Forward variable selection for sparse ultra-high dimensional varying coefficient models

Ming-Yen Cheng, Toshio Honda, and Jin-Ting Zhang

Abstract

Varying coefficient models have numerous applications in a wide scope of scientific areas. While enjoying nice interpretability, they also allow flexibility in modeling dynamic impacts of the covariates. But, in the new era of big data, it is challenging to select the relevant variables when there are a large number of candidates. Recently several works are focused on this important problem based on sparsity assumptions; they are subject to some limitations, however. We introduce an appealing forward variable selection procedure. It selects important variables sequentially according to a reduction in sum of squares criterion and it employs a BIC-based stopping rule. Clearly it is simple to implement and fast to compute, and it possesses many other desirable properties from both theoretical and numerical viewpoints. Notice that the BIC is a special case of the EBIC, when an extra tuning parameter in the latter vanishes. We establish rigorous screening consistency results when either BIC or EBIC is used as the stopping criterion, although the BIC is preferred to the EBIC on the bases of its superior numerical performance and simplicity. The theoretical results depend on some conditions on the eigenvalues related to the design matrices, and we consider the situation where we can relax the conditions on the eigenvalues. Results of an extensive simulation study and a real data example are also presented to show the efficacy and usefulness of our procedure.

Keywords: B-spline; BIC; independence screening; marginal model; semi-varying coefficient models; sub-Gaussion error.

Ming-Yen Cheng is Professor, Department of Mathematics, National Taiwan University, Taipei 106, Taiwan (Email: cheng@math.ntu.edu.tw). Toshio Honda is Professor, Graduate School of Economics, Hitotsubashi University, 2-1 Naka, Kunitachi, Tokyo 186-8601, Japan (Email: t.honda@r.hit-u.ac.jp). Jin-Ting Zhang is Associate Professor, Department of Statistics & Applied Probability, National University of Singapore, 3 Science Drive 2, Singapore 117546 (Email: stazjt@nus.edu.sg). This research is partially supported by the Hitotsubashi International Fellow Program and the Mathematics Division, National Center of Theoretical Sciences (Taipei Office). Cheng is supported by the Ministry of Science and Technology grant 101-2118-M-002-001-MY3. Honda is supported by the JSPS Grant-in-Aids for Scientific Research (A) 24243031 and (C) 25400197. Zhang is supported by the National University of Singapore research grant R-155-000-128-112.

1 Introduction

We consider variable selection problem for the varying coefficient model defined by

$$Y = \sum_{j=0}^{p} \beta_{0j}(T)X_j + \epsilon, \tag{1}$$

where Y is a scalar response variable, $X_0 \equiv 1$, $X_1, \ldots, X_p$ are the candidate covariates, ϵ is the random error, and $T \in [0, 1]$. The coefficient functions β_{0j}, $j = 0, 1, \ldots, p$, are assumed to vary smoothly with T, and are non-zero for only a subset. See Assumption B in Section 3 for the required properties of the β_{0j}'s. The variable T is an influential variable, such as age or income in econometric studies, and is sometimes called the index variable. As for the candidate variables $X_j, \ldots, X_p$, they are uniformly bounded for simplicity of presentation. This assumption can be relaxed as discussed in Remark 3 at the end of Section 3. Since we always include $X_0 \equiv 1$ in the model and consider the sum of squared residuals (see Section 2), we can further assume that $\mathrm{E}\{X_j\} = 0$, $j = 1, \ldots, p$. We also assume that the error ϵ satisfies the sub-Gaussian property uniformly conditionally on the covariates. See Assumption E in Section 3.

The varying coefficient model is a popular and useful structured nonparametric approach to modeling data that may not obey the restrictive form of traditional linear models. While it retains the nice interpretability of the linear models, it also allows for good flexibility in capturing the dynamic impacts of the relevant covariates on the response. In practical applications, some of the true covariates may have simply constant effects while the others have varying effects. Such situation can be easily accommodated by a variant, the so called semi-varying coefficient model [33, 38]. Furthermore, model (1) has been generalized in order to model various data types including count data, binary response, clustered/longitudinal data, time series, etc. We refer to [14] for a comprehensive review.

Due to recent rapid development in technology for data acquisition and storage, nowadays a lot of high-dimensional data sets are collected in various research fields, such as medicine, marketing and so on. In this case, the true model is usually sparse, that is, the number of important covariates is not large even when the dimensionality p is very large. Under this sparsity condition, some effective variable selection procedures are necessary. Existing general penalized variable selection methods include the Lasso [29], group Lasso [23, 36], adaptive Lasso [39], SCAD [8] and Dantzig selector [3].

In ultra-high dimensional cases where p is very large selection consistency, i.e. exactly the true variables are selected with probability tending to 1, becomes challenging and even nearly impossible for existing variable selection methods to achieve. Thus, an additional independence screening step is usually necessary before variable selection is carried out. For example, sure independence screening (SIS) methods are introduced by [9, 11] for linear models and generalized linear models respectively, and nonparametric independence screening (NIS) is suggested for additive models by [7]. Under general parametric models, [12] suggested using the Lasso at the screening stage. Under model (1), there are some existing works on penalized variable selection in several different setups of the dimensionality p, using the Lasso or folded concave penalties such as the SCAD [1, 18, 24, 28, 31, 32, 34]. In ultra-high dimensional cases, for the independence screening purpose, the Lasso is recommended by [32] and NIS is considered by several authors [5, 10, 20, 27]. The NIS procedure of [10] uses marginal spline regression models, and the iterative NIS procedures use group SCAD implicitly. In all of the above mentioned methods, some tuning parameter or threshold value is involved which needs to be determined by the user or by some elaborated means.

More recently, alternative forward selection approaches receive increasing attention in linear regression. This includes the least angle regression [6], the forward iterative regression and shrinkage technique [16], the forward regression [30], the forward Lasso adaptive shrinkage [25], and the sequential Lasso (SLASSO) [22]. Such methods enjoy desirable theoretical properties and have advantages from numerical aspects. The SLASSO employes Lasso in the forward selection of candidate variables and uses the EBIC [4] as the stopping criterion. The consistency result of the EBIC-based model selection in ultra-high dimensional additive models is established by [19] when the number of true covariates is bounded. It assumes some knowledge of the number of true covariates, which may be unrealistic or difficult to obtain in some cases. On the other hand, without this kind of knowledge, the number of all possible subsets of the candidate variables is too large and there is no guarantee that EBIC-based model selection will perform properly. Therefore, it makes sense to consider a forward procedure, which does not require such prior knowledge, and use the EBIC as the stopping criterion.

Motivated by the above facts, we propose and investigate thoroughly a groupwise forward selection procedure for the varying coefficient model in the ultra-high dimensional case. The proposed method is constructed in a spirit similar to that of the SLASSO [22]. However, we recommend to use the BIC rather than the EBIC as the stopping criterion for the following reasons. In our extensive simulation study, we tried the AIC, BIC and EBIC for this purpose, and the BIC yielded the best performance. Also, the

EBIC contains a parameter η which needs to be chosen. Besides, instead of the Lasso we use the reduction in the sum of squared residuals as the selection criterion, as in [30]. This is suggested by our preliminary simulation studies, which showed the latter performs better. It is also more natural as the (E)BIC takes into account the sum of squared residuals while the Lasso is based on the estimated coefficient functions.

Under some assumptions we establish the screening consistency of our forward selection method, that is, all the true variables will be included in the model with probability tending to 1. We consider the effects of eigenvalues related to design matrices explicitly, and the situation where we can relax the conditions on the eigenvalues from a theoretical point of view in Theorem 3.2. Our theoretical results hold when either the EBIC or the BIC is used in the stopping rule, although the latter is preferred. We exploited desirable properties of B-spline bases to drive these theoretical results. After the screening, if necessary, we can identify consistently the true covariates in a second stage by applying the group SCAD or the adaptive group Lasso [5]. Selection consistency of the proposed forward procedure can be obtained under stronger conditions like those in [22]. Such results and the proofs are given in the supplement.

The proposed method has many merits compared to existing methods, from both practical and theoretical viewpoints. First, since the variables are selected sequentially, the final model has good *interpretability* in the sense that we can rank the importance of the variables according to the order they are selected. Second, since no tuning parameters or threshold parameters are present, the implementation and computation are *simple and fast*. Third, we impose assumptions only on the original model, and no faithfulness assumptions in which marginal models reflect the original model is assumed. Fourth, in this article p can be $O(\exp(n^{c_{p2}}))$ for some sufficiently small positive c_{p2}. See Lemma 3.1 and Assumption D in Section 3 for the details. Therefore, the forward procedure can reduce the dimensionality more effectively, as illustrated in Section 4. Finally, our method requires *milder regularity conditions* than the sparse Riesz condition [32] and the restricted eigenvalue conditions [2] for the Lasso, which are related to all the variables.

In Section 2, we describe the proposed forward selection procedure. At each step, it uses the residual sum of squares resulted from spline estimation of an extended marginal model to determine the next candidate feature, and it uses the BIC to decide whether to stop or to include the newly selected feature and continue. In Section 3 we state the assumptions and theoretical results and give some remarks on forward selection for additive coefficient models. Results of simulation studies and a real data example are presented in Section 4. Proofs of the theoretical results are given in Section 5.

