T_{1S}(\boldsymbol{\xi}) := \mathbb{E}_n\left[Y\boldsymbol{\xi}^{\mathrm{T}}\mathbf{X}_S\right]$, $T_{2S}(\boldsymbol{\xi}) := \mathbb{E}_n\left[b\left(\left(\boldsymbol{\beta}_S^* + \boldsymbol{\xi}\right)^{\mathrm{T}}\mathbf{X}_S\right) - b\left(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S\right)\right]$, and $T_S(\boldsymbol{\xi}) := T_{1S}(\boldsymbol{\xi}) - T_{2S}(\boldsymbol{\xi})$.
Given any $\boldsymbol{\xi} \in \mathcal{R}_{|S|}(d)$, there exists $\mathcal{C}(\boldsymbol{\xi}_k) \in \mathcal{C}_{|S|}$ such that $\boldsymbol{\xi} \in \mathcal{C}(\boldsymbol{\xi}_k)$. Then

$$\begin{aligned} |T_S(\boldsymbol{\xi}) - E\left[T_S(\boldsymbol{\xi})\right]| &\leq |T_S(\boldsymbol{\xi}) - T_S(\boldsymbol{\xi}_k)|\,|T_S(\boldsymbol{\xi}_k) - E\left[T_S(\boldsymbol{\xi}_k)\right]| + |E\left[T_S(\boldsymbol{\xi})\right] - E\left[T_S(\boldsymbol{\xi}_k)\right]| \\ &=: I + II + III. \end{aligned}$$

We deal with III first. By the mean value theorem, there exists a $\tilde{\boldsymbol{\xi}}$ between $\boldsymbol{\xi}$ and $\boldsymbol{\xi}_k$ such that

$$\begin{aligned} &|E\left[T_S(\boldsymbol{\xi}_k)\right] - E\left[T_S(\boldsymbol{\xi})\right]| = \left|E\left[Y(\boldsymbol{\xi}_k - \boldsymbol{\xi})^{\mathrm{T}}\mathbf{X}_S\right] + E\left[\mu\left(\left(\boldsymbol{\beta}_S^* + \tilde{\boldsymbol{\xi}}\right)^{\mathrm{T}}\mathbf{X}_S\right)(\boldsymbol{\xi}_k - \boldsymbol{\xi})^{\mathrm{T}}\mathbf{X}_S\right]\right| \\ &\leq E[|Y|]\|\boldsymbol{\xi}_k - \boldsymbol{\xi}\|\|\mathbf{X}_S\| + \mu_{\max}\|\boldsymbol{\xi}_k - \boldsymbol{\xi}\|\|\mathbf{X}_S\| \leq (M + 2\mu_{\max})d\sqrt{|S|}n^{-2} = A_1 d\sqrt{|S|}n^{-2}, \end{aligned} \tag{A8}$$

where the last inequality follows from Lemma A2 and $A_1 = M + 2\mu_{\max}$.

Next, we evaluate II. By Condition (2), $|\mathbf{X}_{iS}^{\mathrm{T}}\boldsymbol{\xi}| \leq \|\mathbf{X}_{iS}\|\|\boldsymbol{\xi}\| \leq d/(K\sqrt{|S|})\sqrt{|S|}K = d$, for all $\boldsymbol{\xi} \in \mathcal{R}_{|S|}(d)$. Then by Lemma A2,

$$E\left[\left|Y\boldsymbol{\xi}_k^{\mathrm{T}}\mathbf{X}_S - E\left[Y\boldsymbol{\xi}_k^{\mathrm{T}}\mathbf{X}_S\right]\right|^m\right] \leq m!(2(M+\mu_{\max})d)^m.$$

By Bernstein's inequality, when n is sufficiently large such that $10\sqrt{q\log p/n} \le 1$.

$$P\left(|T_{1S}(\boldsymbol{\xi}_k) - E\left[T_{1S}(\boldsymbol{\xi}_k)\right]| > 10(M+\mu_{\max})d\sqrt{qn^{-1}\log p}\right)$$
$$\le 2\exp\left(-\frac{1}{2}\frac{100q\log p}{4+20\sqrt{q\log p/n}}\right) \le 2\exp(-10q\log p). \tag{A9}$$

