Graduate School of Economics, Hitotsubashi University
Discussion Paper Series No. 2020-01

FORWARD VARIABLE SELECTION FOR SPARSE ULTRA-HIGH DIMENSIONAL GENERALIZED VARYING COEFFICIENT MODELS

Toshio HONDA and Chien-Tong LIN

First version : January 2020
This version : February 2020

FORWARD VARIABLE SELECTION FOR SPARSE ULTRA-HIGH DIMENSIONAL GENERALIZED VARYING COEFFICIENT MODELS

Toshio Honda[1] and Chien-Tong Lin[2]

Hitotsubashi University[1] *and National Tsing Hua University*[2]

Abstract: In this paper we propose forward variable selection procedures for feature screening in ultra-high dimensional generalized varying coefficient models. We employ regression spline to approximate coefficient functions and then maximize the log-likelihood to select an additional relevant covariate sequentially. If we decide we do not significantly improve the log-likelihood any more by selecting any new covariates from our stopping rule, we terminate the forward procedures and give our estimates of relevant covariates. The effect of the size of the current model has been overlooked in stopping rules for sequential procedures for high-dimensional models. Our stopping rule takes into account the size of the current model suitably. Our forward procedures have screening consistency and some other desirable properties under regularity conditions. We also present the results of numerical studies to show their good finite sample performances.

Key words and phrases: B-spline basis, forward procedure, maximum likelihood, screening consistency, stopping rule, varying coefficient model.

1. Introduction

Suppose we have n i.i.d. observations $(Y_i, \boldsymbol{X}_i, Z_i), i = 1, \ldots, n$, where Y_i is a real response variable and $(\boldsymbol{X}_i, Z_i)$ is a covariate vector such that $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^T \in R^p$, $X_{i1} \equiv 1$, and Z_i is an index variable satisfying $Z_i \in [0, 1]$. Here we assume Y_i follows a high-dimensional sparse generalized varying coefficient model (GVCM) : the conditional density on $(\boldsymbol{X}_i, Z_i)$ w.r.t. some known σ-finite measure ν is given by

$$f(y|\boldsymbol{x}, z) = \exp\{y\boldsymbol{x}^T\boldsymbol{g}^*(z) - b(\boldsymbol{x}^T\boldsymbol{g}^*(z)) + c(y)\}, \tag{1.1}$$

where $b(\theta)$ and $c(y)$ are known functions and

$$\boldsymbol{g}^*(z) = (g_1^*(z), \ldots, g_p^*(z))^T = (g_j^*(z))_{j\in\{1,\ldots,p\}}$$

is a vector of p unknown smooth functions. We consider the setup where p is extremely large compared to n, but $\boldsymbol{g}^*(z)$ is sparse, i.e. most of $g_j^*(z)$ are irrelevant. We denote the set of relevant indecies by $\mathcal{M}$ and $|\mathcal{M}|$ is very small compared to p, where $|S|$ is the number of the elements of $S \subset \{1, \ldots, p\}$.

In such high-dimensional settings, even if the dimension of $\boldsymbol{X}_i$, p, is very large compared to the sample size n, the number of active or relevant covariates are usually much smaller than p and then we need some variable selection procedures for such high-dimensional datasets like the Lasso (e.g.

Tibshirani (1996)), the SCAD (e.g. Fan and Li (2001)), feature screening procedures based on marginal models or some association measure between the dependent variable and individual covariates (e.g. Fan and Lv (2008) and Zheng, Peng and He (2015)), and forward variable selection procedures (e.g. Wang (2009), Ing and Lai (2011), and Luo and Chen (2014)). Liu, Zhong and Li (2015) is an excellent review paper of feature screening procedures. Feature screening procedures are also called just screening procedures. When we use procedures with oracle properties like the SCAD, screening procedures are also necessary because the SCAD does not work for very large p due to its non-convex penalty. If p is extremely large, even the Lasso will not work and these kinds of screening procedures are still necessary. See Bühlmann and van de Geer (2011) and Hastie, Tibshirani and Wainwright (2015) for standard procedures and recent developments on high-dimensional issues.

In this paper, we deal with the ultra-high dimensional generalized varying coefficient model defined in (1.1), propose forward variable selection procedures for feature screening with a stopping rule, establish their desirable properties, and present the results of numerical studies to support the usefulness and significance of the proposed procedures. As far as we know, there is no forward selection screening procedure for generalized varying

coefficient models and our stopping rule has also independent significance.

Sure independence screening (SIS) and nonparametric independence screening (NIS) procedures (e.g. Fan and Lv (2008), Fan, Feng and Song (2011), and Fan, Ma and Dai (2014)) assume that marginal models reflect the true model faithfully. Model-free screening procedures choose an association measure between the response variable and covariates and assume that the association measure reflects the true model faithfully. Those procedures are very easy to understand and implement. However, they often miss relevant covariates and researchers usually employ those procedures iteratively without theory.

On the other hand, model-based forward selection procedures (e.g. Wang (2009), Ing and Lai (2011), Luo and Chen (2014), Cheng, Honda, and Zhang (2016), and Zheng, Hong and Li (2020)) are iterative procedures by nature and have desirable theoretical properties such as model-based assumptions and screening consistency. Especially, Zheng, Hong and Li (2020) deals with generalized linear models and is related to this paper. However, the authors considered only a sequentially conditional approach, not full maximization of the log-likelihood with respect to submodels, and they did not deal with generalized varying coefficient models. We deal with both ML-type and sequentially conditional procedures in a unified and much

simpler way in a more complicated, but still very important setup.

The varying coefficient model is one of the popular and useful structured nonparametric regression models and Fan and Zhang (2008) is an excellent review paper. There are many important papers on screening procedures for high-dimensional varying coefficient models. Fan, Ma and Dai (2014) and Cheng, Honda, and Zhang (2016) considered the NIS and forward screening procedures, respectively. Both of them deal with mean regression models. Xia, Yang and Li (2016) applied NIS to generalized varying coefficient models. Yang, Yang and Li (2020) proposed an approximated log-likelihood method as in their (2.8) for generalized varying coefficient models. The procedure can improve the log-likelihood function when their conditions on the observed high-dimensional Fisher information matrix are satisfied. However, there seems to be no result on how much the log-likelihood function is improved. Their procedure is not forward procedures and we do not rely on such approximations.

Information criteria are very important for variable selection in high-dimensional settings. In addition information criteria are often used as stopping rules as in Ing and Lai (2011), Luo and Chen (2014), Cheng, Honda, and Zhang (2016), and Zheng, Hong and Li (2020). EBIC is proposed in Chen and Chen (2008) and Chen and Chen (2012). The latter deals

with generalized linear models. The authors of Chen and Chen (2008) and Chen and Chen (2012) established model selection consistency among the models whose dimension is bounded. Hence the theory in Chen and Chen (2008) and Chen and Chen (2012) cannot be applied when the true model dimension increases to infinity. Kim and Jeon (2016) considers more general setups for parametric models and Lee, Noh and Park (2014) deals with varying coefficient quantile regression models.

When we consider information criteria for linear regression models, we have explicit expressions of the LS estimators and theoretical analysis is much easier than for our present model. However, there are some challenges for generalized linear models such as uniformity of estimators in wrong or misspecified models as seen in the proof of Lemma 2. Besides, the model dimension at the current step have critical effects on the asymptotics and we have to take the model dimension at the current step into consideration as in our stopping rule given in (2.2). This means that our information criterion and stopping rule have independent significance.

This paper is organized as follows. In Section 2, we present our forward screening procedures together with critical assumptions and main theoretical results. Then the results of our numerical studies are presented in Section 3. In Section 4, we describe technical assumptions and give the

proofs of the main theoretical results. We conclude this paper with Section 5. The proofs of technical lemmas and additional numerical results are given in the supplementary material.

We end this section with general notation used throughout the paper. In this paper, C, C_1, C_2, ..., are generic positive constants and their values may change from line to line. Note that $a_n \sim b_n$ means $C_1 < a_n/b_n < C_2$ and that $a \vee b$ and $a \wedge b$ stand for the maximum and the minimum of a and b, respectively. For a index set $S \subset \{1, \ldots, p\}$, S^c and $|S|$ stand for the complement and the number of the elements, respectively. For a vector a, $\|a\|$ and $\|a\|_1$ are the Euclidean and L_1 norms, respectively. We denote the maximum and minimum eigenvalues of a symmetric matrix A by $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$, respectively. For a matrix A and a vector a, A^T and a^T are their transposes.

2. Forward selection procedure

We begin with notation related to observations and our model. Then we describe our procedures, critical assumptions, and main theoretical results. We treat our ML-type and sequentially conditional procedures in a unified way.

2.1 Notation

Set $p_n = p \vee n$ and write $\mu(\theta) = b'(\theta)$ and $\sigma(\theta) = b''(\theta)$.

Let $(Y, \boldsymbol{X}, Z)$ have the same distribution as $(Y_i, \boldsymbol{X}_i, Z_i)$ and write

$$\boldsymbol{g}(z) = (g_1(z), \ldots, g_p(z))^T = (g_j(z))_{j\in\{1,\ldots,p\}},$$

$$\boldsymbol{X} = (X_1, \ldots, X_p)^T = (X_j)_{j\in\{1,\ldots,p\}}, \quad \text{and} \quad \boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^T = (X_{ij})_{j\in\{1,\ldots,p\}}.$$

We take $\mathrm{E}\{X_j\} = 0$ and $\mathrm{E}\{X_j^2\} = 1$ for $j = 2, \ldots, p$.

For $S \subset \{1, \ldots, p\}$, we write

$$\boldsymbol{g}_S(z) = (g_j(z))_{j\in S}, \quad \boldsymbol{X}_S = (X_j)_{j\in S}, \quad \text{and} \quad \boldsymbol{X}_{iS} = (X_{ij})_{j\in S}.$$

Define $\ell(y, \theta)$ by

$$\ell(y, \theta) = y\theta - b(\theta).$$

Then by sight abuse of notation, we denote the log-likelihood function for $S \subset \{1, \ldots, p\}$ by

$$\ell_n(\boldsymbol{X}_S^T \boldsymbol{g}_S(Z)) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, \boldsymbol{X}_{iS}^T \boldsymbol{g}_S(Z_i)).$$

Note that we omit Y on the LHS.

Next we define the quasi true coefficient function for $S \subset \{1, \ldots, p\}$ by

$$\boldsymbol{g}_S^*(z) = \arg\max_{\boldsymbol{g}_S(z)} \mathrm{E}\{\ell(Y, \boldsymbol{X}_S^T \boldsymbol{g}_S(Z))\},$$

where $\boldsymbol{g}_S^*(z) = (g_{jS}^*(z))_{j\in S}$ and we assume that $\boldsymbol{g}_S^*(z)$ has desirable properties as specified in Section 4. As for the definition of $\boldsymbol{g}_S^*(z)$, note that

$$\mathrm{E}\{\ell(Y, \boldsymbol{X}_S^T \boldsymbol{g}_S(Z))|Z = z\} \tag{2.1}$$

depends on a $|S|$-dimensional vector and $\boldsymbol{g}_S^*(z)$ is just the parameter maximizing (2.1) for that z. Hereafter we assume the model for S and $\boldsymbol{g}_S^*(z)$ are well defined.

We also define $h_{jS}^*(z)$ by

$$h_{jS}^*(z) = \arg\max_{h_{jS}(z)} \mathrm{E}\{\ell(Y, \boldsymbol{X}_S^T \boldsymbol{g}_S^*(Z) + X_j h_{jS}(Z))\}$$

for $j \notin S$.

We employ regression spline to estimate $\boldsymbol{g}_S^*(z)$ and replace $\boldsymbol{X}_S^T \boldsymbol{g}_S(Z)$ and $\boldsymbol{X}_{iS}^T \boldsymbol{g}_S(Z)$ with $\boldsymbol{W}_S^T \boldsymbol{\beta}_S$ and $\boldsymbol{W}_{iS}^T \boldsymbol{\beta}_S$, where $\boldsymbol{W}_S = \boldsymbol{X}_S \otimes B(Z) \in R^{|S|L}$, $\boldsymbol{W}_{iS} = \boldsymbol{X}_{iS} \otimes B(Z_i) \in R^{|S|L}$, and $\boldsymbol{\beta}_S = (\beta_j)_{j\in S}$ with $\beta_j \in R^L$. Note that $\otimes$ means the Kronecker product and that $B(z) = (B_1(z), \ldots, B_L(z))^T \in R^L$ is a suitable L-dimensional spline basis, e.g. the equispaced B-spline basis on $[0, 1]$. In this paper we use the linear or smoother B-spline basis. Schumaker (2007) is an excellent reference on spline functions.

We also write

$$\boldsymbol{W} = \boldsymbol{X} \otimes B(Z) \in R^{pL},\ \boldsymbol{W}_i = \boldsymbol{X}_i \otimes B(Z_i) \in R^{pL},\ \text{and } \boldsymbol{\beta} = (\beta_j)_{j\in\{1,\ldots,p\}}.$$

Then we assume that we can approximate $h^*_{jS}(z)$, $\boldsymbol{g}^*_S(z) = (g^*_{kS}(z))_{k\in S}$, and $\boldsymbol{g}^*(z) = (g^*_k(z))_{k\in\{1,\ldots,p\}}$ by using suitable η^*_{jS}, $\boldsymbol{\beta}^*_S = (\beta^*_{kS})_{k\in S}$, and $\boldsymbol{\beta}^* = (\beta^*_k)_{k\in\{1,\ldots,p\}}$, respectively as

$$h^*_{jS}(z) \approx \eta^{*T}_{jS} B(z), \quad g^*_{kS}(z) \approx \beta^{*T}_{kS} B(z) \quad \text{and} \quad g^*_k(z) \approx \beta^{*T}_k B(z), \quad \text{respectively.}$$

We explain the meaning of $\approx$ in more detail in Section 4.

Our procedure is based on the maximization of the log-likelihood function for $S \subset \{1,\ldots,p\}$. Thus we need to define relevant notation here.

Let $\widehat{\boldsymbol{g}}_S(z)$ be the quasi ML estimator of $\boldsymbol{g}^*_S(z)$ as given in

$$\widehat{\boldsymbol{g}}_S(z) = I_{|S|} \otimes B(z)^T \widehat{\boldsymbol{\beta}}_S,$$

where $\widehat{\boldsymbol{\beta}}_S = \underset{\boldsymbol{\beta}_S \in R^{|S|L}}{\arg\max}\, \ell_n(\boldsymbol{W}^T_S \boldsymbol{\beta}_S)$. The uniform properties of $\widehat{\boldsymbol{g}}_S(z)$ are important and will be investigated in Lemma 2 in Section 4.

As for $h^*_{jS}(z)$,

$$\widehat{h}_{jS}(z) = B(z)^T \widehat{\eta}_{jS},$$

where $\widehat{\eta}_{jS} = \underset{\eta_{jS} \in R^L}{\arg\max}\, \ell_n(\boldsymbol{W}^T_S \widehat{\boldsymbol{\beta}}_S + W^T_j \eta_{jS})$. The convergence rate of this $\widehat{h}_{jS}(z)$ is given in Lemma 5 in Section S3 in the supplementary material.