2 Method

Before we describe the proposed forward feature selection procedure, we introduce some notation. Let $\#A$ denote the number of elements of a set A and write A^c for the complement of A. We write $\|f\|_{L_2}$ and $\|f\|_\infty$ for the L_2 and sup norm of a function f on $[0,1]$, respectively. When g is a random variable or a function of some random variable(s) we define its L_2 norm by $\|g\| = [\mathrm{E}(g^2)]^{1/2}$. For a k-dimensional vector $\mathbf{x}$, $|\mathbf{x}|$ stands for the Euclidean norm and $\mathbf{x}^T$ is the transpose. We use the same symbol for transpose of matrices. Suppose we have n i.i.d. observations $\{(\mathbf{X}_i, T_i, Y_i)\}_{i=1}^n$, where $\mathbf{X}_i = (X_{i0}, X_{i1}, \ldots, X_{ip})$, taken from the varying coefficient model (1):

$$Y_i = \sum_{j=0}^{p} \beta_{0j}(T_i) X_{ij} + \epsilon_i,\ i = 1, \ldots, n. \tag{2}$$

We write $\mathcal{S}_0$ for the set of indexes of the true covariates in model (1), that is, $\beta_{0j} \not\equiv 0$ for $j \in \mathcal{S}_0$ and $\beta_{0j} \equiv 0$ for $j \in \mathcal{S}_0^c$. In addition, we write p_0 for the number of true covariates, i.e. $p_0 \equiv \#\mathcal{S}_0$. In this paper, our model is sparse and p_0 is much smaller than n and p and we assume that $0 \in \mathcal{S}_0$. In addition, the total number of covariates p can be $O(n^{c_{p1}})$ for any positive c_{p1} or $O(\exp(n^{c_{p2}}))$ for some sufficiently small positive c_{p2}. More details on p will be given later in Lemma 3.1, Assumption D, and Corollary 3.1.

Now we discuss how to construct elements of our method. Suppose that we have selected covariates sequentially as follows:

$$S_1 = \{0\} \subset S_2 \subset \cdots \subset S_k \equiv S.$$

That is, S_j is the index set of the selected covariates upon the completion of the jth step, for $j = 1, \ldots, k$. We can also start with a larger set than just $\{0\}$ according to some *a priori* knowledge. Then, at the current $(k+1)$th step, we need to choose another candidate from S_k^c, and then we need to decide whether we should stop or add it to S_k and go to the next step. Our forward feature selection criterion defined in (9) is based on the reduction in sum of squared residuals, and we employ the BIC as the stopping criterion, which is a special case of the EBIC given in (10) with the parameter $\eta = 0$. See [30] and [22] for more details about forward variable selection and [4] about EBIC in linear regression.

To estimate the coefficient functions in the varying coefficient model (2), we employ an equi-spaced B-spline basis $\boldsymbol{B}(x) = (B_1(x), \dots, B_L(x))^T$ on $[0,1]$, where $L = c_L n^{1/5}$ and the order of the B-spline basis is larger than or equal to two. This is due to our smoothness assumption (Assumption B in Section 3) on the coefficient functions. The existence of the second order derivatives of the coefficient functions is a usual one in the nonparametric literature. See [26] for the definition of B-spline bases. The spline regression model used to approximate model (2) is then given by

$$Y_i = \sum_{j=0}^{p} \boldsymbol{W}_{ij}^T \gamma_{0j} + \epsilon'_i, \tag{3}$$

where $\boldsymbol{W}_{ij} = \boldsymbol{B}(T_i)X_{ij} \in \mathbb{R}^L$, $\boldsymbol{\gamma}_{0j} \in \mathbb{R}^L$, and ϵ'_i is different from ϵ_i in (2).

Now we consider spline estimation of the extended marginal model when we add another index to the index set S, which we will make use of in deriving our forward selection criterion. Hereafter we write $S(l)$ for $S \cup \{l\}$ for any $l \in S^c$. Temporarily we consider the following extended marginal model for $S(l), l \in S^c$:

$$Y = \sum_{j \in S(l)} \overline{\beta}_j(T)X_j + \epsilon_{S(l)}. \tag{4}$$

Here, the coefficient functions $\overline{\beta}_j$, $j \in S(l)$, are defined in terms of minimizing the following mean squared error with respect to β_j, $j \in S(l)$,

$$\mathrm{E}\Big\{\Big(Y - \sum_{j \in S(l)} \beta_j(T)X_j\Big)^2\Big\},$$

where the minimization is over the set of L_2 integrable functions on $[0,1]$. Note that $\|\overline{\beta}_j\|_{L_2}$ should be larger when $j \in \mathcal{S}_0 - S$ than when $j \in \mathcal{S}_0^c$.

We write

$$\boldsymbol{W}_{iS} = (\boldsymbol{W}_{ij}^T)_{j \in S}^T \in \mathbb{R}^{L\#S}, \quad \boldsymbol{W}_j = (\boldsymbol{W}_{1j}, \dots, \boldsymbol{W}_{nj})^T \quad \text{and} \quad \boldsymbol{W}_S = (\boldsymbol{W}_{1S}, \dots, \boldsymbol{W}_{nS})^T.$$

Note that $\boldsymbol{W}_j$ and $\boldsymbol{W}_S$ are respectively $n \times L$ and $n \times (L\#S)$ matrices. Then, the spline model to approximate the extended marginal model (4), in which $l \in S^c$, is given by

$$Y_i = \sum_{j \in S(l)} \overline{\boldsymbol{\gamma}}_j^T \boldsymbol{W}_{ij} + \epsilon'_{iS(l)} = \overline{\boldsymbol{\gamma}}_S^T \boldsymbol{W}_{iS} + \overline{\boldsymbol{\gamma}}_l^T \boldsymbol{W}_{il} + \epsilon'_{iS(l)}, \ i = 1, \dots, n, \tag{5}$$

where $\overline{\boldsymbol{\gamma}}_S^T = (\overline{\boldsymbol{\gamma}}_j^T)_{j \in S}$ and $\overline{\boldsymbol{\gamma}}_j$, $j \in S(l)$, are defined by minimizing with respect to $\boldsymbol{\gamma}_j \in \mathbb{R}^L$, $j \in S(l)$, the following mean squared spline approximation error:

$$\mathrm{E}\Big\{\sum_{i=1}^{n}\big(Y_i - \sum_{j \in S(l)} \boldsymbol{\gamma}_j^T \boldsymbol{W}_{ij}\big)^2\Big\} = \mathrm{E}\Big\{\big|\boldsymbol{Y} - \boldsymbol{W}_S\boldsymbol{\gamma}_S - \boldsymbol{W}_l\boldsymbol{\gamma}_l\big|^2\Big\}$$

with $\boldsymbol{\gamma}_S^T = (\boldsymbol{\gamma}_j^T)_{j\in S}$. Note that $\overline{\boldsymbol{\gamma}}_l^T\mathbf{B}(t)$ should be close to the coefficient function $\overline{\beta}_l(t)$ in the extended marginal model (4). In particular, when $l \in \mathcal{S}_0 - S$, $\|\overline{\beta}_l\|_{L_2}$ should be large enough, and thus $|\overline{\boldsymbol{\gamma}}_l|$ should be also large enough.

We can estimate the vector parameters $\overline{\boldsymbol{\gamma}}_j$, $j \in S(l)$, in model (5) by the ordinary least squares estimates, denoted by $\widehat{\boldsymbol{\gamma}}_j$, $j \in S(l)$. Let $\widehat{\boldsymbol{W}}_{lS}$ and $\widehat{\boldsymbol{Y}}_S$ denote respectively the orthogonal projections of $\boldsymbol{W}_{lS}$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ onto the linear space spanned by the columns of $\boldsymbol{W}_S$, that is,

$$\widehat{\boldsymbol{W}}_{lS} = \boldsymbol{W}_S(\boldsymbol{W}_S^T\boldsymbol{W}_S)^{-1}\boldsymbol{W}_S^T\boldsymbol{W}_l \quad \text{and} \quad \widehat{\boldsymbol{Y}}_S = \boldsymbol{W}_S(\boldsymbol{W}_S^T\boldsymbol{W}_S)^{-1}\boldsymbol{W}_S^T\boldsymbol{Y}\,.$$

Note that $\widehat{\boldsymbol{W}}_{lS}$ is an $n \times L$ matrix. Then $\widehat{\boldsymbol{\gamma}}_l$, the ordinary least square estimate of $\overline{\boldsymbol{\gamma}}_l$, can be expressed as

$$\widehat{\boldsymbol{\gamma}}_l = (\widetilde{\boldsymbol{W}}_{lS}^T\widetilde{\boldsymbol{W}}_{lS})^{-1}\widetilde{\boldsymbol{W}}_{lS}^T\widetilde{\boldsymbol{Y}}_S, \tag{6}$$

where $\widetilde{\boldsymbol{W}}_{lS} = \boldsymbol{W}_l - \widehat{\boldsymbol{W}}_{lS}$ and $\widetilde{\boldsymbol{Y}}_S = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}_S$.