Since $|b((\boldsymbol{\beta}_S^* + \boldsymbol{\xi}_k)^{\mathrm{T}}\mathbf{X}_S) - b(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S)| \le \mu_{\max}d$, by the same arguments used for (A9), we have

$$P\left(|T_{2S}(\boldsymbol{\xi}_k) - E\left[T_{2S}(\boldsymbol{\xi}_k)\right]| > 10\mu_{\max}d\sqrt{qn^{-1}\log p}\right) \le 2\exp(-10q\log p). \tag{A10}$$

Combining (A9) and (A10) yields that uniformly over $\boldsymbol{\xi}_k$

$$|T_S(\boldsymbol{\xi}_k) - E\left[T_S(\boldsymbol{\xi}_k)\right]| \le 10A_1d\sqrt{qn^{-1}\log p}, \tag{A11}$$

with probability at least $1 - 2(4n^2)^{|S|}\exp(-10q\log p)$.

We now assess I. Following the same arguments as in Lemma A3,

$$P\Big(\sup_{\boldsymbol{\xi}\in\mathcal{C}(\boldsymbol{\xi}_k)}|T_S(\boldsymbol{\xi}) - T_S(\boldsymbol{\xi}_k)| > (2M+3\mu_{\max})d\sqrt{|S|}n^{-2}\Big) \le 2\exp(-8q\log p). \tag{A12}$$

Since $\sqrt{|S|}n^{-2} = o(\sqrt{qn^{-1}\log p})$, combining (A8), (A11) and (A12) together yields that

$$P\Big(\sup_{\boldsymbol{\xi}\in\mathcal{R}_{|S|}(d)}|T_S(\boldsymbol{\xi}) - E\left[T_S(\boldsymbol{\xi})\right]| \ge 20A_1d\sqrt{qn^{-1}\log p}\Big)$$
$$\le 2(4n^2)^{|S|}\exp(-10q\log p) + 2\exp(-8q\log p) \le 4\exp(-8q\log p).$$

By the combinatoric inequality $\binom{p}{s} \le (ep/s)^s$, we obtain that

$$P\Big(\sup_{|S|\le q,\boldsymbol{\beta}_S\in\mathcal{B}_S^0(d_1)}\left|\mathbb{G}_n\left[L\left(\boldsymbol{\beta}_S^{\mathrm{T}}\mathbf{X}_S, Y\right) - L\left(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S, Y\right)\right]\right| \ge 20A_1d\sqrt{q\log p}\Big)$$
$$\le \sum_{s=1}^{q}(ep/s)^s 4\exp(-8q\log p) \le 4\exp(-6q\log p).$$

(2): We evaluate the mth moment of $L(\boldsymbol{\beta}_S^*\mathbf{X}_S, Y)$.

$$E\left[\left(Y\boldsymbol{\beta}_S^*\mathbf{X}_S - b(\boldsymbol{\beta}_S^*\mathbf{X}_S)\right)^m\right] \le E\left[\sum_{t=0}^{m}\binom{m}{t}|Y|^tK^{2t}b_{\max}^{m-t}\right]$$
$$\le \sum_{t=0}^{m}\binom{m}{t}t!\left((M+\mu_{\max})K^2\right)^t b_{\max}^{m-t} \le m!((M+\mu_{\max})K^2 + b_{\max})^m.$$

Then, by Bernstein's inequality,

$$P\left(|\mathbb{G}_n\left[L(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S, Y)\right]| > 10(A_1K^2 + b_{\max})\sqrt{q\log p}\right) \le 2\exp(-10q\log p).$$

By the same arguments used in (i), we obtain that

$$P\Big(\sup_{|S|\le q}\Big|\mathbb{G}_n\left[L\left(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S,Y\right)\right]\Big|\ge 10(A_1K^2+b_{\max})\sqrt{q\log p}\Big)$$
$$\le\sum_{s=1}^{q}(ep/s)^s 2\exp(-10q\log p)\le 2\exp(-8q\log p).$$