This $\widehat{h}_{jS}(z)$ is just an auxiliary technical tool for the ML-type forward regression procedure and does not appear in its algorithm. Very interestingly, we do not need the theoretical properties of $\widehat{h}_{jS}(z)$ at all in deriving

the theoretical properties of both of our procedures, even for the sequentially conditional screening procedure.

2.2 Forward selection procedures

We state our two forward variable selection procedures, the ML-type and sequentially conditional ones. We present their desirable properties later in Theorems 1-3.

First we take $S_0 = \{1\}$ and begin with $k = 1$.

(1) Put $S = S_{k-1}$. Then

$$j_k = \arg\max_{j \in S^c} \max_{\boldsymbol{\beta}_{S\cup\{j\}}} \ell_n(\boldsymbol{W}_{S\cup\{j\}}^T \boldsymbol{\beta}_{S\cup\{j\}}).$$

(2) Check if we have significantly improved $\ell_n(\boldsymbol{W}_{S_{k-1}}^T \widehat{\boldsymbol{\beta}}_{S_{k-1}})$ by adding j_k. Specifically, if we have with $S = S_{k-1}$,

$$\max_{j \in S^c} \max_{\boldsymbol{\beta}_{S\cup\{j\}}} \ell_n(\boldsymbol{W}_{S\cup\{j\}}^T \boldsymbol{\beta}_{S\cup\{j\}}) - \ell_n(\boldsymbol{W}_S^T \widehat{\boldsymbol{\beta}}_S) > L\xi_n |S| \log p_n/n, \tag{2.2}$$

set $S_k = S_{k-1} \cup \{j_k\}$ and go to **(1)**. If not, set $\widehat{\mathcal{M}} = S_{k-1}$ and end this algorithm.

The above is our ML-type forward regression procedure and is denoted by FR in Section 3.

By replacing j_k in **(1)** with

$$j_k = \arg\max_{j \in S^c} \max_{\eta_{jS}} \ell_n(\boldsymbol{W}_S^T \widehat{\boldsymbol{\beta}}_S + W_j^T \eta_{jS}) \tag{2.3}$$

and

$$\max_{j\in S^c}\max_{\boldsymbol{\beta}_{S\cup\{j\}}}\ell_n(\boldsymbol{W}_{S\cup\{j\}}^T\boldsymbol{\beta}_{S\cup\{j\}}) \text{ with } \max_{\boldsymbol{\beta}_{S\cup\{j_k\}}}\ell_n(\boldsymbol{W}_{S\cup\{j_k\}}^T\boldsymbol{\beta}_{S\cup\{j_k\}})$$

in **(2)**, we can define the sequentially conditional screening procedure, which is denoted by SC in Section 3. From a computational point of view, this is easier to implement.

Some comments on the stopping rule in (2.2) are in place. We define our information criterion as

$$-\ell_n(\boldsymbol{W}_S^T\widehat{\boldsymbol{\beta}}_S)+\frac{L|S|}{2}\frac{q_n\log p_n}{n}, \tag{2.4}$$

where $q_n=\xi_n|S|$ and ξ_n tends to ∞ slowly. From a theoretical point of view, any ξ_n tending to ∞ will work. We emphasize q_n's dependence on $|S|$.

We use this information criterion as our stopping rule. Actually, $|S|$ in (2.2) should be $(2|S|+1)/2$ from (2.4). However, we omitted $1/2$ just for simplicity. We have no theoretical suggestion for ξ_n in Theorem 2 below. Therefore we took $\xi_n=1$ in our numerical studies.

Choosing a suitable stopping rule is very challenging. This is because the true model size can increase to infinity and we have to pay some cost in the uniform convergence rate in S. Besides, our setup is nonlinear and nonparametric. Note that both Chen and Chen (2008) and Chen and Chen

(2012) assume the true model size is bounded and that they proved selection consistency among the models whose dimensions are bounded. The current model size can affect the asymptotics. However, when we go on to a larger model, the penalty increment of traditional information criteria is independent of the current model size such as $|S|$. Thus we have come up with the criterion in (2.4) to cope with these problems. We need to make some non-trivial modifications to the proof of Theorem 2 in Honda, Ing and Wu (2019) to derive our theoretical results on this criterion. If we deal with parametric models, we should remove L in (2.4).

2.3 Critical assumptions and main theorems

Next we present critical assumptions and main theoretical results which cover both of the proposed procedures. The other technical assumptions are relegated to Section 4.

The following assumption is similar to Assumption (E) in Zheng, Hong and Li (2020) and this assumption stipulates how large the signal is when $\mathcal{M} \not\subset S$. If the LHS of (2.5) is small for any $j \in S^c$, the remaining signal is sufficiently small and there may be no need of adding new covariates. In Theorem 1, we relate the LHS of (2.5) to the log-likelihood.

Assumption LB : When $\mathcal{M} \not\subset S$, there is a uniform lower bound as

$$|\mathrm{E}[\{Y - \mu(\boldsymbol{g}_S^{*T}\boldsymbol{X}_S)\}X_j\overline{h}_{jS}(Z)]| > \rho_{LB} \tag{2.5}$$

for some $j \in \mathcal{M} \cap S^c$ and $\overline{h}_{jS}(z)$ satisfying $\mathrm{E}\{\overline{h}_{jS}^2(Z)\} = 1$. This ρ_{LB} can depend on n and is closely related to K_n and p_n as seen in Theorems 1-3. This assumption is equivalent to

$$\mathrm{E}[\mathrm{E}\{(Y - \mu(\boldsymbol{g}_S^{*T}\boldsymbol{X}_S))X_j|Z\}^2]^{1/2} > \rho_{LB}.$$

We consider general setups and have to deal with uniformity in $S \subset \{1, \ldots, p\}$. Hence we need an upper bound on the size of possible S to have sufficiently small convergence rates of the estimators.

Assumption UB : We know an upper limit of $\mathcal{M}$, K_n, and our technical assumptions should hold for $S \subset \{1, \ldots, p\}$ such that $|S| \leq K_n$. We allow both $|\mathcal{M}|$ and K_n to increase to infinity. This K_n satisfies

$$\frac{K_n^{3/2}L}{\sigma_{min}\lambda_L}\sqrt{\frac{\log p_n}{n}} \to 0. \tag{2.6}$$

(2.6) implies we should have $K_n/n^{1/5} \to 0$. This upper limit K_n does not have to be tight w.r.t. $|\mathcal{M}|$, e.g. we allow $|\mathcal{M}|/K_n \to 0$.

Now we state our theoretical results. We present the proofs of these theorems in Section 4 together with Lemmas 1-4. These lemmas are proved in the supplementary material.

We have three critical parameters, K_n, p_n, and ρ_{LB}. There are trade-offs among them as in (2.6) in Assumption UB, (2.8), and (2.9) in Theorem 3.

Theorem 1 relates Assumption LB to possible improvement on the log-likelihood function.

Theorem 1. *Suppose that all the assumptions except for Assumption B(2) hold. Then with probability tending to 1, we have uniformly in k smaller than K_n,*

$$\max_{j\in S^c}\max_{\boldsymbol{\beta}_{S\cup\{j\}}} \ell_n(\boldsymbol{W}_{S\cup\{j\}}^T\boldsymbol{\beta}_{S\cup\{j\}}) - \ell_n(\boldsymbol{W}_S^T\widehat{\boldsymbol{\beta}}_S) \geq C_{LB}\rho_{LB}^2 \quad \textit{with } S = S_{k-1}$$

and

$$\max_{j\in S^c}\max_{\eta_{jS}} \ell_n(\boldsymbol{W}_S^T\widehat{\boldsymbol{\beta}}_S + W_j^T\eta_{jS}) - \ell_n(\boldsymbol{W}_S^T\widehat{\boldsymbol{\beta}}_S) \geq C_{LB}\rho_{LB}^2 \quad \textit{with } S = S_{k-1}$$

if $\mathcal{M} \not\subset S_{k-1}$ and $C_{LB}\rho_{LB}^2/2$ is larger than the RHS's of Lemmas 3 and 4 in Section 4. Note that C_{LB} is from Lemma 1 in Section 4.

Note that

$$\max_{\boldsymbol{\beta}_{S\cup\{j\}}} \ell_n(\boldsymbol{W}_{S\cup\{j\}}^T\boldsymbol{\beta}_{S\cup\{j\}}) \geq \max_{\eta_{jS}} \ell_n(\boldsymbol{W}_S^T\widehat{\boldsymbol{\beta}}_S + W_j^T\eta_{jS}). \tag{2.7}$$

If we have

$$\frac{K_n\sqrt{\log p_n}}{n^{2/5}} = o(\rho_{LB}^2), \tag{2.8}$$

our condition on the RHS's of Lemmas 3 and 4 is satisfied.

In Theorem 2, we establish the theoretical validity of our stopping rule. We have to prove that our procedures stop once we have $\mathcal{M} \subset S_k$. There are many and many S satisfying $\mathcal{M} \subset S$ and $|S|+1 \leq K_n$. This is different from the proof for model selection consistency and we cannot apply the standard arguments for information criteria straightforwardly. In addition recall we allow both $|\mathcal{M}|$ and K_n to tend to infinity as long as all the assumptions and conditions are satisfied. Therefore the arguments in Chen and Chen (2012) do not apply to out setup.

Theorem 2. *Suppose that all the conditions and assumptions in Theorem 1 and Assumption B(2) hold. Then the criterion defined in (2.4) works as a stopping rule in both of our forward selection algorithms. Specifically, as long as* $C_{LB}\rho_{LB}^2 > L\xi_n|S_k|\log p_n/n$ *and* $|S_k| < K_n$*, our algorithms do not stop until* $\mathcal{M} \subset S_k$ *with probability tending to 1. In addition, once* $\mathcal{M} \subset S_k$ *and* $|S_k| < K_n$*, we stop at this step with probability tending to 1.*

The inequality condition in Theorem 2 is much less restrictive since it is true if we have for some positive constant C_1,

$$C_{LB}\rho_{LB}^2 > C_1\xi_n K_n \frac{\log p_n}{n^{4/5}}.$$

According to Theorem 3, both of the proposed procedures enjoy the

screening consistency when ρ_{LB} in (2.5) is large enough.

Theorem 3. *Suppose that all the conditions and assumptions in Theorem 1 hold and set*

$$\Delta = \mathrm{E}\{\ell(Y, \boldsymbol{X}^T \boldsymbol{g}^*(Z))\} - \mathrm{E}\{\ell(Y, \boldsymbol{X}_{S_0}^T \boldsymbol{g}_{S_0}^*(Z))\}.$$

Then $\mathcal{M} \subset S_k$ for some $k \leq K_n$ with probability tending to 1 if

$$\Delta/(C_{LB}\rho_{LB}^2) < K_n - 2. \tag{2.9}$$

3. Numerical studies

In this section, we present our numerical studies. They consist of simulation studies in Subsection 3.1 and an application to the multiple myeloma (MM) data with $p = 44760$ in Subsection 3.2. We did all the computations by using R.

3.1 Simulation studies

We carried out simulation studies to evaluate the finite sample performances of the ML-type forward regression (FR) and sequentially conditional screening (SC) procedures in the high-dimensional GVCM and the results are compared with those of the nonparametric independence screening of Fan, Feng and Song (2011). Recall that the parametric version of the SC pro-

cedure was originally proposed for the generalized linear model in Zheng, Hong and Li (2020). In addition, the results are also compared with those of the group LASSO (gLASSO) and group SCAD (gSCAD) procedures. We implemented both of them by using R package *grpreg* of Breheny and Huang (2015).

We describe our simulation setups. When we generate the index variable $Z \in [0,1]$, the covariate vector $\boldsymbol{X} \in \mathbb{R}^p$ and the response variable Y, we follow the same spirit in the simulation settings of Yang, Yang and Li (2020). We first sample $(Z^*, \boldsymbol{X})^T$ from a $p+1$ dimensional normal distribution $N(0, \Sigma)$, where Σ is a $(p+1) \times (p+1)$ covariance matrix. Two commonly used covariance structures for Σ with parameter ρ are

$$\Sigma_1 : \sigma_{ij} = 1, \ \forall \ i = j; \sigma_{ij} = \rho, \ \forall \ i \neq j, \text{ and}$$

$$\Sigma_2 : \sigma_{ij} = \rho^{|i-j|}, \ \forall \ i \neq j,$$

and they correspond to the equi-correlated and the auto-correlated structures respectively. Here, we take multivariate normal $\boldsymbol{X}$ to check the robustness of our procedure in spite of Assumption X(1) in Section 4. Letting $\Phi(\cdot)$ be the cumulative density function of $N(0,1)$, we set $Z = \Phi(Z^*)$ so that the index variable is correlated with $\boldsymbol{X}$.

Given $\{Z, \boldsymbol{X}\}$, we consider two sets of coefficient functions

$$G_1 : \boldsymbol{g}_1(Z) = \boldsymbol{g}_2(Z) = \boldsymbol{g}_3(Z) = 2 + 2\sin^2(2\pi Z), \boldsymbol{g}_4(Z) = -3\rho\boldsymbol{g}_1(Z)$$

and

$$G_2 : \boldsymbol{g}_1(Z) = -(3 + 2\cos^2(0.5\pi Z)), \boldsymbol{g}_2(Z) = -3(1 + Z),$$
$$\boldsymbol{g}_3(Z) = (2 - Z)^2 + 2, \boldsymbol{g}_4(Z) = 3 + 2\sin^2(0.5\pi Z).$$

We generate the response variable Y following the high-dimensional GVCM in (1.1) by specifying the covariate structure (Σ) and coefficient function (G): (Σ_1, G_1), (Σ_1, G_2), (Σ_2, G_1) and (Σ_2, G_2). We write $\eta(Z, \boldsymbol{X}) = \boldsymbol{g}_1(Z)X_1 + \boldsymbol{g}_2(Z)X_2 + \boldsymbol{g}_3(Z)X_3 + \boldsymbol{g}_4(Z)X_4$. We implicitly include the intercept term as in the algorithm in Subsection 2.2. This means $S_0 = \{0\}$ with $X_0 \equiv 1$.

We deal with normal, logistic and Poisson regression models. In the normal regression models, Y follows a normal distribution with mean $\eta(Z, \boldsymbol{X})$ and variance 1; In the logistic regression models, Y follows a Bernoulli distribution with $P(Y = 1) = \exp(\eta(Z, \boldsymbol{X}))/(1 + \exp(\eta(Z, \boldsymbol{X})))$; In the Poisson regression models, Y follows a Poisson distribution with mean $\exp(\eta(Z, \boldsymbol{X}))$. In particular, we set smaller coefficients $G_1/4$ and $G_2/6$ to avoid large mean values in Poisson regression models. Besides, a kind of special care is necessary to the logistic regression models due to the separation problem (Heinze and Schemper, 2002). See Remark 1 at the end of

this subsection.

We use both the stopping rule (2.2) and the high-dimensional BIC (HBIC) of Yang, Yang and Li (2020) to terminate the iteration of FR and SC. The HBIC criterion is included for comparison although it is not theoretically justified yet for the present setup. In addition, we also give the results on FR and SC with no stopping rule ($K_n = 10$). Hereafter they are called FR.full and SC.full. In this section, we denote the criterion (2.4) as the likelihood information criterion (LIC) in our table. As illustrated by Cheng, Honda, and Zhang (2016), the stopping criterion based on EBIC or BIC will tend to stop the forward regression in varying coefficient model too early. Thus, we continue selection along the path of FR and SC algorithms until the stopping criterion is satisfied m times and they are denoted by FR.m and SC.m, respectively. In our simulation, we set $m = 1, \ldots, 5$. In Cheng, Honda, and Zhang (2016), they terminated their algorithm with $m = 5$.