Recall that at the current step we are given S, the index set of the covariates already selected, and the job is to choose from S^c another candidate and then decide whether we should add it to S or we should not and stop. To select the next candidate, we consider the reduction in the sum of squared residuals (SSR) when adding $l \in S^c$ to the model. Specifically, we compute $n\widehat{\sigma}_S^2 - n\widehat{\sigma}_{S(l)}^2$, where for an index set Q the SSR $n\widehat{\sigma}_Q^2$ is given by

$$n\widehat{\sigma}_Q^2 = \boldsymbol{Y}^T\boldsymbol{Y} - \boldsymbol{Y}^T\boldsymbol{W}_Q(\boldsymbol{W}_Q^T\boldsymbol{W}_Q)^{-1}\boldsymbol{W}_Q^T\boldsymbol{Y} \tag{7}$$

and is the sum of squared residuals from least squares regression of

$$Y_i = \sum_{j\in Q}\boldsymbol{W}_{ij}^T\boldsymbol{\gamma}_j + \epsilon_i'.$$

Using (6), we can rewrite $n(\widehat{\sigma}_S^2 - \widehat{\sigma}_{S(l)}^2)$ as

$$\begin{aligned} n(\widehat{\sigma}_S^2 - \widehat{\sigma}_{S(l)}^2) &= \big(\widetilde{\boldsymbol{W}}_{lS}^T\widetilde{\boldsymbol{Y}}_S\big)^T\big(\widetilde{\boldsymbol{W}}_{lS}^T\widetilde{\boldsymbol{W}}_{lS}\big)^{-1}\big(\widetilde{\boldsymbol{W}}_{lS}^T\widetilde{\boldsymbol{Y}}_S\big) \\ &= \widehat{\boldsymbol{\gamma}}_l^T\Big(\widetilde{\boldsymbol{W}}_{lS}^T\widetilde{\boldsymbol{W}}_{lS}\Big)\widehat{\boldsymbol{\gamma}}_l \;\approx\; n\mathrm{E}\Big\{\big(\overline{\beta}_l(T)\widetilde{X}_{lS}\big)^2\Big\}, \end{aligned} \tag{8}$$

where $\widetilde{X}_{lS} = X_l - \widehat{X}_{lS}$ and $\widehat{X}_{lS}$ is the projection of X_l onto $\{\sum_{j\in S}\beta_j(T)X_j\}$ with respect to the L_2 norm $\|\cdot\|$. As noted earlier, if $l \in \mathcal{S}_0$ then $\|\overline{\beta}_l\|_{L_2}$ will be large enough. Hence, following from expression (8) and noticing that $\widehat{\boldsymbol{\gamma}}_l^T\mathbf{B}(t)$ is the spline estimate of $\overline{\beta}_l(t)$ in the extended marginal model (4), we choose the candidate index as

$$l^* = \underset{l\in S^c}{\operatorname{argmin}}\,\widehat{\sigma}_{S(l)}^2\,. \tag{9}$$

Then, we have high confidence that l^* belongs to $\mathcal{S}_0 - S$ provided that the latter is non-empty.

To determine whether or not to include the candidate feature X_{l^*} in the set of selected ones, we employ the BIC criterion. Since the BIC criterion is a special case of the EBIC, we define the EBIC of a subset of covariates indexed by Q as the following:

$$\text{EBIC}(Q) = n \log(\widehat{\sigma}^2_Q) + \#Q \times L(\log n + 2\eta \log p), \tag{10}$$

where η is a fixed nonnegative constant and $n\widehat{\sigma}^2_Q$ is given in (7). When $\eta = 0$, the EBIC is the BIC. Then, we should include the new candidate covariate X_{l^*}, where l^* is defined in (9), provided that the BIC decreases when we add l^* to S and form $S(l^*)$. Otherwise, if the BIC increases, we should not select any more covariates and stop at the kth step.

In the following, we define formally the proposed forward feature selection algorithm. We can also implement it with the BIC replaced by the EBIC throughout.

Initial step: Take $S_1 = \{0\}$ or a larger set based on some *a priori* knowledge of $\mathcal{S}_0$ and compute BIC(S_1).

Sequential selection: At the $(k+1)$th step, compute $\widehat{\sigma}^2_{S_k(l)}$ for every $l \in S_k^c$, and find

$$l^*_{k+1} = \operatorname*{argmin}_{l \in S_k^c} \widehat{\sigma}^2_{S_k(l)} \,.$$

Then, let $S_{k+1} = S_k \cup \{l^*_{k+1}\}$ and compute BIC(S_{k+1}).Stop and declare S_k as the set of selected covariate indexes if BIC$(S_{k+1}) >$ BIC(S_k); otherwise, change k to $k+1$ and continue to search for the next candidate feature.

Remark 1 *In practice, the forward procedure with the EBIC or the BIC stopping rule may stop a little too early due to rounding errors and fails to select some relevant variables. For example, it may happen that the stopping criterion value drops in one step, then increases in the next step, and then drops again. To avoid interference caused by such small fluctuations, we can continue the forward selection process until the stopping criterion value continuously increases for several consecutive steps before the forward selection procedure is stopped.*

Remark 2 *At first, instead of (9), we considered choosing the candidate index as*

$$l^\dagger = \operatorname*{argmax}_{j \in S^c} \left| \widetilde{\boldsymbol{W}}^T_{lS} \widetilde{\boldsymbol{Y}}_S \right| \tag{11}$$

as the next candidate index, as motivated by the sequential Lasso for linear models proposed by [22]. However, after some preliminary simulation studies we found that (9) performs better. The intuition is that in each step the selection criterion (9) gives the smallest value of the stopping criterion value, while this property is not necessarily true for (11).

3 Assumptions and theoretical properties

In this section, we describe technical assumptions and present theoretical properties of the proposed forward procedure. Specifically, we prove a desirable property of the sum of regression residuals in Theorem 3.1 and then establish the screening consistency in Corollaries 3.1 and 3.2. In Theorem 3.2, we consider relaxing the eigenvalue conditions in Assumption V. Note that we treat the EBIC and the BIC (when $\eta = 0$ in the definition of EBIC) in a unified way. Therefore the theoretical results hold when either of them is used in the stopping rule. The proofs are given in Section 5.

First, we define some notation and symbols. For a vector of regression coefficient functions $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$, we define its L_2 norm by $\|\boldsymbol{\beta}\|_{L_2}^2 = \sum_{j=1}^p \|\beta_j\|_{L_2}^2$. We denote the maximum and minimum eigenvalues of a symmetric matrix $\mathbf{A}$ by $\lambda_{\max}(\mathbf{A})$ and $\lambda_{\min}(\mathbf{A})$, respectively. Let $\mathbf{I}_k$ be the k dimensional identity matrix.

The following assumption assures that regression coefficient functions can be approximated accurately enough by spline functions.

Assumption B: For some positive constant C_{B0}, the coefficient functions $\{\beta_{0j} \mid j \in \mathcal{S}_0\}$ are twice differentiable and satisfy

$$\sum_{j \in \mathcal{S}_0} \|\beta_{0j}\|_\infty < C_{B0}, \quad \sum_{j \in \mathcal{S}_0} \|\beta_{0j}''\|_\infty < C_{B0}, \quad \text{and} \quad L^2 \min_{j \in \mathcal{S}_0} \|\beta_{0j}\|_{L_2} \to \infty.$$

We present another expression of (2) under Assumption B:

$$Y_i = \sum_{j \in \mathcal{S}_0} \boldsymbol{W}_{ij}^T \gamma_j^* + r_i + \epsilon_i, \; i = 1, \dots, n, \tag{12}$$

where, for some positive constant C_r, $\{r_i\}$ satisfies

$$\frac{1}{n} \sum_{i=1}^n r_i^2 \le C_r L^{-4} \tag{13}$$

with probability tending to 1, and for some positive constants C_{B1} and C_{B2}, $\{\gamma_j^* \mid j \in \mathcal{S}_0\}$ satisfies

$$C_{B1} L \|\beta_{0j}\|_{L_2}^2 \leq |\gamma_j^*|^2 \leq C_{B2} L \|\beta_{0j}\|_{L_2}^2 \tag{14}$$

for any $j \in \mathcal{S}_0$. Note that C_r, C_{B1}, and C_{B2} depend on C_{B0}. Properties (13) and (14) follow from Assumption B and are standard results in spline function approximation. For example, see Corollary 6.26 of [26].

The following sub-Gaussian assumption is necessary when we evaluate quadratic forms of $\{\epsilon_i\}$.

Assumption E: There is a positive constant C_ϵ such that, for any $u \in \mathbb{R}$,

$$\mathrm{E}\{\exp(u\epsilon) \mid X_1, \ldots, X_p\} \leq \exp(C_\epsilon u^2/2)\,.$$

Recall that we start with $S_1 = \{0\}$ or a larger set, so hereafter let all the subsets denoted by S satisfy $\{0\} \subset S$. We can say that Assumption V given in the following is the definition of $\tau_{\min}(M)$ and $\tau_{\max}(M)$ except the positivity of $\tau_{\min}(M)$.

Assumption V: For any given integer M larger than p_0, there exist positive functions $\tau_{\min}(M)$ and $\tau_{\max}(M)$ such that

$$\frac{\tau_{\min}(M)}{L} \leq \lambda_{\min}(\mathrm{E}\{\boldsymbol{W}_{1S}\boldsymbol{W}_{1S}^T\}) \leq \lambda_{\max}(\mathrm{E}\{\boldsymbol{W}_{1S}\boldsymbol{W}_{1S}^T\}) \leq \frac{\tau_{\max}(M)}{L}$$

uniformly in $S \subset \{0, 1, \ldots, p\}$ satisfying $\#S \leq M$.