□

Lemma A5. *Given a model S and $r\in S^c$, under Conditions (1), (2) and (5), for any $\|\boldsymbol{\beta}_S-\boldsymbol{\beta}_S^*\|\le K/\sqrt{|S|}$, $\sigma_{\min}\kappa_{\min}\|\boldsymbol{\beta}_S-\boldsymbol{\beta}_S^*\|^2/2\le E\left[\ell_S(\boldsymbol{\beta}_S^*)\right]-E\left[\ell_S(\boldsymbol{\beta}_S)\right]\le\sigma_{\max}\kappa_{\max}\|\boldsymbol{\beta}_S-\boldsymbol{\beta}_S^*\|^2/2$.*

Proof. Due to the concavity of the log-likelihood in GLMs with the canonical link, $E\left[Y\mathbf{X}_S-\mu(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S)\mathbf{X}_S\right]=\mathbf{0}$. Then for any $\|\boldsymbol{\beta}_S-\boldsymbol{\beta}_S^*\|\le K/\sqrt{|S|}$, $|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_S|\le|\boldsymbol{\beta}^{*\mathrm{T}}\mathbf{X}_S|+|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^{\mathrm{T}}\mathbf{X}_S|\le K^2+K/\sqrt{|S|}\times K\sqrt{|S|}=2KL$. Thus, by Taylor's expansion,

$$E\left[\ell_S(\boldsymbol{\beta}_S)\right]-E\left[\ell_S(\boldsymbol{\beta}_S^*)\right]=-\frac{1}{2}(\boldsymbol{\beta}_S-\boldsymbol{\beta}_S^*)^{\mathrm{T}}E\left[\sigma\left(\tilde{\boldsymbol{\beta}}_S^{\mathrm{T}}\mathbf{X}_S\right)\mathbf{X}_S^{\otimes 2}\right](\boldsymbol{\beta}_S-\boldsymbol{\beta}_S^*),$$

where $\tilde{\boldsymbol{\beta}}_S$ is between $\boldsymbol{\beta}_S$ and $\boldsymbol{\beta}_S^*$. By Condition (5), $\sigma_{\min}\kappa_{\min}\|\boldsymbol{\beta}_S-\boldsymbol{\beta}_S^*\|^2/2\le E\left[\ell_S(\boldsymbol{\beta}_S^*)\right]-E\left[\ell_S(\boldsymbol{\beta}_S)\right]\le\sigma_{\max}\kappa_{\max}\|\boldsymbol{\beta}_S-\boldsymbol{\beta}_S^*\|^2/2$.

□

Lemma A6. *Under Conditions (1)–(6), there exist some constants A_2 and A_3 that do not depend on n, such that $\|\hat{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_S^*\|\le A_2K^{-1}\sqrt{q^2\log p/n}$ and $|\ell_S(\hat{\boldsymbol{\beta}}_S)-\ell_S(\boldsymbol{\beta}_S^*)|\le A_3q^2\log p/n$ hold uniformly over $S:|S|\le q$, with probability at least $1-6\exp(-6q\log p)$.*

Proof. Define

$$\Omega(d):=\Big\{\sup_{|S|\le q,\boldsymbol{\beta}_S\in\mathcal{B}_S^0(d)}\Big|\mathbb{G}_n\Big[L\left(\boldsymbol{\beta}_S^{\mathrm{T}}\mathbf{X}_S,Y\right)-L\left(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S,Y\right)\Big]\Big|<20A_1d\sqrt{q\log p}\Big\}.$$

By Lemma A4, the event $\Omega(d)$ holds with probability at least $1-4\exp(-6q\log p)$. Thus, in the proof of Lemma A6, we shall assume $\Omega(d)$ hold with $d=A_2\sqrt{q^3\log p/n}$ for some $A_2>20(\sigma_{\min}\kappa_{\min})^{-1}K^2A_1$.