We describe the parameter setting of algorithms and simulation designs here. The quadratic B-spline with $L = 5$ is employed in all the procedures. We consider $(n, p) = (200, 1000)$ and $(n, p) = (400, 1000)$ for the normal regression models, and consider $(n, p) = (300, 1000)$ and $(n, p) = (500, 1000)$ for the logistic and Poisson regression models. We take $\rho = 0.25$ and

$\rho = 0.5$ for Σ_1 and $\rho = 0.5$ for Σ_2. We select the top-$\lfloor (n/L)/\log(n/L) \rfloor$ ranked variables based on their marginal ranking, in the NIS procedure, where $\lfloor x \rfloor$ represents the integer part of number x, and we set $K_n = 10$ as the maximum iteration number for both the FR and SC procedures. We exploit 10-fold cross validation for tuning parameter selection for the group LASSO (gLASSO) and group SCAD (gSCAD) procedures and set $a = 3.7$ in the SCAD penalty as proposed in Fan and Li (2001).

We evaluate the performances of these procedures by the averaged number of true positives (TP), the averaged number of false positives (FP) and the proportion of attaining the sure screening result (Sure) over 200 replications.

We present the results for (Σ_1, G_1) and (Σ_2, G_2) in Tables 1-6 here and put the other tables in the supplementary material. Note that FR+LIC with some m is our proposed procedure. It is very important for screening procedures not to miss relevant covariates with small or moderate FP rates. We summarize our simulation results from this perspective. Our observations are as follows :

1. For the normal regression models, gSCAD and FR+LIC with $m = 1$ performed equally very well. However, m should be 3 or larger for the logistic and Poisson regression models. Specifically, gSCAD and FR+LIC

with $m = 3$ and 4 performed equally very well. FR+LIC with $m = 3$ has higher TP than gSCAD except only for the logistic regression models in Tables 1, 3 and 4. Note that FR's FP is reasonably small. FR+LIC with $m = 3$ has a lower TP, 3.02, than 4.00 for gSCAD for the logistic regression model in Table 3. However, for $m = 4$, we have 4.00 vs. 4.00(gSCAD). In Tables 1 and 4, the differences between gSCAD and FR+LIC with $m = 3$ are much smaller. FR+LIC with $m = 4$ has the highest TP among the three and FR.full has the highest TP in all the tables.

2. FR performed better than SC and there was no significant difference between them. Thus we may be able to use SC as an alternative for extremely large p although full likelihood maximization w.r.t. submodels may be desirable. HBIC also performed almost as well as LIC although HBIC is not theoretically justified as a stopping rule yet, either.

3. NIS does not seem to work well as a screening procedure for GVCM in terms of poor TP and FP. gLASSO did not perform well with very large FP.

In conclusion, we recommend to use FR+LIC with $m = 3$ or 4. We can adjust the choice of m depending on the situation. We repeat that we should not miss relevant covariates at the stage of screening.

Remark 1. Note that TP + FP are less than $K_n = 10$ in the FR.full

and SC.full rows of the results for the logistic regression models. This is because we stopped the FR and SC procedures early due to the separation problem (Heinze and Schemper, 2002). The separation problem leads to non-convergent coefficient estimates and unbounded log-likelihood functions. When the separation occurred, we used the Firth's bias reduction and terminated the iteration since variables in the selected set was able to perfectly separate the binary response and thus it was not necessary to include any more variables. We also applied this principle to FR.m and SC.m.

3.2 Real data analysis

We apply our proposed method to the analysis of the multiple myeloma (MM) data in Mulligan et al. (2007). One of the main purposes of the study was to identify genes that are relevant to the clinical response of multiple myeloma. In this MM data, 264 patients were subject to replicate gene expression profiling using the Affymetrix 133A/B microarray and 44760 non-overlapped genes were measured for each patient. These MM patients are originally classified into six categories according to the clinical response. Among the six categories, we merge the complete response (CR) and the partial response (PR) to one category and the other categories to

Table 1: Simulation results for the design (Σ_1, G_1) with $\rho = 0.25$, $(n, p) = (200, 1000)$ for the normal model, and $(n, p) = (300, 1000)$ for logistic and Poisson models.

Screen	Stop	m	Normal			Logistic			Poisson		
			Sure	TP	FP	Sure	TP	FP	Sure	TP	FP
NIS	-	-	0	3	7.00	0.00	3.00	11.01	0.00	2.63	11.37
gLASSO	-	-	1	4	26.73	1.00	4.00	55.59	0.19	3.19	23.58
gSCAD	-	-	1	4	3.30	0.96	3.96	0.28	0.10	3.08	0.92
FR.full	-	-	1	4	6.00	1.00	4.00	6.00	0.30	3.31	6.70
SC.full	-	-	1	4	6.00	1.00	4.00	5.87	0.26	3.25	6.75
FR	LIC	1	1	4	0.00	0.00	1.70	0.00	0.00	3.00	0.02
		2	1	4	1.00	0.00	2.70	0.00	0.22	3.22	0.80
		3	1	4	2.00	0.70	3.69	0.00	0.28	3.29	1.74
		4	1	4	3.00	1.00	4.00	0.70	0.29	3.29	2.73
		5	1	4	4.00	1.00	4.00	1.70	0.30	3.30	3.73
	HBIC	1	1	4	0.00	0.00	1.00	0.00	0.00	3.00	0.02
		2	1	4	1.00	0.00	2.00	0.00	0.22	3.22	0.80
		3	1	4	2.00	0.00	3.00	0.00	0.28	3.29	1.74
		4	1	4	3.00	1.00	4.00	0.00	0.29	3.29	2.73
		5	1	4	4.00	1.00	4.00	1.00	0.30	3.30	3.72
SC	LIC	1	1	4	0.00	0.00	1.70	0.00	0.00	2.97	0.16
		2	1	4	1.00	0.00	2.70	0.00	0.16	3.13	1.00
		3	1	4	2.00	0.70	3.69	0.00	0.20	3.19	1.94
		4	1	4	3.00	0.99	3.99	0.71	0.22	3.21	2.92
		5	1	4	4.00	1.00	4.00	1.70	0.24	3.23	3.90
	HBIC	1	1	4	0.00	0.00	1.00	0.00	0.00	2.98	0.16
		2	1	4	1.00	0.00	2.00	0.00	0.16	3.14	1.00
		3	1	4	2.00	0.00	3.00	0.00	0.20	3.19	1.95
		4	1	4	3.00	0.99	3.99	0.01	0.22	3.21	2.92
		5	1	4	4.00	1.00	4.00	1.00	0.24	3.23	3.91

Table 2: Simulation results for the design (Σ_1, G_1) with $\rho = 0.25$, $(n, p) = (400, 1000)$ for the normal model, and $(n, p) = (500, 1000)$ for logistic and Poisson models.

Screen	Stop	m	Normal			Logistic			Poisson		
			Sure	TP	FP	Sure	TP	FP	Sure	TP	FP
NIS	-	-	0	3	15.00	0.00	3.00	18.00	0.00	2.86	18.14
gLASSO	-	-	1	4	14.31	1.00	4.00	77.30	0.42	3.42	24.69
gSCAD	-	-	1	4	1.58	1.00	4.00	0.68	0.30	3.29	0.81
FR.full	-	-	1	4	6.00	1.00	4.00	6.00	0.58	3.58	6.42
SC.full	-	-	1	4	6.00	1.00	4.00	5.96	0.50	3.50	6.50
FR	LIC	1	1	4	0.00	0.00	2.00	0.00	0.00	3.00	0.00
		2	1	4	1.00	0.00	3.00	0.00	0.42	3.42	0.58
		3	1	4	2.00	1.00	4.00	0.00	0.50	3.50	1.50
		4	1	4	3.00	1.00	4.00	1.00	0.54	3.54	2.46
		5	1	4	4.00	1.00	4.00	2.00	0.56	3.56	3.44
	HBIC	1	1	4	0.00	0.00	1.00	0.00	0.00	3.00	0.00
		2	1	4	1.00	0.00	2.00	0.00	0.42	3.42	0.58
		3	1	4	2.00	0.01	3.01	0.00	0.50	3.50	1.50
		4	1	4	3.00	1.00	4.00	0.01	0.54	3.54	2.46
		5	1	4	4.00	1.00	4.00	1.01	0.56	3.56	3.44
SC	LIC	1	1	4	0.00	0.00	2.00	0.00	0.00	3.00	0.02
		2	1	4	1.00	0.00	3.00	0.00	0.32	3.31	0.70
		3	1	4	2.00	1.00	4.00	0.00	0.40	3.40	1.61
		4	1	4	3.00	1.00	4.00	1.00	0.44	3.44	2.58
		5	1	4	4.00	1.00	4.00	2.00	0.48	3.48	3.54
	HBIC	1	1	4	0.00	0.00	1.00	0.00	0.00	3.00	0.02
		2	1	4	1.00	0.00	2.00	0.00	0.32	3.31	0.70
		3	1	4	2.00	0.02	3.02	0.00	0.40	3.40	1.61
		4	1	4	3.00	1.00	4.00	0.02	0.44	3.44	2.58
		5	1	4	4.00	1.00	4.00	1.01	0.48	3.48	3.54

Table 3: Simulation results for the design (Σ_1, G_1) with $\rho = 0.5$, $(n, p) = (200, 1000)$ for the normal model, and $(n, p) = (300, 1000)$ for logistic and Poisson models.

Screen	Stop	m	Normal			Logistic			Poisson		
			Sure	TP	FP	Sure	TP	FP	Sure	TP	FP
NIS	-	-	0	3	7.00	0.00	2.99	11.01	0.00	2.64	11.36
gLASSO	-	-	1	4	26.09	1.00	4.00	56.67	0.42	3.42	21.12
gSCAD	-	-	1	4	2.88	1.00	4.00	0.16	0.37	3.31	1.62
FR.full	-	-	1	4	6.00	1.00	4.00	6.00	0.82	3.83	6.17
SC.full	-	-	1	4	6.00	1.00	4.00	5.62	0.60	3.59	6.41
FR	LIC	1	1	4	0.00	0.00	1.01	0.00	0.00	2.97	0.06
		2	1	4	1.00	0.00	2.02	0.00	0.70	3.68	0.34
		3	1	4	2.00	0.02	3.02	0.00	0.76	3.76	1.26
		4	1	4	3.00	1.00	4.00	0.02	0.78	3.78	2.25
		5	1	4	4.00	1.00	4.00	1.01	0.79	3.79	3.23
	HBIC	1	1	4	0.00	0.00	1.00	0.00	0.03	3.01	0.06
		2	1	4	1.00	0.00	2.00	0.00	0.70	3.70	0.36
		3	1	4	2.00	0.00	3.00	0.00	0.76	3.77	1.30
		4	1	4	3.00	1.00	4.00	0.00	0.78	3.78	2.29
		5	1	4	4.00	1.00	4.00	1.00	0.79	3.79	3.27
SC	LIC	1	1	4	0.00	0.00	1.01	0.00	0.00	2.85	0.31
		2	1	4	1.00	0.00	2.02	0.00	0.36	3.27	0.89
		3	1	4	2.00	0.02	3.02	0.00	0.47	3.42	1.74
		4	1	4	3.00	1.00	4.00	0.02	0.54	3.50	2.66
		5	1	4	4.00	1.00	4.00	1.02	0.57	3.55	3.59
	HBIC	1	1	4	0.00	0.00	1.00	0.00	0.00	2.91	0.32
		2	1	4	1.00	0.00	2.00	0.00	0.40	3.35	0.91
		3	1	4	2.00	0.00	3.00	0.00	0.48	3.44	1.82
		4	1	4	3.00	1.00	4.00	0.00	0.55	3.51	2.75
		5	1	4	4.00	1.00	4.00	1.00	0.57	3.55	3.69

Table 4: Simulation results for the design (Σ_1, G_1) with $\rho = 0.5$, $(n, p) = (400, 1000)$ for the normal model, and $(n, p) = (500, 1000)$ for logistic and Poisson models.

Screen	Stop	m	Normal			Logistic			Poisson		
			Sure	TP	FP	Sure	TP	FP	Sure	TP	FP
NIS	-	-	0	3	15.00	0.00	3.00	18.00	0.00	2.81	18.20
gLASSO	-	-	1	4	10.74	1.00	4.00	77.83	0.70	3.70	18.52
gSCAD	-	-	1	4	0.80	1.00	4.00	0.31	0.70	3.68	1.01
FR.full	-	-	1	4	6.00	1.00	4.00	6.00	1.00	4.00	6.00
SC.full	-	-	1	4	6.00	1.00	4.00	5.97	0.90	3.90	6.09
FR	LIC	1	1	4	0.00	0.00	2.10	0.00	0.03	3.03	0.02
		2	1	4	1.00	0.28	3.15	0.00	0.94	3.94	0.11
		3	1	4	2.00	0.86	3.87	0.28	0.98	3.98	1.07
		4	1	4	3.00	1.00	4.00	1.15	1.00	4.00	2.06
		5	1	4	4.00	1.00	4.00	2.15	1.00	4.00	3.06
	HBIC	1	1	4	0.00	0.00	1.00	0.00	0.10	3.10	0.02
		2	1	4	1.00	0.00	2.05	0.00	0.94	3.94	0.18
		3	1	4	2.00	0.08	3.08	0.00	0.98	3.98	1.14
		4	1	4	3.00	1.00	4.00	0.08	1.00	4.00	2.12
		5	1	4	4.00	1.00	4.00	1.08	1.00	4.00	3.12
SC	LIC	1	1	4	0.00	0.00	1.92	0.00	0.02	3.02	0.08
		2	1	4	1.00	0.24	3.08	0.00	0.78	3.77	0.32
		3	1	4	2.00	0.84	3.84	0.24	0.86	3.86	1.25
		4	1	4	3.00	1.00	4.00	1.08	0.90	3.90	2.20
		5	1	4	4.00	1.00	4.00	2.08	0.90	3.90	3.17
	HBIC	1	1	4	0.00	0.00	1.00	0.00	0.08	3.08	0.08
		2	1	4	1.00	0.00	2.10	0.00	0.78	3.77	0.38
		3	1	4	2.00	0.14	3.13	0.00	0.86	3.86	1.30
		4	1	4	3.00	1.00	4.00	0.14	0.90	3.90	2.25
		5	1	4	4.00	1.00	4.00	1.14	0.90	3.90	3.22

Table 5: Simulation results for the design (Σ_2, G_2) with $\rho = 0.5$, $(n, p) = (200, 1000)$ for the normal model, and $(n, p) = (300, 1000)$ for logistic and Poisson models.