To give an idea when Assumption V will hold, consider the case where there are positive functions $\tau_{\min}^X(M)$ and $\tau_{\max}^X(M)$ such that

$$\tau_{\min}^X(M) \leq \lambda_{\min}(\mathrm{E}\{\boldsymbol{X}_{1S}\boldsymbol{X}_{1S}^T | T\}) \leq \lambda_{\max}(\mathrm{E}\{\boldsymbol{X}_{1S}\boldsymbol{X}_{1S}^T | T\}) \leq \tau_{\max}^X(M)$$

uniformly in $S \subset \{0, 1, \ldots, p\}$ satisfying $\#S \leq M$, where $\boldsymbol{X}_{iS}$ is defined in the same way as $\boldsymbol{W}_{iS}$, we have

$$\tau_{\min}^X(M)/C < \tau_{\min}(M) < C\tau_{\min}^X(M) \text{ and } \tau_{\max}^X(M)/C < \tau_{\max}(M) < C\tau_{\max}^X(M)$$

for some positive constant C.

Our results crucially depend on $\tau_{\min}(M)$ and $\tau_{\max}(M)$ as in [30], [15], and other papers on variable selection. However, it seems to be almost impossible to evaluate $\tau_{\min}(M)$ and $\tau_{\max}(M)$ explicitly. In some papers, for example [30] and [17], $\tau_{\min}(M)$ and

$\tau_{\max}(M)$ are assumed to be constant. In Theorem 3.2, we show that we may be able to relax the uniformity requirement on the eigenvalues in Assumption V.

We evaluate sample covariance matrices of the covariates in Lemma 3.1. We exploit the local properties of B-spline bases in its proof.

Lemma 3.1 *Suppose that Assumption V holds with $M^2 \max\{\log p, \log n\} = o(n^{4/5}(\tau_{\min}(M))^2)$. Then, with probability tending to 1, we have uniformly in S satisfying $\#S \leq M$ that*

$$\frac{\tau_{\min}(M)}{L}(1-\delta_{1n}) \leq \lambda_{\min}(n^{-1}\boldsymbol{W}_S^T\boldsymbol{W}_S) \leq \lambda_{\max}(n^{-1}\boldsymbol{W}_S^T\boldsymbol{W}_S) \leq \frac{\tau_{\max}(M)}{L}(1+\delta_{1n}),$$

where $\{\delta_{1n}\}$ is a sequence of positive numbers tending to 0 sufficiently slowly.

We use the following assumption to determine the lower bound of the reduction in the sum of squared residuals when $\mathcal{S}_0 \not\subset S$, given in Theorem 3.1. First, we set

$$D_M = p_0^{-1}\frac{C_{B1}^2}{C_{B2}}\frac{\tau_{\min}^2(M)}{\tau_{\max}(M)}\frac{\min_{j\in\mathcal{S}_0}\|\beta_{0j}\|_{L_2}^4}{\|\boldsymbol{\beta}_0\|_{L_2}^2}.$$

This D_M is similar to an expression in (B.7) of [30].

Assumption D: For some positive $d_m < 1$,

$$\frac{D_M L^4}{(\log n)^{d_m}} \to \infty \quad \text{and} \quad M\log p = O(L(\log n)^{d_m}).$$

The second condition in Assumption D is necessary in order to evaluate quadratic forms of $\{\epsilon_i\}$ uniformly in S. We can deal with ultra-high dimensional cases if M increases slowly; see the condition on p given in Lemma 3.1.

Theorem 3.1 *Suppose that assumptions B, E, V and D, and those in Lemma 3.1 hold. Then there exists a sequence of positive numbers $\{\delta_{2n}\}$ satisfying $\delta_{2n} \to 0$ and $\sqrt{n}\delta_{2n} \to \infty$ such that, with probability tending to 1,*

$$\max_{l\in S^c}\{n\widehat{\sigma}_S^2 - n\widehat{\sigma}_{S(l)}^2\} \geq n(1-\delta_{2n})D_M$$

uniformly in S satisfying $\#(S\cup\mathcal{S}_0) \leq M$ and $\mathcal{S}_0 \not\subset S$.

Let T_M be the smallest integer which is larger than or equal to

$$\frac{\mathrm{Var}(Y)}{(1-2\delta_{2n})D_M}.$$

Then, following from Theorem 3.1, Corollary 3.1 gives a sufficient condition for screening consistency of the forward procedure.

Corollary 3.1 *Suppose that we have the same assumptions as in Theorem 3.1. If* $T_M \leq M - p_0$ *and* $L(\log n + 2\eta \log p) = o(nD_M)$*, then we have*

$$\mathcal{S}_0 \subset S_k \quad for\ some\ k \leq T_M$$

with probability tending to 1.

To give an idea of when the conditions on M in Corollary 3.1 would hold, suppose p_0 and $\{\beta_{0j} \,|\, j \in \mathcal{S}_0\}$ are independent of n. Then a sufficient condition for the existence of M in Corollary 3.1 is that for some $M = M_n \to \infty$ sufficiently slowly, we have

$$\frac{\tau_{\max}(M_n)}{M_n \tau_{\min}^2(M_n)} \to 0.$$

This condition on $\tau_{\max}(M)$ and $\tau_{\min}(M)$ may be restrictive and difficult to check. Nevertheless, Theorem 3.2 given later suggests that we may be able to relax the uniformity requirement on the eigenvalues in Assumption V in some cases.

The following corollary follows easily from the proof of Theorem 3.1. A similar result is given in [19]. It assures the screening consistency of the proposed forward procedure and we expect it to stop early with $\mathcal{S}_0 \subset S_k$ for some k. Then, if necessary, following the screening we can carry out variable selection using adaptive Lasso or SCAD procedures to find exactly the true model or equivalently $\mathcal{S}_0$.

Corollary 3.2 *Suppose that we have the same assumptions as in Theorem 3.1. If* $\mathcal{S}_0 \not\subset S_{k-1}$*,* $\mathcal{S}_0 \subset S_k$*, and* $k \leq M$*, our forward selection procedure stops at the kth step with probability tending to 1.*

Next we consider relaxing the uniformity condition in Assumption V. Suppose the covariate indexes can be divided into $\mathcal{V}$ and $\mathcal{V}^c$ such that $\mathcal{S}_0 \subset \mathcal{V}$ and

$$\mathrm{E}\{X_{1j}B_l(T_1)X_{1s}B_m(T_1)\} = \begin{cases} 0, & B_l(t)B_m(t) \equiv 0 \\ O(\kappa_{0n}/L), & B_l(t)B_m(t) \not\equiv 0 \end{cases}, \tag{15}$$

uniformly in l, m, $j \in \mathcal{V}^c$, and $s \in \mathcal{V}$, where $\{\kappa_{0n}\}$ is a sequence satisfying $\kappa_{0n} \to 0$. We define $\tau_{\min}^{\mathcal{V}}(M)$ and $\tau_{\max}^{\mathcal{V}}(M)$ in the same way as $\tau_{\min}(M)$ and $\tau_{\max}(M)$ in Assumption V but consider only $S \subset \mathcal{V}$. Besides, we define $\tau_{\max}^{\mathcal{S}_0}$ by

$$\frac{\tau_{\max}^{\mathcal{S}_0}}{L} = \lambda_{\max}(\mathrm{E}\{\boldsymbol{W}_{1\mathcal{S}_0}\boldsymbol{W}_{1\mathcal{S}_0}^T\}).$$

The following assumption is about the correlations between covariates indexed by $\mathcal{V}$ and those indexed by $\mathcal{V}^c$.

Assumption Z : Assume that there exists some $\kappa_n \to 0$ such that

$$\frac{1}{n}\sum_{i=1}^{n} X_{ij}B_l(T_i)X_{is}B_m(T_i) = \begin{cases} 0, & B_l(t)B_m(t) \equiv 0 \\ O_p(\kappa_n/L), & B_l(t)B_m(t) \not\equiv 0 \end{cases}$$

and

$$\lambda_{\min}(n^{-1}\boldsymbol{W}_j^T\boldsymbol{W}_j) > \frac{C_m}{L}(1+o_p(1))$$

uniformly in $j \in \mathcal{V}^c$ and $s \in \mathcal{V}$, where C_m is a positive constant. We assume $\kappa_n^2 M/\tau_{\min}^{\mathcal{V}}(M) \to 0$ for simplicity of presentation.

Assumption Z may look a little general and we give an example when it holds. When $\{X_j \,|\, j \in \mathcal{V}\} \cup \{T\}$ and $\{X_j \,|\, j \in \mathcal{V}^c\}$ are mutually independent and $p = O(n^{c_{p1}})$ for some positive fixed constant c_{p1}, we have $\kappa_{0n} = 0$ and

$$\kappa_n = n^{-2/5}\sqrt{\log n}$$

with some regularity conditions. Recall that $\mathrm{E}\{X_j\} = 0$, $j = 1, \dots, p$, in this paper. In general we have $\kappa_n = \kappa_{0n} + n^{-2/5}\sqrt{\log n}$.

Next we give an upper bound for the reduction in the sum of squared residuals.

Theorem 3.2 *Suppose that Assumptions B, E, and Z and those in Lemma 3.1 for $\mathcal{V}$ hold. Then we have, with probability tending to 1,*

$$\max_{l \in S^c \cap \mathcal{V}^c}\{n\widehat{\sigma}_S^2 - n\widehat{\sigma}_{S(l)}^2\} \le C\Big(\frac{nLp_0\kappa_n^2}{C_m} + \frac{nL\tau_{\max}^{\mathcal{S}_0}M\kappa_n^2}{C_m\tau_{\min}^{\mathcal{V}}(M)} + \frac{n}{L^4} + o(L\log n)\Big)$$

uniformly in S satisfying $\#S \le M$ and $S \subset \mathcal{V}$, where C is some positive constant.