For any $\|\boldsymbol{\beta}_S-\boldsymbol{\beta}_S^*\|=A_2K^{-1}\sqrt{q^2\log p/n}$, since $\sqrt{q^2\log p/n}\le\sqrt{q^3\log p/n}/\sqrt{|S|}$, $\boldsymbol{\beta}_S\in\mathcal{B}_S^0(d)$. By Lemma A5,

$$\begin{aligned}&\ell_S(\boldsymbol{\beta}_S^*)-\ell_S(\boldsymbol{\beta}_S)\\&=\Big(\ell_S(\boldsymbol{\beta}_S^*)-E\left[\ell_S(\boldsymbol{\beta}_S^*)\right]-\left(\ell_S(\boldsymbol{\beta}_S)-E\left[\ell_S(\boldsymbol{\beta}_S)\right]\right)\Big)+\left(E\left[\ell_S(\boldsymbol{\beta}_S^*)\right]-E\left[\ell_S(\boldsymbol{\beta}_S)\right]\right)\\&\ge\sigma_{\min}\kappa_{\min}\|\boldsymbol{\beta}_S-\boldsymbol{\beta}_S^*\|^2/2-20A_1d\sqrt{q\log p/n}\\&=\sigma_{\min}\kappa_{\min}A_2^2q^2\log p/(K^2n)-20A_1A_2q^2\log p/n>0.\end{aligned}$$

Thus,

$$\inf_{|S|\le q,\|\boldsymbol{\beta}_S-\boldsymbol{\beta}_S^*\|=A_2K^{-1}\sqrt{q^2\log p/n}}\ell_S(\boldsymbol{\beta}_S^*)-\ell_S(\boldsymbol{\beta}_S)>0.$$

Then by the concavity of $\ell_S(\cdot)$, we obtain that $\max_{|S|\le q}\left\|\hat{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_S^*\right\|\le A_2K^{-1}\sqrt{q^2n^{-1}\log p}$.

On the other hand, for any $\|\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*\| \leq A_2 K^{-1}\sqrt{q^2 \log p/n}$,

$$\begin{aligned}
&|\ell_S(\boldsymbol{\beta}_S^*) - \ell_S(\boldsymbol{\beta}_S)| \\
&\leq \left|\ell_S(\boldsymbol{\beta}_S^*) - E\left[\ell_S(\boldsymbol{\beta}_S^*)\right] - \left(\ell_S(\boldsymbol{\beta}_S) - E\left[\ell_S(\boldsymbol{\beta}_S)\right]\right)\right| + \left|E\left[\ell_S(\boldsymbol{\beta}_S^*)\right] - E\left[\ell_S(\boldsymbol{\beta}_S)\right]\right| \\
&\leq \sigma_{\max}\kappa_{\max}\|\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*\|^2/2 + 20A_1 d\sqrt{q \log p/n} \leq A_3 q^2 n^{-1} \log p,
\end{aligned}$$

where $A_3 := 4\sigma_{\max}\lambda_{\max}A_2^2 K^{-2} + 20A_1A_2$. □

Appendix B. Additional Results in the Applications

Table A1. Comparison of genes selected by each competing method from the mammalian eye data set.

	STEPWISE	FR	LASSO	SIS+LASSO	SC	dgLARS
STEPWISE	3	3	2	2	2	0
FR		4	2	2	2	0
LASSO			16	5	2	0
SIS+LASSO				9	2	0
SC					4	0
dgLARS						7

Note: Diagonal and off-diagonal elements of the table represent the model sizes for each method and the number of overlapping genes selected by the two methods corresponding to the row and column, respectively.

Table A2. Selected miRNAs for ESCC training dataset.