Screen	Stop	m	Normal			Logistic			Poisson		
			Sure	TP	FP	Sure	TP	FP	Sure	TP	FP
NIS	-	-	0	2.96	7.04	0.00	2.94	11.05	0.00	2.46	11.54
gLASSO	-	-	1	4.00	20.23	1.00	4.00	74.03	1.00	4.00	41.24
gSCAD	-	-	1	4.00	2.13	0.88	3.88	0.76	0.92	3.86	0.92
FR.full	-	-	1	4.00	6.00	1.00	4.00	6.00	1.00	4.00	6.00
SC.full	-	-	1	4.00	6.00	1.00	4.00	5.92	0.94	3.90	6.11
FR	LIC	1	1	4.00	0.00	0.00	2.00	0.00	0.19	2.75	0.00
		2	1	4.00	1.00	0.00	3.00	0.00	0.62	3.62	0.19
		3	1	4.00	2.00	1.00	4.00	0.00	1.00	4.00	0.80
		4	1	4.00	3.00	1.00	4.00	1.00	1.00	4.00	1.80
		5	1	4.00	4.00	1.00	4.00	2.00	1.00	4.00	2.81
	HBIC	1	1	4.00	0.00	0.00	1.00	0.00	0.40	2.88	0.00
		2	1	4.00	1.00	0.00	2.00	0.00	0.90	3.90	0.40
		3	1	4.00	2.00	0.00	3.00	0.00	1.00	4.00	1.29
		4	1	4.00	3.00	1.00	4.00	0.00	1.00	4.00	2.29
		5	1	4.00	4.00	1.00	4.00	1.00	1.00	4.00	3.29
SC	LIC	1	1	4.00	0.00	0.00	2.00	0.00	0.16	2.62	0.02
		2	1	4.00	1.00	0.00	2.98	0.02	0.52	3.36	0.34
		3	1	4.00	2.00	0.98	3.96	0.04	0.84	3.77	0.93
		4	1	4.00	3.00	0.98	3.97	1.03	0.92	3.86	1.84
		5	1	4.00	4.00	0.98	3.98	2.02	0.94	3.88	2.82
	HBIC	1	1	4.00	0.00	0.00	1.00	0.00	0.32	2.73	0.02
		2	1	4.00	1.00	0.00	2.00	0.00	0.76	3.60	0.50
		3	1	4.00	2.00	0.00	2.98	0.02	0.88	3.81	1.34
		4	1	4.00	3.00	0.98	3.96	0.04	0.92	3.86	2.29
		5	1	4.00	4.00	0.98	3.97	1.03	0.94	3.88	3.27

Table 6: Simulation results for the design (Σ_2, G_2) with $\rho = 0.5$, $(n, p) = (400, 1000)$ for the normal model, and $(n, p) = (500, 1000)$ for logistic and Poisson models.

Screen	Stop	m	Normal			Logistic			Poisson		
			Sure	TP	FP	Sure	TP	FP	Sure	TP	FP
NIS	-	-	0	3	15.00	0.00	3.00	18.00	0.00	2.84	18.16
gLASSO	-	-	1	4	2.38	1.00	4.00	102.50	1.00	4.00	45.38
gSCAD	-	-	1	4	0.17	0.98	3.98	0.20	1.00	4.00	0.29
FR.full	-	-	1	4	6.00	1.00	4.00	6.00	1.00	4.00	6.00
SC.full	-	-	1	4	6.00	1.00	4.00	5.96	1.00	4.00	6.00
FR	LIC	1	1	4	0.00	0.00	2.01	0.00	0.96	3.96	0.00
		2	1	4	1.00	0.06	3.06	0.00	1.00	4.00	0.96
		3	1	4	2.00	1.00	4.00	0.06	1.00	4.00	1.97
		4	1	4	3.00	1.00	4.00	1.05	1.00	4.00	2.96
		5	1	4	4.00	1.00	4.00	2.06	1.00	4.00	3.96
	HBIC	1	1	4	0.00	0.00	1.53	0.00	0.96	3.93	0.00
		2	1	4	1.00	0.42	2.94	0.00	1.00	4.00	0.96
		3	1	4	2.00	0.88	3.88	0.42	1.00	4.00	1.97
		4	1	4	3.00	1.00	4.00	1.29	1.00	4.00	2.96
		5	1	4	4.00	1.00	4.00	2.29	1.00	4.00	3.96
SC	LIC	1	1	4	0.00	0.00	2.02	0.00	0.94	3.92	0.00
		2	1	4	1.00	0.08	3.08	0.00	0.98	3.97	0.96
		3	1	4	2.00	1.00	4.00	0.08	0.99	3.99	1.94
		4	1	4	3.00	1.00	4.00	1.07	1.00	4.00	2.93
		5	1	4	4.00	1.00	4.00	2.08	1.00	4.00	3.93
	HBIC	1	1	4	0.00	0.00	1.67	0.00	0.94	3.88	0.00
		2	1	4	1.00	0.55	3.22	0.00	0.99	3.98	0.96
		3	1	4	2.00	0.90	3.90	0.55	1.00	4.00	1.94
		4	1	4	3.00	1.00	4.00	1.45	1.00	4.00	2.94
		5	1	4	4.00	1.00	4.00	2.45	1.00	4.00	3.94

another category. Hence we have only two categories. 90 patients are in the CR+PR category. In brief, CR and PR requires at least 100% and 50% decrease in paraprotein, respectively, and CR+PR represents high decrease in paraprotein. This group merging is also employed in the analysis of Mulligan et al. (2007). We select relevant genes to classify patients into the two categories, CR+PR and the others, by exploiting our proposed screening procedure for GVCM, where we adopt AGE as the index variable. All the genes are normalized with mean 0 and variance 1.

We consider both the logistic regression model and the varying coefficient logistic regression model with AGE as the index variable in our analysis. We use the quadratic spline with $L = 5$. For the naive logistic regression model, we apply the sure independence screening and the sequentially conditional screening, which are denoted by SIS and nSC respectively. For the varying coefficient logistic regression model, we consider NIS and SC. The latter one is one of the proposed procedures in this paper. 44760 is too large for FR and besides SC performed almost as well in the simulation studies. Therefore we consider only SC here.

To measure the prediction performances, we compare the models with top-10 genes selected by these methods. Note that we select 10 covariates for each case and don't use any stopping rules. We report the area under

the curve (AUC) and the leave-one-out prediction error. The latter one is defined as the mean of the absolute difference between the predicted probability and the response. The results are reported in Table 7. The table implies that using the varying coefficient model (NIS/SC) leads to larger AUC than using the naive logistic regression model (SIS/nSC), and using SC yields the smallest prediction error among all methods. This implication indicates that genes has dynamic impacts on the clinical response through AGE. Further, SC with 10 covariates on hand performs the best in this data and the AUC is very close to 1 .

Next we apply SC with the stopping rule to the varying coefficient logistic model. We take $m = 4$ in LIC as suggested in our simulation, and then four genes are selected by SC+LIC. The coefficient functions of the selected covariates are illustrated in Figure 1 and the corresponding AUC is around 0.85. These four genes are good candidates for further biological research and deserve more attention in conjunction with patients' age.

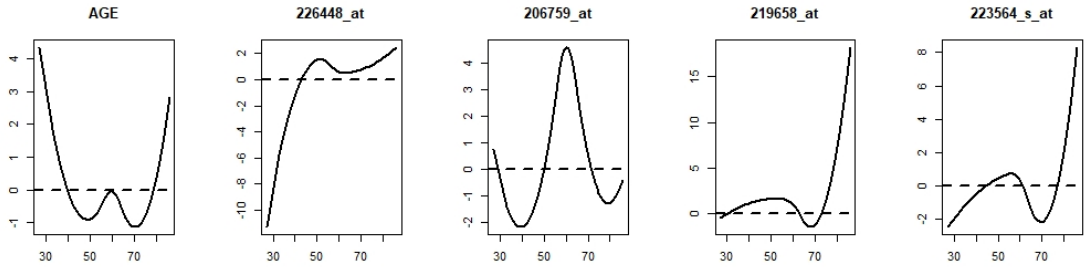
4. Assumptions and proofs

In this section, we describe technical assumptions first and state technical lemmas that are verified in the supplementary material. Finally we prove Theorems 1-3.

Table 7: AUC and leave-one-out prediction error.

	SIS	NIS	nSC	SC
Prediction error	0.36	0.49	0.24	0.22
AUC	0.78	0.80	0.91	0.99

Figure 1: The AGE versus β(AGE) plot for the index variable AGE and 4 genes selected by SC in the MM dataset.



4.1 Technical assumptions

First we state technical assumptions. Note that all S in the following technical assumptions satisfy $|S| \leq K_n$. All the constants in the assumptions are fixed except for L. We need those assumptions because we consider generalized varying coefficient models and have to deal with $b(\boldsymbol{X}_S^T \boldsymbol{g}_S(Z))$, $b(\boldsymbol{W}_S^T \boldsymbol{\beta}_S)$, $\mu(\boldsymbol{X}_S^T \boldsymbol{g}_S(Z))$, $\mu(\boldsymbol{W}_S^T \boldsymbol{\beta}_S)$, $\sigma(\boldsymbol{X}_S^T \boldsymbol{g}_S(Z))$, and $\sigma(\boldsymbol{W}_S^T \boldsymbol{\beta}_S)$. If $\boldsymbol{X}_S^T \boldsymbol{g}_S(Z)$ or $\boldsymbol{W}_S^T \boldsymbol{\beta}_S$ take very small or large values, we have to put stringent assump-

tions on $b(\theta)$, $\mu(\theta)$, and $\sigma(\theta)$ and this will make the setup and proofs very complicated. Therefore we impose Assumptions X(1) and CF(1) as Assumptions (A) and (B) in Zheng, Hong and Li (2020).

Assumption X :

(1) $|X_j| < C_{X1}$ uniformly in j for some positive C_{X1}.

(2) $C_{X2} < \mathrm{E}\{X_j^2|Z\}$ uniformly in j for some positive constant C_{X2}.

We need Assumption CF(2) to approximate coefficient functions sufficiently by employing the B-spline basis uniformly in S and $j \in S$.

Assumption CF :

(1) $\max_z \|\boldsymbol{g}_S^*(z)\|_1 \leq C_g$ uniformly in S and $\max_z |h_{jS}^*(z)| < C_h$ uniformly in $j \in S^c$ and S for some positive constants C_g and C_h.

(2) We take $L \sim n^{1/5}$. Then uniformly in S,

$$\max_z \|\boldsymbol{g}_S^*(z) - I_{|S|} \otimes B^T(z)\boldsymbol{\beta}_S^*\|_1 < \frac{C_{A1}}{L^2}$$

and uniformly in $j \notin S$ and S,

$$\max_z |h_{jS}^*(z) - B^T(z)\eta_{jS}^*| < \frac{C_{A2}}{L^2}$$

for some positive constants C_{A1} and C_{A2}.

Then we have $|\boldsymbol{X}_S^T \boldsymbol{g}_S^*(z)| < C_{X1}C_g$ and related variables are in a sufficiently large bounded interval, say $[\theta_L, \theta_U]$, with probability tending to 1.

A sufficient condition for the former of Assumption CF(2) is

$$\sup_z \{\|g_S^*(z)\|_1 + \|g_S^{*\prime}(z)\|_1 + \|g_S^{*(2)}(z)\|_1\} < C_1$$

for some positive constant C_1 uniformly in S by Corollary 6.26 of Schumaker (2007).

Assumption B :

(1) $b(\theta)$ is twice differentiable on $[\theta_L, \theta_U]$ and

$$|\mu(\theta)| < \mu_{max} \quad \text{and} \quad \sigma_{min} < \sigma(\theta) < \sigma_{max}$$

for $\theta \in [\theta_L, \theta_U]$. μ_{max}, σ_{min}, and σ_{max} are positive constants constants.

(2) $b(\theta)$ is also three times differentiable and $|b^{(3)}(\theta)| < \sigma'_{max}$ for $\theta \in [\theta_L, \theta_U]$. σ'_{max} is a positive constant.

Assumption B(1) has the following implication : As long as $K_n^{3/2} L \sqrt{\log p_n / n} \to 0$, we can fix a sufficiently large bounded interval $[\theta_L, \theta_U]$ depending on Assumptions X(1) and CF(1)(2) and there are suitable μ_{max}, σ_{min}, and σ_{max} for this $[\theta_L, \theta_U]$. Note that this $[\theta_L, \theta_U]$ contains all the related variables with probability tending to 1.

As in Lee, Noh and Park (2014) and other papers on varying coefficient models, we impose Assumption W. These authors assume the population version and derive the sample version given in our Assumption W in their technical lemmas. However, such arguments are very common and we di-

rectly use the sample version for simplicity of presentation.

Assumption W : For some positive constants λ_L and λ_U,

$$\frac{\lambda_L}{L} < \lambda_{\min}\Big(n^{-1}\sum_{i=1}^{n} \boldsymbol{W}_{iS}\boldsymbol{W}_{iS}^T\Big) \leq \lambda_{\max}\Big(n^{-1}\sum_{i=1}^{n} \boldsymbol{W}_{iS}\boldsymbol{W}_{iS}^T\Big) < \frac{\lambda_U}{L}$$

uniformly in S with probability tending to 1.

The following assumption is also a standard one for varying coefficient models. If we do not have any observations on some part of $[0, 1]$, we cannot identify or estimate coefficient functions there.

Assumption Z : For some positive constants C_{Z1} and C_{Z2}, $C_{Z1} < f_Z(z) < C_{Z2}$ on $[0, 1]$, where $f_Z(z)$ is the density of Z.

When we prove necessary uniform properties of $\widehat{\boldsymbol{\beta}}_S$ in S, we exploit standard arguments based on Bernstein's inequality (e.g. Lemma 2.2.11 of van der Vaart and Wellner (1996)). We use the next assumption when we apply Bernstein's inequality in the proofs of Theorem 3 and Lemma 2. This kind assumption is a standard one in the literature, e.g. Zheng, Hong and Li (2020).

Assumption E : Write $e = Y - \mu(\boldsymbol{X}^T\boldsymbol{g}^*(Z))$ and $e_i = Y_i - \mu(\boldsymbol{X}_i^T\boldsymbol{g}^*(Z_i))$. Then for some positive constant M_e,

$$\mathrm{E}\{|e|^m|\boldsymbol{X}, Z\} \leq m!M_e^m, \quad m \geq 2$$

uniformly in $\boldsymbol{X}$ and Z.

4.2 Technical lemmas

We present technical lemmas necessary to the proofs of Theorems 1-3. We prove these lemmas in the supplementary material.

In the lemmas, we assume $|S| + 1 \leq K_n$. Note $C_1, C_2, \ldots$ are generic positive constants which are large enough. They may depend on fixed positive constants in the assumptions, but are independent of n. In these lemmas, we don't suppress the constants, λ_L, λ_U, σ_{min}, and σ_{max} to see their potential effects on the lemmas.

Lemma 1 evaluates how much $\mathrm{E}\{\ell(Y, \boldsymbol{X}_S^T \boldsymbol{g}_S^*(Z))\}$ increases when $\mathcal{M} \not\subset S$.

Lemma 1. *Suppose Assumptions LB, X(1)(2), CF(1), B(1), and Z hold. Then we have (i) and (ii).*

(i) Assumption LB implies that there is a positive constant C_1 such that

$$\mathrm{E}\{h_{jS}^{*2}(Z)\} \geq C_1 \frac{\rho_{LB}^2}{\sigma_{max}^2}$$

for some $j \in \mathcal{M} \cap S^c$ if $\mathcal{M} \not\subset S$.