We define $D_M^{\mathcal{V}}$ by replacing $\tau_{\min}(M)$ and $\tau_{\max}(M)$ with $\tau_{\min}^{\mathcal{V}}(M)$ and $\tau_{\max}^{\mathcal{V}}(M)$ in the definition of D_M and confine the uniformity requirement in Assumption V to $\mathcal{V}$. If

$$L(\log n + 2\eta\log p) = o(nD_M^{\mathcal{V}}), \quad \frac{p_0L\kappa_n^2}{C_m} = o(D_M^{\mathcal{V}}), \quad \text{and} \quad \frac{\tau_{\max}^{\mathcal{S}_0}LM\kappa_n^2}{C_m\tau_{\min}^{\mathcal{V}}(M)} = o(D_M^{\mathcal{V}})$$

under the assumptions of Theorem 3.1, Theorem 3.2 implies we will never choose any covariates from $\mathcal{V}^c$ with probability tending to 1. Then we actually have only to consider the uniformity requirement in Assumption V on $\mathcal{V}$. When $\#\mathcal{V}$ grows slowly compared to p, $D_M^{\mathcal{V}}$ also decreases slowly compared to M and the assumption on T_M in Corollary 3.1 will hold with D_M replaced by $D_M^{\mathcal{V}}$. Then we will still have screening consistency.

Before closing this section, we discuss the boundedness condition on the covariates in Remark 3 and we comment on forward selection for additive models in Remark 4.

Remark 3 *The boundedness condition on X_j, $j = 1, \ldots, p$, can be relaxed. We use the condition only for the evaluation of $\sum_{i=1}^n r_i^2$ and when applying the Bernstein inequality. Recall d_m is defined in Assumption D. If instead we have*

$$\mathrm{E}\Big\{\sum_{j\in\mathcal{S}_0} X_j^2\Big\} = O((\log n)^{d_m}) \qquad \text{and} \qquad \max_{1\le j\le p} |X_j| \le C\sqrt{\frac{n}{L\log n}},$$

then we will still have the same theoretical results. Since we have from Assumption B and the former condition in the above that

$$\frac{1}{n}\sum_{i=1}^n r_i^2 = O(L^{-4})\frac{1}{n}\sum_{i=1}^n\sum_{j\in\mathcal{S}_0} X_{ij}^2 = O_p(L^{-4}(\log n)^{d_m}),$$

$\sum_{i=1}^n r_i^2$ is still negligible. Besides, the latter condition assures the same uniform convergence rate based on the Bernstein inequality.

Remark 4 *We can handle other models that can be represented as in (12) in almost the same way, for example, additive models as in [17]. When we consider the additive model:*

$$Y_i = \sum_{j\in\mathcal{S}_0} f_j(X_{ij}) + \epsilon_i,\ i = 1, \ldots, n,$$

we have an expression similar to (12):

$$Y_i = \mu^* + \sum_{j\in\mathcal{S}_0} \boldsymbol{W}_{ij}^T\boldsymbol{\gamma}_j^* + r_i + \epsilon_i,\ i = 1, \ldots, n,$$

where $\boldsymbol{W}_{ij}$ are defined conformably with the additive model. Therefore we can deal with additive models in almost the same way as in this paper, except that we cannot fully use the local properties of B-spline bases due to the identification issue.

4 Simulation and empirical studies

We carried out some simulation studies and a real data analysis based on a genetic dataset to assess the performance of the proposed forward feature selection method with AIC, BIC or EBIC as the stopping criterion. For simplicity, we denote these three variants by fAIC, fBIC and fEBIC respectively. At the initial step of the forward selection, we let $S_1 = \{0\}$. To prevent the forward procedures from stopping too early and missing some true variables, we terminated the sequential selection only if the stopping criterion value increases for five consecutive steps, as suggested in Remark 1. The value of the parameter η in the definition of EBIC was taken as $\eta = 1 - \log(n)/(3\log p)$, as suggested by [4]. Notice that BIC can be regarded as a special version of EBIC when $\eta = 0$ while AIC can be obtained from BIC via replacing $\log(n)$ with 2. This shows that the penalty terms of AIC, BIC, and EBIC are getting larger in turns and hence the model selected by fAIC is the largest, followed by the one selected by fBIC, while the one selected by fEBIC is the smallest. Also notice that the model selected by fAIC is usually not consistent while, under some regularity conditions, the models selected by fBIC and fEBIC can be consistent when n is sufficiently large.

In the simulation studies, we generated data from the two varying coefficient models studied by [10] under different levels of correlation. Following the paper, we used the cubic B-spline with $L = [2n^{1/5}]$ where $[\cdot]$ denotes the function rounding to the nearest integer. We set the sample size as $n = 200, 300$, or 400 and the number of covariates as $p = 1000$, and we repeated each of the simulation configurations for $N = 1000$ times. The experiments were run on a Dell PC intel Core i7 vPro, in Matlab and in Windows 7 environment.

4.1 Simulation Studies

In this section, we compare the finite sample performance of the fAIC, fBIC and fEBIC with that of the greedy-INIS algorithm of [10] (denoted as gNIS for simplicity of notation). Note that, based on the simulation results presented in [10], their conditional-INIS approach performs similarly to the greedy-INIS approach. Thus, we shall not include it in the comparison for the shake of time saving.

Example 1 *Following Example 3 of [10], we generated N samples from the following varying coefficient model:*

$$Y = 2 \cdot X_1 + 3T \cdot X_2 + (T+1)^2 \cdot X_3 + \frac{4\sin(2\pi T)}{2 - \sin(2\pi T)} \cdot X_4 + \sigma\epsilon,$$

where $X_j = (Z_j + t_1 U_1)/(1+t_1), j = 1, 2, \cdots, p$, and $T = (U_2 + t_2 U_1)/(1+t_2)$, with $Z_1, Z_2, \cdots, Z_p \overset{i.i.d}{\sim} N(0,1), U_1, U_2 \overset{i.i.d}{\sim} U(0,1)$, and $\epsilon \sim N(0,1)$ being all mutually independent with each other. The noise variance σ is used to control the noise level.

Table 1: Correlations between the covariates X_j's and the index variable T.

$[t_1, t_2]$	$[0, 0]$	$[2, 0]$	$[2, 1]$	$[3, 1]$	$[4, 5]$	$[6, 8]$
$\mathrm{corr}(X_j, X_k)$	0	0.25	0.25	0.43	0.57	0.75
$\mathrm{corr}(X_j, T)$	0	0	0.36	0.46	0.74	0.86

In this example, the number of true covariates p_0 is four. The parameters t_1 and t_2 control the correlations between the covariates X_j's and the index covariate T. It is easy to show that $\mathrm{corr}(X_j, X_k) = t_1^2/(12 + t_1^2)$ for any $j \neq k$, and $\mathrm{corr}(X_j, T) = t_1 t_2/[(12+t_1^2)(1+t_2^2)]^{1/2}$, independent of j. Table 1 lists the values of $[t_1, t_2]$ which define six cases of the correlations. In the first case the X_j's and T are all uncorrelated. In the second case the X_j's are correlated but they are uncorrelated with T. In the last four cases the X_j's are increasingly correlated and the correlations between the X_j's and T are also increasing. Notice that the correlations in the last two cases are rather big and it is quite challenging to correctly identify the true covariates.

In Table 2, we report the average numbers of true positive (TP) and false positive (FP) selections, and their robust standard deviations (in parentheses) for all the four considered approaches for $n = 200, 300$ or 400, when $\sigma = 1$ and $p = 1000$. The tuning parameters $[t_1, t_2]$, the signal-to-noise-ratio, denoted by SNR and defined as $\mathrm{Var}(\beta^T(T)\mathbf{X})/\mathrm{Var}(\epsilon)$, and the approaches considered are listed in the first three columns. Notice that the larger the TP values and the smaller the FP values are, the better the associated approach performs. From this point of view, it is seen that with the correlations between X_j's and the correlations between X_j and T increasing, the performance of the approaches is generally getting worse. This is not surprising since when the correlations are getting larger, it is more challenging to distinguish the true covariates from the false covariates. At the same time, with n increasing from 200 to 400, the performance of all the approaches is generally getting better. This is also expected since larger sample sizes generally provide more information and better estimation of the underlying quantities.

We now compare the performance of the four approaches. From Table 2, it is seen that for each setting, the fAIC selects more true covariates than the other three approaches but it also selects much more false covariates. In addition, with increasing n, the fAIC performs worse in terms of selecting more false covariates. This is due to the fact that in the penalty term of AIC the sample size n is not involved while in the penalty terms of BIC and EBIC n is involved to different degrees. From this point of

Table 2: Average numbers of true positive (TP) and false positive (FP) over 1000 repetitions and their robust standard deviations (in parentheses) for the gNIS, fAIC, fBIC and fEBIC approaches under the varying coefficient model defined in Example 1 with $\sigma = 1$ and $p = 1000$.