Methods	Selected miRNAs
STEPWISE	*miR-4783-3p*; *miR-320*b; *miR-1225-3p*
FR	*miR-4783-3p*; *miR-320b*; *miR-1225-3p*; *6789-5p*
SC	*miR-4783-3p*; *miR-320b*; *miR-1225-3p*; *6789-5p*
LASSO	*miR-6789-5p*; *miR-6781-5p*; *miR-1225-3p*; *miR-1238-5p*; *miR-320b*; *miR-6794-5p*; *miR-6877-5p*; *miR-6785-5p*; *miR-718*; *miR-195-5p*
SIS+LASSO	*miR-6785-5p*; *miR-1238-5p*; *miR-1225-3p*; *miR-6789-5p*; *miR-320b*; *miR-6875-5p*; *miR-6127*; *miR-1268b*; *miR-6781-5p*; *miR-125a-3p*
dgLARS	*miR-891b*; *miR-6127*; *miR-151a-5p*; *miR-195-5p*; ; *miR-3688-5p* *miR-125b-1-3p*; *miR-1273c*; *miR-6501-5p*; *miR-4666a-5p*; *miR-514a-3p*

Note: LASSO, SIS+LASSO, dgLARS selected 20, 17 and 33 miRNAs, respectively, and we only reported top 10 miRNAs.

References

1. Prosperi, M.; Min, J.S.; Bian, J.; Modave, F. Big data hurdles in precision medicine and precision public health. *BMC Med. Inform. Decis. Mak.* **2018**, *18*, 139. [CrossRef] [PubMed]
2. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B-Stat. Methodol.* **1996**, *58*, 267–288. [CrossRef]
3. Flynn, C.J.; Hurvich, C.M.; Simonoff, J.S. On the sensitivity of the lasso to the number of predictor variables. *Stat. Sci.* **2017**, *32*, 88–105. [CrossRef]
4. van de Geer, S.A. On the asymptotic variance of the debiased Lasso. *Electron. J. Stat.* **2019**, *13*, 2970–3008. [CrossRef]
5. Fan, J.; Lv, J. Sure independence screening for ultrahigh dimensional feature space (with discussion). *J. R. Stat. Soc. Ser. B-Stat. Methodol.* **2008**, *70*, 849–911. [CrossRef]
6. Barut, E.; Fan, J.; Verhasselt, A. Conditional sure independence screening. *J. Am. Stat. Assoc.* **2016**, *111*, 1266–1277. [CrossRef] [PubMed]