(ii) There is a positive constant C_2 such that for any $j \in \mathcal{M} \cap S^c$,

$$\mathrm{E}\{\ell(Y, \boldsymbol{X}_S^T \boldsymbol{g}_S^*(Z) + X_j h_{jS}^*(Z))\} - \mathrm{E}\{\ell(Y, \boldsymbol{X}_S^T \boldsymbol{g}_S^*(Z))\} \geq C_2 \sigma_{min} \mathrm{E}\{h_{jS}^{*2}(Z)\}.$$

We denote $C_1 C_2 \sigma_{min}/\sigma_{max}^2$ from (i) and (ii) by $2C_{LB}$ in Theorems 1-3.

We need the uniform convergence rate of $\widehat{\boldsymbol{\beta}}_S$ even for wrong or misspecified models ($\mathcal{M} \not\subset S$). Then we have to deal with all $S \subset \{1, \ldots, p\}$ satisfying $|S| \leq K_n$. This is why there is $|S| = |S|^{1/2}|S|^{1/2}$ on the RHS in (4.1). The first $|S|^{1/2}$ is from the size of S and the second $|S|^{1/2}$ is from the uniformity in S. This latter $|S|^{1/2}$ is an additional cost for dealing with the uniformity. The B-spline basis dimension L in the convergence rate is a standard one.

Lemma 2. *Suppose Assumptions X(1)(2), CF(1)(2), B(1), W, Z, and E hold. Then for some positive constant C_1, we have*

$$\|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^*\| \leq C_1 \frac{L|S|}{\sigma_{min}\lambda_L}\sqrt{\frac{\log p_n}{n}} \tag{4.1}$$

uniformly in S with probability tending to 1.

By Lemma 2, we have $|(\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^*)^T \boldsymbol{W}_{iS}| \to 0$ uniformly in S when Assumption UB is satisfied. This and Assumptions X(1) and CF(1) imply that all the related variables are in some sufficiently large bounded interval.

In the next lemma, we evaluate the difference between the log-likelihood for $\widehat{\boldsymbol{\beta}}_S$ and its theoretical counterpart. Lemma 4 deals with a similar problem.

Lemma 3. *For some positive constants C_1 and C_2, we have*

$$\begin{aligned} &|\ell_n(\boldsymbol{W}_S^T\widehat{\boldsymbol{\beta}}_S) - \mathrm{E}\{\ell(Y, \boldsymbol{X}_S^T\boldsymbol{g}_S^*(Z))\}| \\ &\leq \; C_1\Big(\frac{\lambda_U^{1/2}|S|}{\sigma_{min}\lambda_L}\sqrt{\frac{\log p_n}{n^{4/5}}} + \sqrt{\frac{|S|\log p_n}{n}}\Big) + C_2\frac{1}{L^2} \end{aligned} \tag{4.2}$$

uniformly in S with probability tending to 1.

The first term is dominant on the RHS and the RHS tends to 0 by Assumption UB.

Lemma 4. *For some positive constants C_1 and C_2, we have*

$$\begin{aligned} &|\ell_n(\boldsymbol{W}_S^T\widehat{\boldsymbol{\beta}}_S + W_j^T\eta_{jS}^*) - \mathrm{E}\{\ell(Y, \boldsymbol{X}_S^T\boldsymbol{g}_S^*(Z) + X_j h_{jS}^*(Z))\}| \\ &\leq \; C_1\Big(\frac{\lambda_U^{1/2}|S|}{\sigma_{min}\lambda_L}\sqrt{\frac{\log p_n}{n^{4/5}}} + \sqrt{\frac{|S|\log p_n}{n}}\Big) + C_2\frac{1}{L^2}. \end{aligned} \tag{4.3}$$

uniformly in S with probability tending to 1.

The first term is dominant on the RHS and the RHS tends to 0 by Assumption UB.

4.3 Proofs of Theorems 1-3

We prove Theorems 1-3 by employing Lemmas 1-4. We prove Theorem 1 first. We verify Theorem 2 by following the proof of Theorem 2 of Honda, Ing and Wu (2019). However, the proof is lengthy and complicated. We put that of Theorem 3 before that of Theorem 2.

Note that we deal with S satisfying $|S|+1 \leq K_n$ in the proofs.

Proof of Theorem 1. The idea of the proof is as follows : The unconditional maximum w.r.t. $\boldsymbol{g}_{S\cup\{j\}}^T(Z)\boldsymbol{X}_{S\cup\{j\}}$ is larger than or equal to the conditional or sequential maximum w.r.t. $h_{jS}(Z)X_j$ in $\boldsymbol{g}_S^{*T}(Z)\boldsymbol{X}_S + h_{jS}(Z)X_j$. Thus our assumptions guarantee a lower bound of the amount of the increase of $\ell_n(\boldsymbol{g}_{S\cup\{j\}}^T(Z)\boldsymbol{X}_{S\cup\{j\}})$ when $\mathcal{M} \not\subset S$.

First notice the following inequalities : Uniformly in $j \notin S$ and S,

$$\ell_n(\boldsymbol{W}_{S\cup\{j\}}^T\widehat{\boldsymbol{\beta}}_{S\cup\{j\}}) \geq \ell_n(\boldsymbol{W}_S^T\widehat{\boldsymbol{\beta}}_S + W_j^T\widehat{\eta}_{jS}) \geq \ell_n(\boldsymbol{W}_S^T\widehat{\boldsymbol{\beta}}_S + W_j^T\eta_{jS}^*). \quad (4.4)$$

We have with probability tending to 1,

$$\begin{aligned} &|\ell_n(\boldsymbol{W}_S^T\widehat{\boldsymbol{\beta}}_S + W_j^T\eta_{jS}^*) - \mathrm{E}\{\ell(Y, \boldsymbol{X}_S^T\boldsymbol{g}_S^*(Z) + X_j h_{jS}^*(Z))\}| \qquad (4.5)\\ &\qquad \leq \textit{the RHS of } (4.3) < \frac{C_{LB}\rho_{LB}^2}{2} \end{aligned}$$

uniformly in $j \notin S$ and S as shown in Lemma 4.

Lemma 1 and Assumption LB imply

$$\mathrm{E}\{\ell(Y, \boldsymbol{X}_S^T\boldsymbol{g}_S(Z) + X_j h_{jS}^*(Z))\} - \mathrm{E}\{\ell(Y, \boldsymbol{X}_S^T\boldsymbol{g}_S^*(Z))\} \geq 2C_{LB}\rho_{LB}^2 \quad (4.6)$$

for some $j \in \mathcal{M} \cap S^c$ if $\mathcal{M} \not\subset S$. Recall that C_{LB} is defined in Lemma 1.

We also have with probability tending to 1,

$$|\ell_n(\boldsymbol{W}_S^T\widehat{\boldsymbol{\beta}}_S) - \mathrm{E}\{\ell(Y, \boldsymbol{X}_S^T\boldsymbol{g}_S^*(Z))\}| \leq \textit{the RHS of } (4.2) < \frac{C_{LB}\rho_{LB}^2}{2} \quad (4.7)$$

uniformly in S as shown in Lemma 3. Thus (4.4)-(4.6) yield the desired result. Hence the proof is complete.

Proof of Theorem 3. Notice that if $(k+1) \leq K_n$, we have by Lemma 3,

$$\begin{aligned} \mathrm{E}\{\ell(Y, \boldsymbol{X}_{S_0}^T \boldsymbol{g}_{S_0}^*(Z))\} - \frac{C_{LB}\rho_{LB}^2}{2} \leq \ell_n(\boldsymbol{W}_{S_0}^T \widehat{\boldsymbol{\beta}}_{S_0}) \leq \ell_n(\boldsymbol{W}_{S_k}^T \widehat{\boldsymbol{\beta}}_{S_k}) \qquad (4.8) \\ \leq \mathrm{E}\{\ell(Y, \boldsymbol{X}_{S_k}^T \boldsymbol{g}_{S_k}^*(Z))\} + \frac{C_{LB}\rho_{LB}^2}{2} \end{aligned}$$

with probability tending to 1.

If $\mathcal{M} \not\subset S_k$ and $(k+1) \leq K_n$,

$$\ell_n(\boldsymbol{W}_{S_k}^T \widehat{\boldsymbol{\beta}}_{S_k}) - \ell_n(\boldsymbol{W}_{S_0}^T \widehat{\boldsymbol{\beta}}_{S_0}) \geq k C_{LB}\rho_{LB}^2 \qquad (4.9)$$

with probability tending to 1 by Theorem 1.

By (4.8) and (4.9), we obtain

$$(k-1)C_{LB}\rho_{LB}^2 \leq \mathrm{E}\{\ell(Y, \boldsymbol{X}_{S_k}^T \boldsymbol{g}_{S_k}^*(Z))\} - \mathrm{E}\{\ell(Y, \boldsymbol{X}_{S_0}^T \boldsymbol{g}_{S_0}^*(Z))\} \qquad (4.10)$$

Besides, we have for any S_k,

$$\mathrm{E}\{\ell(Y, \boldsymbol{X}_{S_k}^T \boldsymbol{g}_{S_k}^*(Z))\} \leq \mathrm{E}\{\ell(Y, \boldsymbol{X}^T \boldsymbol{g}^*(Z))\}. \qquad (4.11)$$

Thus (4.10) and (4.11) imply

$$(k-1)C_{LB}\rho_{LB}^2 \leq \Delta \qquad (4.12)$$

with probability tending to 1. If we have $k+1 = K_n$, (4.12) contradicts the assumption on Δ and K_n. Hence the result of Theorem 3 follows.

Proof of Theorem 2. The former half is trivial from Theorem 1, (2.7), and (2.2).

We exploit the idea of the proof of Theorem 2 of Honda, Ing and Wu (2019) to prove the latter half.

First notice that

$$\boldsymbol{X}_S^T \boldsymbol{g}_S^*(Z) = \boldsymbol{X}^T \boldsymbol{g}^*(Z) \quad \text{and} \quad \boldsymbol{W}_S^T \boldsymbol{\beta}_S^* = \boldsymbol{W}^T \boldsymbol{\beta}^*$$

if $\mathcal{M} \subset S$ and $|S| \leq K_n$. Hence we have for these S,

$$\|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^*\| \leq C_1 \frac{L}{\sigma_{min}\lambda_L} \sqrt{\frac{|S| \log p_n}{n}}$$

uniformly in S with probability tending to 1. The above rate is different from that in Lemma 2 and this is because $p_n^{|S|}$ is not necessary when we employ the arguments based on Bernstein's inequality for these S. See (S1.5) and (S1.6) in the supplementary material.

The proof consists of three steps.

Step 1: Derivation of an expression of $\ell_n(\boldsymbol{W}_S^T \boldsymbol{\beta}_S) - \ell_n(\boldsymbol{W}_S^T \boldsymbol{\beta}_S^*)$.

Step 2: Derivation of an expression of $\ell_n(\boldsymbol{W}_S^T \widehat{\boldsymbol{\beta}}_S) - \ell_n(\boldsymbol{W}_S^T \boldsymbol{\beta}_S^*)$.

Step 3: Evaluation of $\ell_n(\widehat{\boldsymbol{\beta}}_{S_+}) - \ell_n(\widehat{\boldsymbol{\beta}}_S)$, where $S_+ = S \cup \{j\}$ for S and j such that $\mathcal{M} \subset S$, $|S| < K_n$, and $j \notin S$.

Step 1) We derive (4.16), which is similar to (44) of Honda, Ing and Wu (2019).

Let

$$\|\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*\| \leq L\sqrt{\frac{\eta_n |S| \log p_n}{n}}, \tag{4.13}$$

where $\eta_n \to \infty$ at any rate. Then we evaluate

$$\ell_n(\boldsymbol{W}_S^T \boldsymbol{\beta}_S) - \ell_n(\boldsymbol{W}_S^T \boldsymbol{\beta}_S^*) = \ell_n(\boldsymbol{W}_S^T \boldsymbol{\beta}_S) - \ell_n(\boldsymbol{W}^T \boldsymbol{\beta}^*).$$

By Taylor's theorem, this is equal to

$$\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*)^T \boldsymbol{W}_{iS} \ell'(Y_i, \boldsymbol{W}_{iS}^T \boldsymbol{\beta}_S^*)$$
$$+ \frac{1}{2n}\sum_{i=1}^{n}(\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*)^T \boldsymbol{W}_{iS} \boldsymbol{W}_{iS}^T (\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*) \ell''(Y_i, \boldsymbol{W}_{iS}^T \boldsymbol{\beta}_S^* + \theta_i \boldsymbol{W}_{iS}^T (\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*))$$
$$= A_1 + A_2 \qquad \text{(say)},$$

where $|\theta_i| \leq 1$. Recall that $\boldsymbol{W}_S^T \boldsymbol{\beta}_S^* = \boldsymbol{W}^T \boldsymbol{\beta}^*$.

A_1 : We should deal with

$$\frac{1}{n}\sum_{i=1}^{n} X_{ij} B_k(Z_i)[e_i + \{\mu(\boldsymbol{X}_i^T \boldsymbol{g}^*(Z_i)) - \mu(\boldsymbol{W}_i^T \boldsymbol{\beta}^*)\}] = B_1 + B_2 \qquad \text{(say)}.$$

B_1 is the main term and we have

$$\frac{1}{n}\sum_{i=1}^{n} X_{ij} B_k(Z_i) e_i = \frac{1}{n}\sum_{i=1}^{n} X_{ij} B_k(Z_i) \ell'(Y_i, \boldsymbol{X}_i^T \boldsymbol{g}^*(Z_i)) = O_p((\log p_n/(nL))^{1/2}) \tag{4.14}$$

uniformly in j and k from the arguments based on Bernstein's inequality.

As for B_2, there are positive constants C_1 and C_2 such that

$$\Big|\frac{1}{n}\sum_{i=1}^{n} X_{ij}B_k(Z_i)\{\mu(\boldsymbol{X}_i^T\boldsymbol{g}^*(Z_i)) - \mu(\boldsymbol{W}_i^T\boldsymbol{\beta}^*)\}\Big| \tag{4.15}$$
$$\leq \frac{C_1}{n}\sum_{i=1}^{n} B_k(Z_i)\sigma_{max}|X_{ij}||\boldsymbol{X}_i^T\boldsymbol{g}^*(Z_i) - \boldsymbol{W}_i^T\boldsymbol{\beta}^*| \leq \frac{C_2C_XC_{A1}}{L^3}$$

uniformly in j and k with probability tending to 1. Here we used Assumptions X(1) and CF(2) and the fact that $n^{-1}\sum_{i=1}^{n} B_k(Z_i) = O_p(L^{-1})$ uniformly in k.

Hence we have

$$A_1 = \Big\{\frac{1}{n}\sum_{i=1}^{n} \ell'(Y_i, \boldsymbol{X}_i^T\boldsymbol{g}^*(Z_i))\boldsymbol{W}_{iS}^T\Big\}(\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*) + O_p(\|\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*\|\sqrt{L|S|}L^{-3}).$$

Note that by (4.13),

$$\|\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*\|\sqrt{L|S|}L^{-3} = O(n^{-1}L|S|\sqrt{\eta_n \log p_n}).$$

A_2 : Write $\widehat{\Sigma}_S = n^{-1}\sum_{i=1}^{n} \sigma(\boldsymbol{W}_i^T\boldsymbol{\beta}^*)\boldsymbol{W}_{iS}\boldsymbol{W}_{iS}^T$.