			$n = 200$		$n = 300$		$n = 400$	
$[t_1, t_2]$	SNR	Approach	TP	FP	TP	FP	TP	FP
$[0, 0]$	17.15	gNIS	3.99(0.00)	3.10(1.49)	4.00(0.00)	0.37(0.75)	4.00(0.00)	0.08(0.00)
		fAIC	4.00(0.00)	65.8(11.2)	4.00(0.00)	83.2(13.4)	4.00(0.00)	106(16.40)
		fBIC	4.00(0.00)	5.73(0.00)	4.00(0.00)	0.00(0.00)	4.00(0.00)	0.00(0.00)
		fEBIC	4.00(0.00)	0.00(0.00)	4.00(0.00)	0.00(0.00)	4.00(0.00)	0.00(0.00)
$[2, 0]$	3.71	gNIS	3.84(0.00)	0.31(0.00)	3.95(0.00)	0.00(0.00)	3.99(0.00)	0.00(0.00)
		fAIC	3.99(0.00)	68.2(13.4)	4.00(0.00)	85.3(14.9)	4.00(0.00)	106(17.91)
		fBIC	3.99(0.00)	6.16(0.00)	4.00(0.00)	0.01(0.00)	4.00(0.00)	0.00(0.00)
		fEBIC	1.64(0.00)	0.00(0.00)	2.56(0.00)	0.00(0.00)	3.73(0.00)	0.00(0.00)
$[2, 1]$	3.24	gNIS	3.50(0.00)	1.72(2.23)	3.81(0.00)	1.16(1.49)	3.96(0.00)	0.34(0.00)
		fAIC	4.00(0.00)	82.9(14.9)	4.00(0.00)	105(16.41)	4.00(0.00)	134(20.89)
		fBIC	3.98(0.00)	2.24(0.00)	4.00(0.00)	0.01(0.00)	4.00(0.00)	0.00(0.00)
		fEBIC	1.34(0.75)	0.00(0.00)	2.24(0.00)	0.00(0.00)	3.41(0.75)	0.00(0.00)
$[3, 1]$	2.85	gNIS	3.22(0.75)	0.82(0.75)	3.63(0.75)	0.53(0.75)	3.84(0.00)	0.19(0.00)
		fAIC	3.88(0.00)	82.1(14.2)	3.99(0.00)	106(18.66)	4.00(0.00)	132(23.88)
		fBIC	3.58(0.75)	1.99(0.00)	3.89(0.00)	0.01(0.00)	3.99(0.00)	0.00(0.00)
		fEBIC	1.10(0.00)	0.01(0.00)	1.68(0.75)	0.00(0.00)	2.14(0.00)	0.00(0.00)
$[4, 5]$	2.96	gNIS	1.55(0.75)	0.11(0.00)	2.39(0.75)	0.05(0.00)	3.28(0.75)	0.04(0.00)
		fAIC	3.47(0.75)	74.6(16.4)	3.91(0.00)	92.7(18.7)	3.99(0.00)	115.7(20)
		fBIC	2.82(0.75)	2.65(0.75)	3.40(0.75)	0.02(0.00)	3.80(0.00)	0.00(0.00)
		fEBIC	0.98(0.00)	0.02(0.00)	1.09(0.00)	0.00(0.00)	1.72(0.75)	0.00(0.00)
$[6, 8]$	3.16	gNIS	0.26(0.00)	0.02(0.00)	0.47(0.75)	0.01(0.00)	1.09(1.49)	0.00(0.00)
		fAIC	2.37(0.75)	73.7(14.9)	3.26(0.75)	90.7(19.4)	3.75(0.00)	111(22.76)
		fBIC	1.18(0.00)	0.99(0.75)	1.85(0.75)	0.09(0.00)	2.58(0.75)	0.04(0.00)
		fEBIC	0.86(0.00)	0.14(0.00)	0.98(0.00)	0.02(0.00)	1.03(0.00)	0.01(0.00)

view, the fAIC is not recommended. The fEBIC, on the other hand, selects much fewer true covariates than the other three approaches in various settings although it also selects slightly fewer false covariates. This is because the penalty term of EBIC involves the dimensionality p and is much larger than the penalty terms of AIC and BIC. For the forward selection procedures, the sub-models under consideration are usually much smaller than the full model with p covariates, thus the EBIC may be less appropriate in this context. It is not difficult to see that the fBIC generally outperforms the fAIC and fEBIC in terms of selecting more true covariates and less false covariates. In addition, under various settings, we can see that the fBIC in general outperforms the gNIS approach. When $n = 300$ and 400 the fBIC generally selects more true covariates and fewer false covariates than the gNIS does, and when $n = 200$ it selects more false covariates but it also selects much more true covariates.

The varying coefficient model in Example 1 has only four true underlying covariates. In the varying coefficient model defined in the following example, there are eight true underlying covariates.

Example 2 *Following Example 4 of [10], we generated N samples from the following varying coefficient model:*

$$\begin{aligned} Y = \quad & 3T \cdot X_1 + (T+1)^2 \cdot X_2 + (T-2)^3 \cdot X_3 + 3 \cdot \sin(2\pi T) \cdot X_4 \\ & + \exp(T) \cdot X_5 + 2 \cdot X_6 + 2 \cdot X_7 + 3\sqrt{T} \cdot X_8 + \sigma\epsilon, \end{aligned}$$

while T, $\mathbf{X}$, Y and ϵ were generated in the same way as described in Example 1 and again the noise variance σ is used to control the noise level.

Table 3 displays the simulation results under the varying coefficient model defined in Example 2 with $\sigma = 1$ and $p = 1000$. The conclusions are similar to those drawn based on the simulation results summarized in Table 2. That is, the fAIC selects too many false covariates while the fEBIC selects too fewer true covariates. The fBIC approach generally outperforms the other three.

We also repeated the above simulation studies with $\sigma = 2$. To save space the simulation results are not presented here and are place in the supplement. The above conclusions can also be drawn similarly except now all the four approaches perform worse than they do when $\sigma = 1$ as presented in Tables 2 and 3.

In addition, we would like to mention that the gNIS approach often uses more computational time than the fAIC approach does. They both sometimes take 10 times

Table 3: The same caption as that of Table 2 but now under the varying coefficient model defined in Example 2 with $\sigma = 1$ and $p = 1000$.

			$n = 200$		$n = 300$		$n = 400$	
$[t_1, t_2]$	SNR	Approach	TP	FP	TP	FP	TP	FP
$[0, 0]$	47.94	gNIS	7.94(0.00)	7.58(11.9)	7.98(0.00)	1.64(0.75)	8.00(0.00)	0.17(0.00)
		fAIC	8.00(0.00)	61.1(11.2)	8.00(0.00)	78.7(12.7)	8.00(0.00)	102(16.42)
		fBIC	8.00(0.00)	48.1(10.5)	8.00(0.00)	0.34(0.00)	8.00(0.00)	0.00(0.00)
		fEBIC	1.61(0.00)	0.00(0.00)	5.48(5.22)	0.00(0.00)	8.00(0.00)	0.00(0.00)
$[2, 0]$	9.46	gNIS	7.41(0.00)	3.37(2.24)	7.92(0.00)	0.04(0.00)	7.98(0.00)	0.00(0.00)
		fAIC	7.96(0.00)	64.7(13.4)	8.00(0.00)	82.4(15.7)	8.00(0.00)	104(17.91)
		fBIC	7.94(0.00)	53.3(14.9)	8.00(0.00)	0.85(0.00)	8.00(0.00)	0.00(0.00)
		fEBIC	0.96(0.00)	0.05(0.00)	1.40(0.00)	0.01(0.00)	5.32(3.73)	0.00(0.00)
$[2, 1]$	8.68	gNIS	4.71(5.97)	4.87(2.24)	5.78(5.97)	2.43(2.24)	7.85(0.00)	1.05(1.49)
		fAIC	7.96(0.00)	78.7(14.5)	8.00(0.00)	103(18.66)	8.00(0.00)	13.2(20.9)
		fBIC	7.92(0.00)	43.5(54.5)	7.99(0.00)	0.13(0.00)	8.00(0.00)	0.01(0.00)
		fEBIC	0.98(0.00)	0.02(0.00)	1.08(0.00)	0.00(0.00)	3.51(4.48)	0.00(0.00)
$[3, 1]$	7.66	gNIS	4.68(4.48)	2.97(2.24)	6.19(0.75)	1.99(1.49)	7.73(0.00)	1.06(0.75)
		fAIC	7.38(0.75)	80.3(15.7)	7.98(0.00)	103(17.91)	8.00(0.00)	130(23.13)
		fBIC	6.53(2.24)	32.2(52.2)	7.60(0.75)	0.08(0.00)	7.94(0.00)	0.01(0.00)
		fEBIC	0.98(0.00)	0.02(0.00)	1.01(0.00)	0.00(0.00)	1.44(0.00)	0.00(0.00)
$[4, 5]$	9.20	gNIS	2.43(2.24)	0.16(0.00)	4.14(2.98)	0.10(0.00)	6.64(1.49)	0.10(0.00)
		fAIC	6.11(1.49)	71.4(14.9)	7.73(0.00)	90.2(19.4)	7.97(0.00)	112(20.89)
		fBIC	4.34(2.24)	16.5(2.24)	5.94(1.49)	0.11(0.00)	7.27(0.75)	0.01(0.00)
		fEBIC	0.93(0.00)	0.07(0.00)	0.99(0.00)	0.01(0.00)	1.04(0.00)	0.00(0.00)
$[6, 8]$	10.26	gNIS	0.79(0.75)	0.05(0.00)	1.25(1.49)	0.01(0.00)	2.38(1.49)	0.00(0.00)
		fAIC	4.17(1.49)	72.1(14.9)	6.20(1.49)	88.1(18.7)	7.32(0.75)	109(22.39)
		fBIC	2.00(1.49)	3.97(0.75)	2.88(0.00)	0.19(0.00)	3.86(0.75)	0.05(0.00)
		fEBIC	0.80(0.00)	0.20(0.00)	0.96(0.00)	0.04(0.00)	0.99(0.00)	0.01(0.00)

more computational time than the fBIC and fEBIC do. For example, for the setting "$[t_1, t_2] = [0, 0]$ and $n = 200$" in Table 2, on average the gNIS, fAIC, fBIC, and fEBIC approaches took $1041, 126, 19$ and 10 minutes to finish the 1000 runs, respectively. This is because the gNIS iterates many times before it stops while the fAIC often selects too many false covariates (often in hundreds). While the time used by the fBIC and fEBIC are comparable, the fEBIC uses less computational time than the fBIC does since it selects much fewer true and false covariates.