7. Wang, H. Forward regression for ultra-high dimensional variable screening. *J. Am. Stat. Assoc.* **2009**, *104*, 1512–1524. [CrossRef]
8. Zheng, Q.; Hong, H.G.; Li, Y. Building generalized linear models with ultrahigh dimensional features: A sequentially conditional approach. *Biometrics* **2019**, *76*, 1–14. [CrossRef]
9. Hong, H.G.; Zheng, Q.; Li, Y. Forward regression for Cox models with high-dimensional covariates. *J. Multivar. Anal.* **2019**, *173*, 268–290. [CrossRef]
10. Efron, B.; Hastie, T.; Johnstone, I.; Tibshirani, R. Least angle regression. *Ann. Stat.* **2004**, *32*, 407–499.
11. Augugliaro, L.; Mineo, A.M.; Wit, E.C. Differential geometric least angle regression: a differential geometric approach to sparse generalized linear models. *J. R. Stat. Soc. Ser. B-Stat. Methodol.* **2013**, *75*, 471–498. [CrossRef]
12. Pazira, H.; Augugliaro, L.; Wit, E. Extended differential geometric LARS for high-dimensional GLMs with general dispersion parameter. *Stat. Comput.* **2018**, *28*, 753–774. [CrossRef]
13. An, H.; Huang, D.; Yao, Q.; Zhang, C.H. Stepwise searching for feature variables in high-dimensional linear regression. 2008. Available online: http://eprints.lse.ac.uk/51349/ (accessed on 20 August 2020).
14. Ing, C.K.; Lai, T.L. A stepwise regression method and consistent model selection for high-dimensional sparse linear models. *Stat. Sin.* **2011**, *21*, 1473–1513. [CrossRef]
15. Hwang, J.S.; Hu, T.H. A stepwise regression algorithm for high-dimensional variable selection. *J. Stat. Comput. Simul.* **2015**, *85*, 1793–1806. [CrossRef]
16. McCullagh, P. *Generalized Linear Models*; Routledge: Abingdon-on-Thames, UK, 1989.
17. Chen, J.; Chen, Z. Extended BIC for small-*n*-large-*P* sparse GLM. *Stat. Sin.* **2012**, 22, 555–574. [CrossRef]
18. Bühlmann, P.; Yu, B. Sparse boosting. *J. Mach. Learn. Res.* **2006**, *7*, 1001–1024.
19. van de Geer, S.A. High-dimensional generalized linear models and the lasso. *Ann. Stat.* **2008**, *36*, 614–645. [CrossRef]
20. Chen, J.; Chen, Z. Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **2008**, *95*, 759–771. [CrossRef]
21. Fan, Y.; Tang, C.Y. Tuning parameter selection in high dimensional penalized likelihood. *J. R. Stat. Soc. Ser. B-Stat. Methodol.* **2013**, *75*, 531–552. [CrossRef]
22. Cheng, M.Y.; Honda, T.; Zhang, J.T. Forward variable selection for sparse ultra-high dimensional varying coefficient models. *J. Am. Stat. Assoc.* **2016**, *111*, 1209–1221. [CrossRef]
23. Zhao, S.D.; Li, Y. Principled sure independence screening for Cox models with ultra-high-dimensional covariates. *J. Multivar. Anal.* **2012**, *105*, 397–411. [CrossRef] [PubMed]
24. Kwemou, M. Non-asymptotic oracle inequalities for the Lasso and group Lasso in high dimensional logistic model. *ESAIM-Prob. Stat.* **2016**, *20*, 309–331. [CrossRef]
25. Jiang, Y.; He, Y.; Zhang, H. Variable selection with prior information for generalized linear models via the prior LASSO method. *J. Am. Stat. Assoc.* **2016**, *111*, 355–376. [CrossRef]
26. Zhang, C.H.; Huang, J. The sparsity and bias of the Lasso selection in high-dimensional linear regression. *Ann. Stat.* **2008**, *36*, 1567–1594. [CrossRef]
27. Fan, J.; Song, R. Sure independence screening in generalized linear models with NP-dimensionality. *Ann. Stat.* **2010**, *38*, 3567–3604. [CrossRef]
28. Luo, S.; Chen, Z. Sequential Lasso cum EBIC for feature selection with ultra-high dimensional feature space. *J. Am. Stat. Assoc.* **2014**, *109*, 1229–1240. [CrossRef]
29. Luo, S.; Xu, J.; Chen, Z. Extended Bayesian information criterion in the Cox model with a high-dimensional feature space. *Ann. Inst. Stat. Math.* **2015**, *67*, 287–311. [CrossRef]
30. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data mining, Inference, and Prediction*; Springer: Berlin/Heidelberger, Germany, 2009.
31. Simon, N.; Friedman, J.; Hastie, T.; Tibshirani, R. Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *J. Stat. Softw.* **2011**, *39*, 1–13. [CrossRef]
32. Breheny, P.; Huang, J. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Ann. Appl. Stat.* **2011**, *5*, 232–253. [CrossRef]
33. Wang, X.; Leng, C. R package: screening, 2016. Available Online: https://github.com/wwrechard/screening (accessed on 20 August 2020).
34. Augugliaro, L.; Mineo, A.M.; Wit, E.C. dglars: An R Package to Estimate Sparse Generalized Linear Models. *J. Stat. Softw.* **2014**, *59*, 1–40. [CrossRef]