Then

$$A_2 = -\frac{1}{2}(\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*)^T\widehat{\Sigma}_S(\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*) + \frac{\lambda_U}{L}\|\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*\|^2 O_p(\delta_n),$$

where $\delta_n = \|\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*\||S|\sigma'_{max}$. The second term is negligible compared to $n^{-1}L|S|\sqrt{\eta_n \log p_n}$.

Hence we have

$$\begin{aligned}
&\ell_n(\boldsymbol{W}_S^T\boldsymbol{\beta}_S) - \ell_n(\boldsymbol{W}_S^T\boldsymbol{\beta}_S^*) \qquad (4.16)\\
&= \Big\{\frac{1}{n}\sum_{i=1}^n \ell'(Y_i, \boldsymbol{X}_i^T\boldsymbol{g}^*(Z_i))\boldsymbol{W}_{iS}^T\Big\}(\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*)\\
&\qquad -\frac{1}{2}(\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*)^T\widehat{\Sigma}_S(\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*) + O_p(n^{-1}L|S|\sqrt{\eta_n \log p_n}).
\end{aligned}$$

Step 2) We will use (4.16) to derive a useful expression of $\ell_n(\boldsymbol{W}_S^T\widehat{\boldsymbol{\beta}}_S)$ in (4.20). Define $\boldsymbol{a}_S$ and $\overline{\boldsymbol{\beta}}_S$ by

$$\boldsymbol{a}_S = \frac{1}{n}\sum_{i=1}^n \boldsymbol{W}_{iS}e_i \quad \text{and} \quad \overline{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^* = \widehat{\Sigma}_S^{-1}\boldsymbol{a}_S. \qquad (4.17)$$

(4.14) and Assumption W imply

$$|\boldsymbol{a}_S|^2 = O_p(n^{-1}|S|\log p_n) \quad \text{and} \quad \overline{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^* = O_p(L(n^{-1}|S|\log p_n)^{1/2}). \qquad (4.18)$$

If for some $\boldsymbol{\delta}_S \in R^{L|S|}$,

$$\|\overline{\boldsymbol{\beta}}_S + \boldsymbol{\delta}_S - \boldsymbol{\beta}_S^*\| \le L\sqrt{\frac{\eta_n|S|\log p_n}{n}},$$

we have from (4.16) that uniformly in $\boldsymbol{\delta}_S$ and S,

$$\begin{aligned}
&\ell_n(\boldsymbol{W}_S^T(\overline{\boldsymbol{\beta}}_S + \boldsymbol{\delta}_S)) - \ell_n(\boldsymbol{W}_S^T\boldsymbol{\beta}_S^*) \qquad (4.19)\\
&= -\frac{1}{2}\boldsymbol{a}_S^T\widehat{\Sigma}_S^{-1}\boldsymbol{a}_S + \frac{1}{2}\boldsymbol{\delta}_S^T\widehat{\Sigma}_S\boldsymbol{\delta}_S + O_p\Big(\frac{|S|L(\eta_n\log p_n)^{1/2}}{n}\Big).
\end{aligned}$$

Because of the optimality of $\ell_n(\widehat{\boldsymbol{\beta}}_S)$ and (4.19), we obtain

$$\ell_n(\boldsymbol{W}_S^T\widehat{\boldsymbol{\beta}}_S) - \ell_n(\boldsymbol{W}_S^T\boldsymbol{\beta}_S^*) = -\frac{1}{2}\boldsymbol{a}_S^T\widehat{\Sigma}_S^{-1}\boldsymbol{a}_S + O_p\Big(\frac{|S|L(\eta_n\log p_n)^{1/2}}{n}\Big) \qquad (4.20)$$

uniformly in S. Recall that $\ell_n(\boldsymbol{W}_S^T\boldsymbol{\beta}_S^*) = \ell_n(\boldsymbol{W}^T\boldsymbol{\beta}^*)$.

Step 3) Hereafter we write $S_+ = S \cup \{j\}$ for S and j such that $\mathcal{M} \subset S$, $|S| < K_n$, and $j \notin S$. Then we evaluate $\ell_n(\widehat{\boldsymbol{\beta}}_{S_+}) - \ell_n(\widehat{\boldsymbol{\beta}}_S)$ by using (4.20).

We write

$$\widehat{\Sigma}_{S_+} = \begin{pmatrix} \widehat{\Sigma}_S & \widehat{\Sigma}_{Sj} \\ \widehat{\Sigma}_{jS} & \widehat{\Sigma}_{jj} \end{pmatrix} \quad \text{and} \quad \boldsymbol{a}_{S_+} = \begin{pmatrix} \boldsymbol{a}_S \\ a_j \end{pmatrix}. \tag{4.21}$$

Thus due to (4.20), we have only to consider the difference

$$\begin{aligned} \boldsymbol{a}_{S_+}^T \widehat{\Sigma}_{S_+}^{-1} \boldsymbol{a}_{S_+} - \boldsymbol{a}_S^T \widehat{\Sigma}_S^{-1} \boldsymbol{a}_S = \boldsymbol{a}_S^T \widehat{\Sigma}_S^{-1} \widehat{\Sigma}_{Sj} \widehat{F}_{Sj} \widehat{\Sigma}_{jS} \widehat{\Sigma}_S^{-1} \boldsymbol{a}_S \\ - 2\boldsymbol{a}_S^T \widehat{\Sigma}_S^{-1} \widehat{\Sigma}_{jS} \widehat{F}_{Sj} a_j + a_j^T \widehat{F}_{Sj} a_j, \end{aligned} \tag{4.22}$$

where $\widehat{F}_{Sj} = (\widehat{\Sigma}_S - \widehat{\Sigma}_{jS}\widehat{\Sigma}_S^{-1}\widehat{\Sigma}_{Sj})^{-1}$, when we evaluate $\ell_n(\widehat{\boldsymbol{\beta}}_{S_+}) - \ell_n(\widehat{\boldsymbol{\beta}}_S)$.

We will demonstrate that the RHS of (4.22) has the stochastic order of $L|S|O_p(n^{-1}\log p_n)$ uniformly in S and j. Then the latter half of Theorem 2 is established with (4.20).

By Assumptions B(1) and W, we have for some positive C_1, C_2, and C_3,

$$C_1 L \le \lambda_{\min}(\widehat{F}_{Sj}) \le \lambda_{\max}(\widehat{F}_{Sj}) \le C_2 L \quad \text{and} \quad \lambda_{\max}(\widehat{\Sigma}_{jS}\widehat{\Sigma}_{Sj}) \le C_3 L^{-2} \tag{4.23}$$

uniformly in S with probability tending to 1. (4.14) implies that uniformly

in j,

$$|a_j|^2 = LO_p\Big(\frac{\log p_n}{nL}\Big). \tag{4.24}$$

Hence (4.23) and (4.24) yield that the third term on the RHS of (4.22) satisfies

$$a_j^T \widehat{F}_{Sj} a_j = LO_p(n^{-1}\log p_n) \quad \text{uniformly in } j \text{ and } S. \tag{4.25}$$

Next we evaluate the first and second terms on the RHS of (4.22) :

$$(\boldsymbol{a}_S^T \widehat{\Sigma}_S^{-1}\widehat{\Sigma}_{Sj})\widehat{F}_{Sj}(\widehat{\Sigma}_{jS}\widehat{\Sigma}_S^{-1}\boldsymbol{a}_S) \quad \text{and} \quad (\boldsymbol{a}_S^T\widehat{\Sigma}_S^{-1}\widehat{\Sigma}_{Sj})\widehat{F}_{Sj}a_j. \tag{4.26}$$

To establish (4.31) below, we carefully evaluate

$$\widehat{\Sigma}_{jS}\widehat{\Sigma}_S^{-1}\boldsymbol{a}_S = \widehat{\Sigma}_{jS}\widehat{\Sigma}_S^{-1}\frac{1}{n}\sum_{i=1}^n \boldsymbol{W}_{iS}e_i. \tag{4.27}$$

Then we need to examine $\widehat{\Sigma}_{Sj} = \widehat{\Sigma}_{jS}^T$ closely. We write

$$\widehat{\Sigma}_{Sj} = (\boldsymbol{s}_1, \ldots, \boldsymbol{s}_L)$$

and notice from (4.23) that

$$\boldsymbol{s}_m^T\boldsymbol{s}_m = O_p(L^{-2}) \text{ and } \lambda_{\max}(\widehat{\Sigma}_{jS}\widehat{\Sigma}_S^{-1}\widehat{\Sigma}_S\widehat{\Sigma}_S^{-1}\widehat{\Sigma}_{Sj}) = O_p(L^{-1}) \tag{4.28}$$

uniformly in $m \in \{1, \ldots, L\}$, j, and S with probability tending to 1. Besides, we have for some positive C_4 and C_5,

$$\max_m |\boldsymbol{s}_m^T\widehat{\Sigma}_S^{-1}\boldsymbol{W}_{iS}| \le C_4 L\|\boldsymbol{s}_m\|\|\boldsymbol{W}_{iS}\| \le C_5 L|S|\|\boldsymbol{s}_m\|C_{X1} = O_p(|S|C_{X1}) \tag{4.29}$$

uniformly in i and S with probability tending to 1.

Here we employ the standard arguments based on Bernstein's inequality conditionally on $\boldsymbol{s}_m^T \widehat{\Sigma}_S^{-1} \boldsymbol{W}_{iS}$ by using Assumption E and the above properties of $\boldsymbol{s}_m^T \widehat{\Sigma}_S^{-1} \boldsymbol{W}_{iS}$ and obtain

$$\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{s}_m^T \widehat{\Sigma}_S^{-1} \boldsymbol{W}_{iS} e_i = O_p(\{(nL)^{-1} |S| \log p_n\}^{1/2}) \tag{4.30}$$

uniformly in m, S, and j. Note that $|S|$ in $O_p(\{(nL)^{-1}|S| \log p_n\}^{1/2})$ is necessary since $\{\boldsymbol{s}_m^T \widehat{\Sigma}_S^{-1} \boldsymbol{W}_{iS}\}_{i=1}^n$ depends on m, j, and S and we have to take into account all S and j satisfying $\mathcal{M} \subset S$, $|S| < K_n$, and $j \notin S$. See also the arguments around (S1.6) in the proof of Lemma 2.

Therefore (4.30) yields that uniformly in S and j,

$$|\widehat{\Sigma}_{jS} \widehat{\Sigma}_S^{-1} \boldsymbol{a}_S|^2 = L O_p((nL)^{-1} |S| \log p_n). \tag{4.31}$$

Thus (4.23), (4.24), (4.26), and (4.31) imply that the first and second terms on the RHS of (4.22) have the stochastic order of $O_p(n^{-1}|S|L \log p_n)$ uniformly in S and j. We have demonstrated that the RHS of (4.22) has the stochastic order of $O_p(n^{-1}|S|L \log p_n)$ uniformly in S and j.

Hence the proof of Theorem 2 is complete.

5. Conclusions

We proposed two forward screening procedures with a stopping rule and established their desirable properties such as screening consistency in Section 2. When we constructed our stopping rule, we took uniformity in the convergence rate of estimators into consideration. Such important uniformity has been overlooked in the literature. Thus the stopping rule and the related information criterion have their own significance.

Our simulation studies showed good finite sample performances of the proposed procedures as screening tools. We also applied one of our procedures to a real data set with $p = 44760$.

Our ML-type forward regression procedure, denoted by FR in Section 3, is based on maximizing the log-likelihood function without any approximation except for spline function approximation to coefficient functions. Our simulation studies in Subsection 3.1 showed a simpler procedure called SC there performed closely to our FR procedure. Likelihood maximization as in the FR procedure may be desirable. However, the SC screening procedure can be a useful and faster alternative for extremely large p.

Supplementary Materials

The supplementary material includes the proofs of all the lemmas in

the paper.

Acknowledgements

We appreciate comments and help from Prof. Ching-Kang Ing very much. Lin's research was supported in part by the Science Vanguard Research Program of the Ministry of Science and Technology, Taiwan.

References

Breheny, P. and Huang, J. (2015). Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Statistics and Computing 25*, pp. 173-187.

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

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

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

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

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

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 95*, pp. 1348–1360.

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

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

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

Fan, J. and Zhang, W. (2008). Statistical methods with varying coefficient models. *Statistics and its Interface 1*, pp. 179–195.

Hastie, T., Tibshirani, R. and Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Boca Raton: Chapman & Hall/CRC.

Heinze G. and Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in medicine 21*, pp. 2409-2419.

Honda, T., Ing, C. K. and Wu, W. Y. (2019). Adaptively weighted group Lasso for semipara-

metric quantile regression models. *Bernoulli 25*, pp. 3311–3338.

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

Kim, Y. and Jeon, J. J. (2016). Consistent model selection criteria for quadratically supported risks.. *The Annals of Statistics 44*, pp. 2467–2496.

Lee, E. R., Noh, H. and Park, B. U. (2014). Model selection via Bayesian information criterion for quantile regression models. *Journal of the American Statistical Association 109*, pp. 216–229.

Liu, J., Zhong, W. and Li, R. (2015). A selective overview of feature screening for ultrahigh-dimensional data. *Science China Mathematics 58*, pp. 1–22.

Luo, S. and Chen, Z. (2014). Sequential Lasso cum EBIC for feature selection with ultra-high dimensional feature space. *Journal of the American Statistical Association 109*, pp. 1229–1240.

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

Schumaker, L. (2007). *Spline Functions: Basic Theory.* 3rd Edition. Cambridge: Cambridge University Press.

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

Xia, X., Yang, H., and Li, J. (2016). Feature screening for generalized varying coefficient models with application to dichotomous responses. *Computational Statistics & Data Analysis 102*, pp. 85–97.

Yang, G., Yang, S. and Li, R. (2020). Feature screening in ultrahigh dimensional generalized varying-coefficient models. Forthcoming in *Statistica Sinica*.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. New York: Springer.

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

Zheng, Q., Hong, H. G. and Li, Y. (2020). Building Generalized Linear Models with Ultrahigh Dimensional Features: A Sequentially Conditional Approach. Forthcoming in *Biometrics*.

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

Graduate School of Economics, Hitotsubashi University, Japan

E-mail: t.honda@r.hit-u.ac.jp

Institute of Statistics, National Tsing Hua University, Taiwan

E-mail: ctlin@mx.nthu.edu.tw

FORWARD VARIABLE SELECTION FOR SPARSE ULTRA-HIGH DIMENSIONAL GENERALIZED VARYING COEFFICIENT MODELS

Toshio Honda[1] and Chien-Tong Lin[2]

Hitotsubashi University[1] *and National Tsing Hua University*[2]

Supplementary Material

This supplementary material contains the proofs of Lemmas 1-4 and additional simulation results.

S1 Proofs of Lemmas 1-4

Recall that $|S| \leq K_n$ in this section.

Proof of Lemma 1.