4.2 Illustrative examples using a breast cancer dataset

In the United States, breast cancer is one of the leading causes of deaths from cancer among women. To help the breast cancer diagnosis, it is important to predict the metastatic behavior of breast cancer tumor jointly using clinical risk factors and gene expression profiles. In the breast cancer study reported by [21], expression levels for 24481 gene probes and clinical risk factors (age, tumor size, histological grade, angioinvasion, lymphocytic infiltration, estrogen receptor, and progesterone receptor status) were collected for 97 lymph node-negative breast cancer patients 55 years old or younger. Among them, 46 developed distant metastases within 5 years and 51 remained metastases free for more than 5 years. Recently, [35] proposed a ROC based approach to rank the genes via adjusting for the clinical risk factors. They removed genes with severe missingness, leading to an effective number of 24188 genes. The gene expression data are normalized such that they all have sample mean 0 and standard deviation 1. In this section, we illustrate our forward selection procedures using two examples based on this breast cancer dataset. However, they are not meant to serve as formal analyses of the data.

In the first illustrative example, we are interested in selecting some useful genes whose expressions can be used to predict the tumor size (TS). To set up a varying coefficient regression model, we use the estrogen receptor (ER) as the index variable, aiming to see if it has strong impact on the effects of gene expression profiles. That is to say, we are interested if the tumor size is determined by some genes with effects adjusted by the value of estrogen receptor. The resulting varying coefficient regression model can be expressed as

$$\mathrm{TS}_i = \beta_0(\mathrm{ER}_i) + \textstyle\sum_{j=1}^{24188} \beta_j(\mathrm{ER}_i)\mathrm{gene}_{ij}, i = 1, 2, \cdots, 97.$$

It is expected that not all of the 24188 genes can have impact on the tumor size. Applying the fBIC approach with the cubic B-spline with 2 knots (selected by BIC) to the data, we

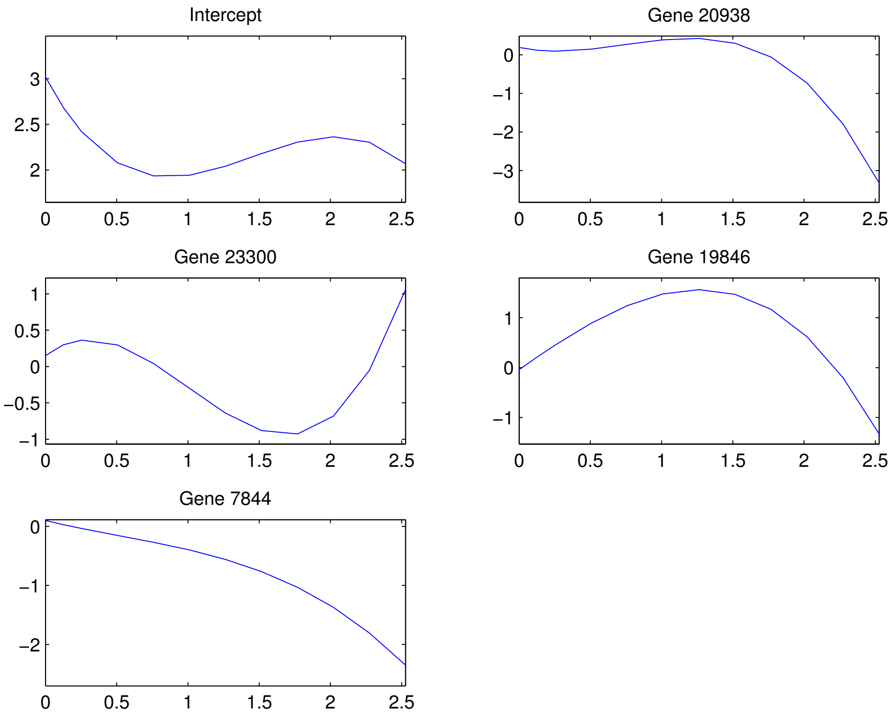

Figure 1: *First illustrative example based on the breast cancer data: fitted coefficient functions.*

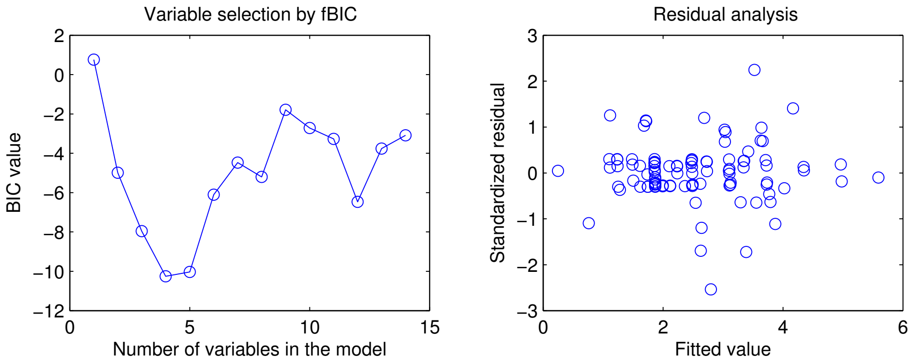

Figure 2: *First illustrative example based on the breast cancer data: variable selection by fBIC and residual analysis.*

identified 4 genes, 20938, 23300, 19846, and 7844, with strong impact on the tumor size. The estimated coefficient functions are presented in Figure 1. Applying the generalized likelihood ratio test of [13] to compare this selected varying coefficient model against the linear regression model obtained via replacing all the coefficient functions with constants shows that the varying coefficient model is indeed statistically significant with p-value 0. The left panel of Figure 2 presents the BIC curve for variable selection by the fBIC approach and the right panel shows the standardized residuals from the fitted varying coefficient model are generally scattered in a proper range. It seems that the varying coefficient model generally fits the data well. We would also like to mention that the fEBIC selected gene 20938 only, the fAIC selected 9 more genes than those obtained by fBIC, while the gNIS approach of [10] selected 14 genes which are different from what we have obtained.

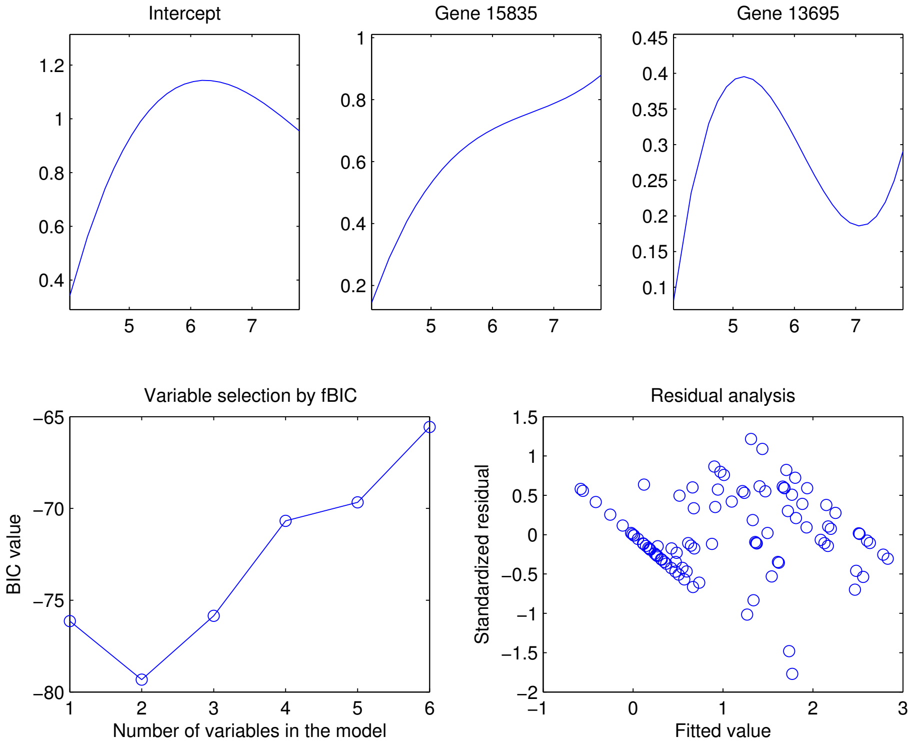

Figure 3: *Second illustrative example based on the breast cancer data.*

In the second illustrative example, we are interested in finding some useful genes whose expressions can be used to predict the values of estrogen receptor (ER) and this time we used the clinical risk factor age as the index variable since the effects of the gene profiles on the estrogen receptor may change over age. The resulting varying coefficient regression model can now be expressed as

$$\mathrm{ER}_i = \beta_0(\mathrm{age}_i) + \textstyle\sum_{j=1}^{24188} \beta_j(\mathrm{age}_i)\mathrm{gene}_{ij}, i = 1, 2, \cdots, 97.$$

Similarly, by applying the fBIC approach with the cubic B-spline with 2 knots (selected by BIC), we found 2 genes, 15835 and 13695, have strong impact on the estrogen receptor. The estimated coefficient functions are presented in the upper panels of Figure 3. The BIC curve for variable selection and the residual analysis are presented in the lower panels. It is seen that the varying coefficient model fits the data reasonably well. Applying the generalized likelihood ratio test of [13] to compare this model against the linear model obtained via replacing the three coefficient functions with constants shows that the varying coefficient model is statistically significant with p-value 0.011. In this example, the fEBIC selected gene 15835 only, the fAIC selected 11 more genes than those obtained by fBIC, while the gNIS approach of [10] did not select any genes.