35. Scheetz, T.E.; Kim, K.Y.A.; Swiderski, R.E.; Philp, A.R.; Braun, T.A.; Knudtson, K.L.; Dorrance, A.M.; DiBona, G.F.; Huang, J.; Casavant, T.L.; et al. Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 14429–14434. [CrossRef] [PubMed]
36. Chiang, A.P.; Beck, J.S.; Yen, H.J.; Tayeh, M.K.; Scheetz, T.E.; Swiderski, R.E.; Nishimura, D.Y.; Braun, T.A.; Kim, K.Y.A.; Huang, J.; et al. Homozygosity mapping with SNP arrays identifies TRIM32, an E3 ubiquitin ligase, as a Bardet–Biedl syndrome gene (BBS11). *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 6287–6292. [CrossRef] [PubMed]
37. He, S.; Peng, J.; Li, L.; Xu, Y.; Wu, X.; Yu, J.; Liu, J.; Zhang, J.; Zhang, R.; Wang, W. High expression of cytokeratin CAM5.2 in esophageal squamous cell carcinoma is associated with poor prognosis. *Medicine* **2019**, *98*, e17104. [CrossRef]
38. Li, B.X.; Yu, Q.; Shi, Z.L.; Li, P.; Fu, S. Circulating microRNAs in esophageal squamous cell carcinoma: association with locoregional staging and survival. *Int. J. Clin. Exp. Med.* **2015**, *8*, 7241–7250.
39. Sudo, K.; Kato, K.; Matsuzaki, J.; Boku, N.; Abe, S.; Saito, Y.; Daiko, H.; Takizawa, S.; Aoki, Y.; Sakamoto, H.; et al. Development and validation of an esophageal squamous cell carcinoma detection model by large-scale microRNA profiling. *JAMA Netw. Open* **2019**, *2*, e194573–e194573. [CrossRef]
40. Zhang, Y. Epidemiology of esophageal cancer. *World J. Gastroenterol* **2013**, *19*, 5598–5606. [CrossRef]
41. Mathieu, L.N.; Kanarek, N.F.; Tsai, H.L.; Rudin, C.M.; Brock, M.V. Age and sex differences in the incidence of esophageal adenocarcinoma: results from the Surveillance, Epidemiology, and End Results (SEER) Registry (1973–2008). *Dis. Esophagus* **2014**, *27*, 757–763. [CrossRef]
42. Zhou, J.; Zhang, M.; Huang, Y.; Feng, L.; Chen, H.; Hu, Y.; Chen, H.; Zhang, K.; Zheng, L.; Zheng, S. MicroRNA-320b promotes colorectal cancer proliferation and invasion by competing with its homologous microRNA-320a. *Cancer Lett.* **2015**, *356*, 669–675. [CrossRef]
43. Lieb, V.; Weigelt, K.; Scheinost, L.; Fischer, K.; Greither, T.; Marcou, M.; Theil, G.; Klocker, H.; Holzhausen, H.J.; Lai, X.; et al. Serum levels of miR-320 family members are associated with clinical parameters and diagnosis in prostate cancer patients. *Oncotarget* **2018**, *9*, 10402–10416. [CrossRef]
44. Mullany, L.E.; Herrick, J.S.; Wolff, R.K.; Stevens, J.R.; Slattery, M.L. Association of cigarette smoking and microRNA expression in rectal cancer: insight into tumor phenotype. *Cancer Epidemiol.* **2016**, *45*, 98–107. [CrossRef]
45. Zheng, H.; Zhang, F.; Lin, X.; Huang, C.; Zhang, Y.; Li, Y.; Lin, J.; Chen, W.; Lin, X. MicroRNA-1225-5p inhibits proliferation and metastasis of gastric carcinoma through repressing insulin receptor substrate-1 and activation of β-catenin signaling. *Oncotarget* **2016**, *7*, 4647–4663. [CrossRef]
46. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2018.
47. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*; Springer: Berlin/Heidelberger, Germany, 2016.
48. Zhao, P.; Yu, B. On model selection consistency of Lasso. *J. Mach. Learn. Res.* **2006**, *7*, 2541–2563.
49. Bühlmann, P.; Van De Geer, S. *Statistics for High-dimensional Data: Methods, Theory and Applications*; Springer: Berlin/Heidelberger, Germany, 2011.
50. Vaart, A.W.; Wellner, J.A. *Weak Convergence and Empirical Processes: with Applications to Statistics*; Springer: Berlin/Heidelberger, Germany, 1996.
51. Talagrand, M. Sharper bounds for Gaussian and empirical processes. *Ann. Probab.* **1994**, *22*, 28–76. [CrossRef]