(i) Take $\overline{h}_{jS}(z)$ in Assumption LB. Then we have by the definition (i.e.

optimality) of $h^*_{jS}(z)$ and Taylor's theorem,

$$
\begin{aligned}
&|\mathrm{E}[\{Y-\mu(\boldsymbol{X}_S^T\boldsymbol{g}_S^*(Z))\}X_j\overline{h}_{jS}(Z)]| \\
&= |\mathrm{E}[\{\mu(\boldsymbol{X}^T\boldsymbol{g}^*(Z))-\mu(\boldsymbol{X}_S^T\boldsymbol{g}_S^*(Z))\}X_j\overline{h}_{jS}(Z)] \\
&\qquad -\mathrm{E}[\{\mu(\boldsymbol{X}^T\boldsymbol{g}^*(Z))-\mu(\boldsymbol{X}_S^T\boldsymbol{g}_S^*(Z)+X_jh^*_{jS}(Z))\}X_j\overline{h}_{jS}(Z)]| \\
&= |\mathrm{E}[\{\mu(\boldsymbol{X}_S^T\boldsymbol{g}_S^*(Z)+X_jh^*_{jS}(Z))-\mu(\boldsymbol{X}_S^T\boldsymbol{g}_S^*(Z))\}X_j\overline{h}_{jS}(Z)]| \\
&= |\mathrm{E}\{X_j^2h^*_{jS}(Z)\overline{h}_{jS}(Z)\sigma(\boldsymbol{X}_S^T\boldsymbol{g}_S^*(Z)+\delta_1X_jh^*_{jS}(Z))\}| \\
&\le \sigma_{max}\mathrm{E}[\mathrm{E}\{X_j^2\,|\,Z\}|h^*_{jS}(Z)\overline{h}_{jS}(Z)|] \\
&\le C_1\sigma_{max}\mathrm{E}\{h^{*2}_{jS}(Z)\}^{1/2},
\end{aligned}
$$

where $|\delta_1|\le 1$, for some positive constant C_1. Here we used the following facts :

$$\mathrm{E}\{Y|\boldsymbol{X},Z\}=\mu(\boldsymbol{X}^T\boldsymbol{g}^*(Z)) \text{ and } \mathrm{E}\{X_j\ell'(\boldsymbol{X}_S^T\boldsymbol{g}_S^*(Z)+X_jh^*_{jS}(Z))|Z\}=0.$$

We also used Assumptions B(1), CF(1), X(1), and Z. Hence (i) is established.

(ii) We also have by the definition of $h^*_{jS}(z)$ and Taylor's theorem,

$$
\begin{aligned}
&\mathrm{E}\{\ell(\boldsymbol{X}_S^T\boldsymbol{g}_S^*(Z)+X_jh^*_{jS}(Z))-\ell(\boldsymbol{X}_S^T\boldsymbol{g}_S^*(Z))\} \\
&= \mathrm{E}[\mathrm{E}\{X_j\ell'(\boldsymbol{X}_S^T\boldsymbol{g}_S^*(Z)+X_jh^*_{jS}(Z))|Z\}h^*_{jS}(Z)] \\
&\qquad +\frac{1}{2}\mathrm{E}\{X_j^2h^{*2}_{jS}(Z)\sigma(\boldsymbol{X}_S^T\boldsymbol{g}_S^*(Z)+\delta_2X_jh^*_{jS}(Z))\} \\
&\ge C_2\sigma_{min}\mathrm{E}\{h^{*2}_{jS}(Z)\},
\end{aligned}
$$

where $|\delta_2| \leq 1$, for some positive constant C_2. Note that we used Assumptions B(1), CF(1), X(1)(2), and Z here. Hence (ii) is established.

Hence the proof of Lemma 1 is complete.

In the following proof, we keep their values when we use M_j repeatedly.

Proof of Lemma 2. We consider $\boldsymbol{\beta}_S$ satisfying

$$\|\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*\| = 2C_{tmp}\frac{L|S|}{\sigma_{min}\lambda_L}\sqrt{\frac{\log p_n}{n}}, \tag{S1.1}$$

where $|S| \leq K_n$ and C_{tmp} is to be specified later in this proof.

Then we have by Taylor's theorem,

$$\begin{aligned}
&\ell_n(\boldsymbol{W}_S^T\boldsymbol{\beta}_S) - \ell_n(\boldsymbol{W}_S^T\boldsymbol{\beta}_S^*) \\
&= \Big\{\frac{1}{n}\sum_{i=1}^n \ell'(Y_i, \boldsymbol{W}_{iS}^T\boldsymbol{\beta}_S^*)\boldsymbol{W}_{iS}^T\Big\}(\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*) \\
&\quad + \frac{1}{2n}\sum_{i=1}^n \ell''(Y_i, \boldsymbol{W}_{iS}^T\boldsymbol{\beta}_S^* + \delta_{3i}\boldsymbol{W}_{iS}^T(\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*))(\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*)^T\boldsymbol{W}_{iS}\boldsymbol{W}_{iS}^T(\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*) \\
&= A_1 + A_2 \quad \text{(say)},
\end{aligned} \tag{S1.2}$$

where $|\delta_{3i}| \leq 1$.

Recall that $\ell'(y,\theta) = y - \mu(\theta)$ and $\ell''(y,\theta) = -\sigma(\theta)$. Now we evaluate A_1 and A_2.

A_1 : An element of the left side of A_1 is divided into

$$B_{1jk} = \frac{1}{n}\sum_{i=1}^n \{\ell'(Y_i, \boldsymbol{W}_{iS}^T\boldsymbol{\beta}_S^*) - \ell'(Y_i, \boldsymbol{X}_{iS}^T\boldsymbol{g}_{iS}^*(Z_i))\}X_{ij}B_k(Z_i)$$

and

$$B_{2jk} = \frac{1}{n}\sum_{i=1}^{n} \ell'(Y_i, \boldsymbol{X}_{iS}^T \boldsymbol{g}_{iS}^*(Z_i)) X_{ij} B_k(Z_i)$$

for $1 \le j \le p$ and $1 \le k \le L$.

Since

$$|\boldsymbol{W}_{iS}^T \boldsymbol{\beta}_S^* - \boldsymbol{X}_{iS}^T \boldsymbol{g}_{iS}^*(Z_i)| \le \frac{C_X C_{A1}}{L^2}$$

and

$$\frac{1}{n}\sum_{i=1}^{n} B_k(Z_i) = O_p(L^{-1}) \tag{S1.3}$$

uniformly in k, we have with probability tending to 1,

$$|B_{1jk}| \le C_1 \frac{C_X^2 C_{A1}}{L^3} \tag{S1.4}$$

uniformly in j and k for some positive constant C_1. Note that (S1.3) follows from the properties of the B-spline basis and Bernstein's inequality. We also used Assumptions X(1), CF(1)(2), and B(1) here.

Next we evaluate B_{2jk} by using Assumption E and Bernstein's inequality. First recall that $\mathrm{E}\{\ell'(Y, \boldsymbol{X}_S^T \boldsymbol{g}_S^*(Z)) X_j B_k(Z)\} = 0$ for $j \in S$ and notice that

$$\{Y - \mu(\boldsymbol{X}_S^T \boldsymbol{g}_S(Z))\} X_j B_k(Z) = \{e + \mu(\boldsymbol{X}^T \boldsymbol{g}^*(Z)) - \mu(\boldsymbol{X}_S^T \boldsymbol{g}_S^*(Z))\} X_j B_k(Z).$$

Thus Assumptions X(1), CF(1)(2), B, and E imply that there are some

positive constants M_1 and M_2 such that

$$|\mathrm{E}[|\{Y-\mu(\boldsymbol{X}_S^T\boldsymbol{g}_S(Z))\}X_jB_k(Z)|^m] \le \frac{M_1}{L}m!M_2^{m-2} \tag{S1.5}$$

for $m \ge 2$.

Applying Bernstein's inequality with (S1.5), we obtain

$$\begin{aligned}
&\mathrm{P}\Big\{\Big|n^{-1}\sum_{i=1}^n \ell'(\boldsymbol{X}_{iS}^T\boldsymbol{g}_S^*(Z_i))X_{ij}B_k(Z_i)\Big| \ge \frac{\kappa_n}{\sqrt{nL}}\Big\} \\
&\le \ 2\exp\Big\{-\frac{n^2\kappa_n^2/(2nL)}{M_1n/L+M_2n\kappa_n/\sqrt{nL}}\Big\} \\
&\le \ 2\exp(-M_3\kappa_n^2)
\end{aligned}$$

for some positive constant M_3. We specify κ_n later.

To guarantee the uniformity of

$$\Big|n^{-1}\sum_{i=1}^n \ell'(\boldsymbol{X}_{iS}^T\boldsymbol{g}_S^*(Z_i))X_{ij}B_k(Z_i)\Big| \ge \frac{\kappa_n}{\sqrt{nL}}$$

in $|S|$ satisfying $|S|=q$ and in $q \le K_n$, we have to deal with

$$2\exp(-M_3\kappa_n^2)\exp(q\log p_n) \tag{S1.6}$$

since $\boldsymbol{g}_S^*(z)$ depends on S. Hence we take $\kappa_n^2 = 4q\log p_n/M_3$ and have

$$2\exp(-M_3\kappa_n^2)\exp(q\log p_n) \le 2\exp(-3q\log p_n).$$

Since we have

$$\sum_{q=1}^\infty 2\exp(-3q\log p_n) \le 2\sum_{q=1}^\infty n^{-3q} \le \frac{16}{7n^3}$$

for $n \geq 2$, we obtain

$$\begin{aligned} |A_1| &\leq C_3\sqrt{|S|L}\|\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*\|\Big(\sqrt{\frac{|S|\log p_n}{nL}} + \frac{1}{L^3}\Big) \\ &\leq M_4\|\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*\|\sqrt{\frac{|S|^2\log p_n}{n}} \end{aligned} \tag{S1.7}$$

uniformly in S with probability tending to 1 for some positive constants C_3 and M_4.

A_2 : Assumptions B and W imply that we have

$$\begin{aligned} A_2 &\leq -\frac{1}{2}\sigma_{min}(\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*)^T\Big(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{W}_{iS}\boldsymbol{W}_{iS}^T\Big)(\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*) \\ &\leq -\frac{\sigma_{min}\lambda_L}{2L}\|\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*\|^2 \end{aligned} \tag{S1.8}$$

uniformly in S with probability tending to 1.

Combining (S1.7) and (S1.8), we have

$$A_1 + A_2 \leq \|\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*\|(M_4 - C_{tmp})\sqrt{\frac{|S|^2\log p_n}{n}} < 0 \tag{S1.9}$$

uniformly in S with probability tending to 1 if we take $C_{tmp} = 2M_4$. Note that (S1.9) holds uniformly in $\boldsymbol{\beta}_S$ in (S1.1). Thus the concavity of $\ell_n(\boldsymbol{W}_S^T\boldsymbol{\beta}_S)$, (S1.2), and (S1.9) yield the desired result. Hence the proof of Lemma 2 is complete.

We can prove Lemmas 3 and 4 in the same way and the proof of Lemma 4 is omitted.

Proof of Lemma 3. We have

$$\begin{aligned}
&\ell_n(\boldsymbol{W}_S^T\widehat{\boldsymbol{\beta}}_S) - \mathrm{E}\{\ell(Y, \boldsymbol{X}_S^T\boldsymbol{g}_S^*(Z))\} \\
&= \ell_n(\boldsymbol{W}_S^T\widehat{\boldsymbol{\beta}}_S) - \ell_n(\boldsymbol{W}_S^T\boldsymbol{\beta}_S^*) + \ell_n(\boldsymbol{W}_S^T\boldsymbol{\beta}_S^*) - \ell_n(\boldsymbol{X}_S^T\boldsymbol{g}_S^*(Z)) \\
&\qquad +\ell_n(\boldsymbol{X}_S^T\boldsymbol{g}_S^*(Z)) - \mathrm{E}\{\ell(Y, \boldsymbol{X}_S^T\boldsymbol{g}_S^*(Z))\} \\
&= A_1 + A_2 + A_3 \quad \text{(say)}. \qquad \text{(S1.10)}
\end{aligned}$$

We evaluate A_1, A_2, and A_3.

A_1 : By the LLN for Y_i^2, Assumption B, and Lemma 2, we have for some positive constants C_1, C_2, and C_3,

$$\begin{aligned}
|A_1| &\le \frac{1}{n}\sum_{i=1}^{n}(|Y_i| + \mu_{max})|\boldsymbol{W}_{iS}^T(\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^*)| \\
&\le \Big\{n^{-1}\sum_{i=1}^{n}(|Y_i| + \mu_{max})^2\Big\}^{1/2}\Big\{n^{-1}\sum_{i=1}^{n}|\boldsymbol{W}_{iS}^T(\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^*)|^2\Big\}^{1/2} \\
&\le C_1\Big(\frac{\lambda_U}{L}\Big)^{1/2}\|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^*\| \le C_2\Big(\frac{\lambda_U}{L}\Big)^{1/2}\frac{L|S|}{\sigma_{min}\lambda_L}\sqrt{\frac{\log p_n}{n}} \\
&\le C_3\frac{\lambda_U^{1/2}|S|}{\sigma_{min}\lambda_L}\sqrt{\frac{\log p_n}{n^{4/5}}} \qquad \text{(S1.11)}
\end{aligned}$$

uniformly in S with probability tending to 1.

A_2 : By the LLN for $|Y_i|$ and Assumptions X(1), CF(2) and B, we have for some positive constant C_4,

$$\begin{aligned}
|A_2| &\le \frac{1}{n}\sum_{i=1}^{n}(|Y_i| + \mu_{max})|\boldsymbol{W}_{iS}^T\boldsymbol{\beta}_S^* - \boldsymbol{X}_{iS}^T\boldsymbol{g}_S^*(Z_i)| \\
&\le C_4(\mu_{max} + \mathrm{E}\{|Y|\})\frac{C_X C_{A1}}{L^2} \qquad \text{(S1.12)}
\end{aligned}$$

uniformly in S with probability tending to 1.

A_3 : We should apply the argument based on Bernstein's inequality to

$$\frac{1}{n}\sum_{i=1}^{n}[\ell(Y_i, \boldsymbol{X}_{iS}\boldsymbol{g}_S^*(Z_i)) - \mathrm{E}\{\ell(Y_i, \boldsymbol{X}_{iS}\boldsymbol{g}_S^*(Z_i))\}].$$

As in the proof of Lemma 2, we can prove that for some positive constant C_5,

$$|A_3| \leq C_5\sqrt{\frac{|S|\log p_n}{n}} \tag{S1.13}$$

uniformly in S with probability tending to 1.

(S1.10)-(S1.13) yield the desired inequality. Hence the proof of Lemma 3 is complete.

S2 Additional numerical studies

Here we give the simulation results for design (Σ_1, G_2) and (Σ_2, G_1). We made almost the same conclusions as in Subsection 3.1.

Table 1: Simulation results for the design (Σ_1, G_2) with $\rho = 0.25$, $(n, p) = (200, 1000)$ for the normal model, and $(n, p) = (300, 1000)$ for logistic and Poisson models.

Screen	Stop	m	Normal			Logistic			Poisson		
			Sure	TP	FP	Sure	TP	FP	Sure	TP	FP
NIS	-	-	0	2.99	7.01	0.00	2.99	11.01	0.00	2.75	11.26
gLASSO	-	-	1	4.00	26.02	1.00	4.00	73.67	1.00	4.00	41.20
gSCAD	-	-	1	4.00	3.37	1.00	4.00	0.08	1.00	4.00	0.24
FR.full	-	-	1	4.00	6.00	1.00	4.00	6.00	1.00	4.00	6.00
SC.full	-	-	1	4.00	6.00	1.00	4.00	5.96	1.00	4.00	6.00
FR	LIC	1	1	4.00	0.00	0.00	1.90	0.00	0.36	3.29	0.00
		2	1	4.00	1.00	0.00	2.90	0.00	0.93	3.93	0.37
		3	1	4.00	2.00	0.90	3.90	0.00	1.00	4.00	1.30
		4	1	4.00	3.00	1.00	4.00	0.90	1.00	4.00	2.30
		5	1	4.00	4.00	1.00	4.00	1.90	1.00	4.00	3.30
	HBIC	1	1	4.00	0.00	0.00	1.00	0.00	0.78	3.64	0.00
		2	1	4.00	1.00	0.00	2.00	0.00	0.94	3.94	0.79
		3	1	4.00	2.00	0.10	3.10	0.00	1.00	4.00	1.73
		4	1	4.00	3.00	1.00	4.00	0.10	1.00	4.00	2.73
		5	1	4.00	4.00	1.00	4.00	1.09	1.00	4.00	3.73
SC	LIC	1	1	4.00	0.00	0.00	1.88	0.00	0.38	3.30	0.00
		2	1	4.00	1.00	0.00	2.88	0.00	0.92	3.92	0.38
		3	1	4.00	2.00	0.88	3.88	0.00	1.00	4.00	1.31
		4	1	4.00	3.00	1.00	4.00	0.88	1.00	4.00	2.31
		5	1	4.00	4.00	1.00	4.00	1.87	1.00	4.00	3.31
	HBIC	1	1	4.00	0.00	0.00	1.00	0.00	0.78	3.63	0.00
		2	1	4.00	1.00	0.00	2.00	0.00	0.94	3.94	0.78
		3	1	4.00	2.00	0.10	3.10	0.00	1.00	4.00	1.72
		4	1	4.00	3.00	1.00	4.00	0.10	1.00	4.00	2.72
		5	1	4.00	4.00	1.00	4.00	1.10	1.00	4.00	3.72

Table 2: Simulation results for the design (Σ_1, G_2) with $\rho = 0.25$, $(n, p) = (400, 1000)$ for the normal model, and $(n, p) = (500, 1000)$ for logistic and Poisson models.

Screen	Stop	m	Normal			Logistic			Poisson		
			Sure	TP	FP	Sure	TP	FP	Sure	TP	FP
NIS	-	-	0	3	15.00	0.00	3.00	18.00	0	2.98	18.02
gLASSO	-	-	1	4	9.30	1.00	4.00	104.14	1	4.00	41.59
gSCAD	-	-	1	4	1.21	1.00	4.00	0.32	1	4.00	0.11
FR.full	-	-	1	4	6.00	1.00	4.00	6.00	1	4.00	6.00
SC.full	-	-	1	4	6.00	1.00	4.00	5.98	1	4.00	6.00
FR	LIC	1	1	4	0.00	0.01	2.06	0.00	1	4.00	0.00
		2	1	4	1.00	0.57	3.57	0.01	1	4.00	1.00
		3	1	4	2.00	1.00	4.00	0.58	1	4.00	2.00
		4	1	4	3.00	1.00	4.00	1.58	1	4.00	3.00
		5	1	4	4.00	1.00	4.00	2.58	1	4.00	4.00
	HBIC	1	1	4	0.00	0.00	1.05	0.00	1	4.00	0.00
		2	1	4	1.00	0.06	2.12	0.00	1	4.00	1.00
		3	1	4	2.00	1.00	4.00	0.06	1	4.00	2.00
		4	1	4	3.00	1.00	4.00	1.05	1	4.00	3.00
		5	1	4	4.00	1.00	4.00	2.06	1	4.00	4.00
SC	LIC	1	1	4	0.00	0.01	2.06	0.00	1	4.00	0.00
		2	1	4	1.00	0.57	3.57	0.01	1	4.00	1.00
		3	1	4	2.00	1.00	4.00	0.58	1	4.00	2.00
		4	1	4	3.00	1.00	4.00	1.58	1	4.00	3.00
		5	1	4	4.00	1.00	4.00	2.58	1	4.00	4.00
	HBIC	1	1	4	0.00	0.00	1.06	0.00	1	4.00	0.00
		2	1	4	1.00	0.07	2.15	0.00	1	4.00	1.00
		3	1	4	2.00	1.00	4.00	0.07	1	4.00	2.00
		4	1	4	3.00	1.00	4.00	1.07	1	4.00	3.00
		5	1	4	4.00	1.00	4.00	2.07	1	4.00	4.00

Table 3: Simulation results for the design (Σ_1, G_2) with $\rho = 0.5$, $(n, p) = (200, 1000)$ for the normal model, and $(n, p) = (300, 1000)$ for logistic and Poisson models.

Screen	Stop	m	Normal			Logistic			Poisson		
			Sure	TP	FP	Sure	TP	FP	Sure	TP	FP
NIS	-	-	0	2.96	7.04	0.00	2.95	11.05	0.00	2.77	11.23
gLASSO	-	-	1	4.00	36.51	1.00	4.00	72.28	1.00	4.00	40.38
gSCAD	-	-	1	4.00	4.29	0.99	3.99	0.08	1.00	4.00	0.16
FR.full	-	-	1	4.00	6.00	1.00	4.00	6.00	1.00	4.00	6.00
SC.full	-	-	1	4.00	6.00	1.00	4.00	5.81	1.00	4.00	6.00
FR	LIC	1	1	4.00	0.00	0.00	1.97	0.00	0.00	2.08	0.03
		2	1	4.00	1.00	0.00	2.97	0.00	0.08	3.08	0.03
		3	1	4.00	2.00	0.97	3.97	0.00	0.99	3.99	0.12
		4	1	4.00	3.00	1.00	4.00	0.97	1.00	4.00	1.10
		5	1	4.00	4.00	1.00	4.00	1.97	1.00	4.00	2.10
	HBIC	1	1	4.00	0.00	0.00	1.00	0.00	0.02	1.98	0.03
		2	1	4.00	1.00	0.00	2.00	0.00	0.12	3.01	0.05
		3	1	4.00	2.00	0.05	3.05	0.00	0.90	3.89	0.17
		4	1	4.00	3.00	1.00	4.00	0.05	0.99	3.99	1.07
		5	1	4.00	4.00	1.00	4.00	1.05	1.00	4.00	2.06
SC	LIC	1	1	4.00	0.00	0.00	1.97	0.00	0.00	2.08	0.03
		2	1	4.00	1.00	0.00	2.97	0.00	0.09	3.08	0.04
		3	1	4.00	2.00	0.97	3.97	0.00	0.99	3.99	0.12
		4	1	4.00	3.00	1.00	4.00	0.97	1.00	4.00	1.12
		5	1	4.00	4.00	1.00	4.00	1.97	1.00	4.00	2.12
	HBIC	1	1	4.00	0.00	0.00	1.00	0.00	0.02	1.98	0.03
		2	1	4.00	1.00	0.00	2.00	0.00	0.13	3.02	0.06
		3	1	4.00	2.00	0.04	3.04	0.00	0.90	3.89	0.18
		4	1	4.00	3.00	1.00	4.00	0.04	0.99	3.99	1.08
		5	1	4.00	4.00	1.00	4.00	1.04	1.00	4.00	2.08

Table 4: Simulation results for the design (Σ_1, G_2) with $\rho = 0.5$, $(n, p) = (400, 1000)$ for the normal model, and $(n, p) = (500, 1000)$ for logistic and Poisson models.

Screen	Stop	m	Normal			Logistic			Poisson		
			Sure	TP	FP	Sure	TP	FP	Sure	TP	FP
NIS	-	-	0	3	15.00	0.00	3.00	18.00	0.00	2.98	18.02
gLASSO	-	-	1	4	21.16	1.00	4.00	99.98	1.00	4.00	39.21
gSCAD	-	-	1	4	2.41	1.00	4.00	0.27	1.00	4.00	0.08
FR.full	-	-	1	4	6.00	1.00	4.00	6.00	1.00	4.00	6.00
SC.full	-	-	1	4	6.00	1.00	4.00	6.00	1.00	4.00	6.00
FR	LIC	1	1	4	0.00	0.00	2.00	0.00	0.22	2.87	0.00
		2	1	4	1.00	0.28	3.28	0.00	0.68	3.67	0.22
		3	1	4	2.00	1.00	4.00	0.28	1.00	4.00	0.89
		4	1	4	3.00	1.00	4.00	1.28	1.00	4.00	1.89
		5	1	4	4.00	1.00	4.00	2.28	1.00	4.00	2.89
	HBIC	1	1	4	0.00	0.00	1.12	0.00	0.39	2.78	0.00
		2	1	4	1.00	0.12	2.25	0.00	0.84	3.82	0.39
		3	1	4	2.00	0.98	3.98	0.12	0.99	3.99	1.23
		4	1	4	3.00	1.00	4.00	1.10	1.00	4.00	2.21
		5	1	4	4.00	1.00	4.00	2.10	1.00	4.00	3.21
SC	LIC	1	1	4	0.00	0.00	2.00	0.00	0.24	2.90	0.00
		2	1	4	1.00	0.28	3.27	0.00	0.70	3.70	0.24
		3	1	4	2.00	1.00	4.00	0.28	1.00	4.00	0.94
		4	1	4	3.00	1.00	4.00	1.27	1.00	4.00	1.94
		5	1	4	4.00	1.00	4.00	2.27	1.00	4.00	2.94
	HBIC	1	1	4	0.00	0.00	1.16	0.00	0.42	2.86	0.00
		2	1	4	1.00	0.15	2.31	0.00	0.84	3.83	0.42
		3	1	4	2.00	0.98	3.98	0.15	0.99	3.99	1.26
		4	1	4	3.00	1.00	4.00	1.13	1.00	4.00	2.25
		5	1	4	4.00	1.00	4.00	2.13	1.00	4.00	3.25

Table 5: Simulation results for the design (Σ_2, G_1) with $\rho = 0.5$, $(n, p) = (200, 1000)$ for the normal model, and $(n, p) = (300, 1000)$ for logistic and Poisson models.

Screen	Stop	m	Normal			Logistic			Poisson		
			Sure	TP	FP	Sure	TP	FP	Sure	TP	FP
NIS	-	-	0	3	7.00	0.00	2.98	11.02	0.00	2.95	11.05
gLASSO	-	-	1	4	27.86	1.00	4.00	68.46	0.26	3.23	21.48
gSCAD	-	-	1	4	3.41	0.88	3.88	0.50	0.36	3.12	2.05
FR.full	-	-	1	4	6.00	1.00	4.00	6.00	0.94	3.94	6.06
SC.full	-	-	1	4	6.00	0.98	3.98	5.84	0.34	2.75	7.25
FR	LIC	1	1	4	0.00	0.00	1.36	0.00	0.00	2.98	0.02
		2	1	4	1.00	0.00	2.36	0.00	0.84	3.84	0.17
		3	1	4	2.00	0.36	3.36	0.00	0.92	3.92	1.08
		4	1	4	3.00	1.00	4.00	0.36	0.93	3.93	2.08
		5	1	4	4.00	1.00	4.00	1.36	0.93	3.93	3.08
	HBIC	1	1	4	0.00	0.00	1.00	0.00	0.02	2.98	0.02
		2	1	4	1.00	0.00	2.00	0.00	0.82	3.83	0.18
		3	1	4	2.00	0.00	3.00	0.00	0.92	3.92	1.08
		4	1	4	3.00	1.00	4.00	0.00	0.93	3.93	2.08
		5	1	4	4.00	1.00	4.00	1.00	0.93	3.93	3.08
SC	LIC	1	1	4	0.00	0.00	1.36	0.00	0.00	1.99	0.94
		2	1	4	1.00	0.00	2.36	0.00	0.22	2.29	1.66
		3	1	4	2.00	0.34	3.35	0.01	0.26	2.45	2.50
		4	1	4	3.00	0.98	3.98	0.38	0.28	2.56	3.39
		5	1	4	4.00	0.98	3.98	1.37	0.32	2.64	4.30
	HBIC	1	1	4	0.00	0.00	1.00	0.00	0.00	2.02	0.98
		2	1	4	1.00	0.00	2.00	0.00	0.22	2.40	1.73
		3	1	4	2.00	0.00	3.00	0.00	0.27	2.52	2.65
		4	1	4	3.00	0.98	3.98	0.02	0.30	2.62	3.57
		5	1	4	4.00	0.98	3.98	1.02	0.33	2.65	4.52

Table 6: Simulation results for the design (Σ_2, G_1) with $\rho = 0.5$, $(n, p) = (400, 1000)$ for the normal model, and $(n, p) = (500, 1000)$ for logistic and Poisson models.

Screen	Stop	m	Normal			Logistic			Poisson		
			Sure	TP	FP	Sure	TP	FP	Sure	TP	FP
NIS	-	-	0	3	15.00	0.00	3.00	18.00	0.00	2.99	18.01
gLASSO	-	-	1	4	9.07	1.00	4.00	95.22	0.60	3.60	23.64
gSCAD	-	-	1	4	1.01	1.00	4.00	0.46	0.78	3.75	1.62
FR.full	-	-	1	4	6.00	1.00	4.00	6.00	1.00	4.00	6.00
SC.full	-	-	1	4	6.00	1.00	4.00	5.97	0.85	3.76	6.24
FR	LIC	1	1	4	0.00	0.00	2.02	0.00	0.08	3.08	0.01
		2	1	4	1.00	0.07	3.05	0.00	1.00	4.00	0.09
		3	1	4	2.00	0.98	3.98	0.07	1.00	4.00	1.08
		4	1	4	3.00	1.00	4.00	1.05	1.00	4.00	2.08
		5	1	4	4.00	1.00	4.00	2.05	1.00	4.00	3.08
	HBIC	1	1	4	0.00	0.00	1.02	0.00	0.28	3.28	0.01
		2	1	4	1.00	0.00	2.08	0.00	1.00	4.00	0.30
		3	1	4	2.00	0.10	3.10	0.00	1.00	4.00	1.29
		4	1	4	3.00	1.00	4.00	0.10	1.00	4.00	2.29
		5	1	4	4.00	1.00	4.00	1.09	1.00	4.00	3.29
SC	LIC	1	1	4	0.00	0.00	2.06	0.00	0.02	2.68	0.77
		2	1	4	1.00	0.12	3.10	0.00	0.66	3.38	1.12
		3	1	4	2.00	0.98	3.98	0.12	0.75	3.52	2.00
		4	1	4	3.00	1.00	4.00	1.09	0.77	3.60	2.92
		5	1	4	4.00	1.00	4.00	2.10	0.80	3.66	3.87
	HBIC	1	1	4	0.00	0.00	1.02	0.00	0.14	2.85	0.88
		2	1	4	1.00	0.00	2.09	0.00	0.68	3.48	1.34
		3	1	4	2.00	0.12	3.12	0.00	0.79	3.66	2.22
		4	1	4	3.00	1.00	4.00	0.12	0.81	3.69	3.13
		5	1	4	4.00	1.00	4.00	1.12	0.85	3.76	3.97