5 Proofs

In this section, we prove Theorem 3.1, Corollaries 3.1 and 3.2, and Theorem 3.2. We postpone the proof of Lemma 3.1 to the end of this section.

Proof of Theorem 3.1. First we define some notation. Let

$$\mathbf{Q}_S = \mathbf{I}_n - \mathbf{H}_S, \quad \mathbf{H}_S = \boldsymbol{W}_S(\boldsymbol{W}_S^T\boldsymbol{W}_S)^{-1}\boldsymbol{W}_S^T, \quad \text{and } \widetilde{\mathbf{H}}_{lS} = \widetilde{\boldsymbol{W}}_{lS}(\widetilde{\boldsymbol{W}}_{lS}^T\widetilde{\boldsymbol{W}}_{lS})^{-1}\widetilde{\boldsymbol{W}}_{lS}^T.$$

They are orthogonal projection matrices. Notice that

$$\widetilde{\mathbf{H}}_{lS}\mathbf{Q}_S = \widetilde{\mathbf{H}}_{lS}, \quad n\widehat{\sigma}_S^2 - n\widehat{\sigma}_{S(l)}^2 = |\widetilde{\mathbf{H}}_{lS}\boldsymbol{Y}|^2,$$

and

$$\max_{l\in S^c}\{n\widehat{\sigma}_S^2 - n\widehat{\sigma}_{S(l)}^2\} \geq \max_{l\in\mathcal{S}_0-S}|\widetilde{\mathbf{H}}_{lS}\boldsymbol{Y}|^2. \tag{16}$$

We evaluate the right-hand side of (16). Writing $\boldsymbol{\gamma}^* = (\gamma_j^*)_{j\in\mathcal{S}_0}$, we have

$$\begin{aligned}|\widetilde{\mathbf{H}}_{lS}\boldsymbol{Y}| &\geq |\widetilde{\mathbf{H}}_{lS}\boldsymbol{W}_{\mathcal{S}_0}\boldsymbol{\gamma}^*| - |\widetilde{\mathbf{H}}_{lS}\boldsymbol{r}| - |\widetilde{\mathbf{H}}_{lS}\boldsymbol{\epsilon}| \qquad (17)\\ &= A_{1l} - A_{2l} - A_{3l} \quad \text{(say)},\end{aligned}$$

where $\boldsymbol{r} = (r_1, \ldots, r_n)^T$ and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$. Since $\widetilde{\mathbf{H}}_{lS}$ is an orthogonal projection matrix whose rank is L, we have

$$|A_{2l}| \leq |\boldsymbol{r}| \leq (nC_rL^{-4})^{1/2} \tag{18}$$

uniformly in l and S. On the other hand, we have from Assumptions D and E that uniformly in $l \in \mathcal{S}_0 - S$ and S,

$$A_{3l}^2 = \boldsymbol{\epsilon}^T \widetilde{\mathbf{H}}_{lS} \boldsymbol{\epsilon} = O_p(L(\log n)^{d_m}). \tag{19}$$

Here we applied Proposition 3 of [37] with $x = C(\log n)^{d_m}$ conditionally on all the covariates. Then we have

$$\mathrm{P}\Big(\frac{\boldsymbol{\epsilon}^T \widetilde{\mathbf{H}}_{lS} \boldsymbol{\epsilon}}{LC_\epsilon} \geq \frac{1+x}{\{1 - 2/(e^{x/2}\sqrt{1+x} - 1)\}_+^2}\Big) \leq \exp\{-Lx/2\}(1+x)^{L/2},$$

where $\{a\}_+ = \max\{0, a\}$. We should take a sufficiently large C.

We evaluate A_{1l} as in the proof of Theorem 1 of [30]. With probability tending to 1, we have that

$$\begin{aligned} A_{1l}^2 &= |\widetilde{\mathbf{H}}_{lS} \mathbf{Q}_S \boldsymbol{W}_{\mathcal{S}_0} \boldsymbol{\gamma}^*|^2 = \boldsymbol{\gamma}^{*T} \boldsymbol{W}_{\mathcal{S}_0}^T \mathbf{Q}_S \widetilde{\boldsymbol{W}}_{lS} (\widetilde{\boldsymbol{W}}_{lS}^T \widetilde{\boldsymbol{W}}_{lS})^{-1} \widetilde{\boldsymbol{W}}_{lS}^T \mathbf{Q}_S \boldsymbol{W}_{\mathcal{S}_0} \boldsymbol{\gamma}^* \\ &\geq \lambda_{\min}((\widetilde{\boldsymbol{W}}_{lS}^T \widetilde{\boldsymbol{W}}_{lS})^{-1}) |\widetilde{\boldsymbol{W}}_{lS}^T \mathbf{Q}_S \boldsymbol{W}_{\mathcal{S}_0} \boldsymbol{\gamma}^*|^2 \\ &\geq \frac{L}{n\tau_{\max}(M)(1+\delta_{1n})} |\widetilde{\boldsymbol{W}}_{lS}^T \mathbf{Q}_S \boldsymbol{W}_{\mathcal{S}_0} \boldsymbol{\gamma}^*|^2 \end{aligned} \tag{20}$$

uniformly in $l \in \mathcal{S}_0 - S$ and S. We will give a lower bound of

$$\max_{l \in \mathcal{S}_0 - S} |\widetilde{\boldsymbol{W}}_{lS}^T \mathbf{Q}_S \boldsymbol{W}_{\mathcal{S}_0} \boldsymbol{\gamma}^*|^2 = \max_{l \in \mathcal{S}_0 - S} |\boldsymbol{W}_l^T \mathbf{Q}_S \boldsymbol{W}_{\mathcal{S}_0} \boldsymbol{\gamma}^*|^2.$$

We have

$$\begin{aligned} |\mathbf{Q}_S \boldsymbol{W}_{\mathcal{S}_0} \boldsymbol{\gamma}^*|^2 = \boldsymbol{\gamma}^{*T} \boldsymbol{W}_{\mathcal{S}_0}^T \mathbf{Q}_S \boldsymbol{W}_{\mathcal{S}_0} \boldsymbol{\gamma}^* &= \sum_{l \in \mathcal{S}_0} \gamma_l^{*T} \boldsymbol{W}_l^T \mathbf{Q}_S \boldsymbol{W}_{\mathcal{S}_0} \boldsymbol{\gamma}^* \\ &\leq \Big(\sum_{l \in \mathcal{S}_0} |\gamma_l^*|^2\Big)^{1/2} \Big(\sum_{l \in \mathcal{S}_0} |\boldsymbol{W}_l^T \mathbf{Q}_S \boldsymbol{W}_{\mathcal{S}_0} \boldsymbol{\gamma}^*|^2\Big)^{1/2} \\ &\leq (C_{B2} L \|\boldsymbol{\beta}_0\|^2)^{1/2} p_0^{1/2} \max_{l \in \mathcal{S}_0 - S} |\boldsymbol{W}_l^T \mathbf{Q}_S \boldsymbol{W}_{\mathcal{S}_0} \boldsymbol{\gamma}^*|. \end{aligned}$$

Here we used (14). Therefore we have

$$\max_{l \in \mathcal{S}_0 - S} |\boldsymbol{W}_l^T \mathbf{Q}_S \boldsymbol{W}_{\mathcal{S}_0} \boldsymbol{\gamma}^*|^2 \geq (C_{B2} L \|\boldsymbol{\beta}_0\|_{L_2}^2)^{-1} p_0^{-1} |\mathbf{Q}_S \boldsymbol{W}_{\mathcal{S}_0} \boldsymbol{\gamma}^*|^4. \tag{21}$$

Note that we derived (21) by just using elementary linear algebra and calculus. On the other hand, we have with probability tending to 1 that

$$\begin{aligned} |\mathbf{Q}_S \boldsymbol{W}_{\mathcal{S}_0} \boldsymbol{\gamma}^*|^2 &= \boldsymbol{\gamma}^{*T} \boldsymbol{W}_{\mathcal{S}_0}^T \mathbf{Q}_S \boldsymbol{W}_{\mathcal{S}_0} \boldsymbol{\gamma}^* = (\gamma_j^{*T})_{j \in \mathcal{S}_0 - S}^T \boldsymbol{W}_{\mathcal{S}_0 - S}^T \mathbf{Q}_S \boldsymbol{W}_{\mathcal{S}_0 - S} (\gamma_j^{*T})_{j \in \mathcal{S}_0 - S}^T \\ &\geq C_{B1} n \sum_{j \in \mathcal{S}_0 - S} \|\beta_{0j}\|_2^2 \tau_{\min}(M)(1 - \delta_{1n}) \end{aligned} \tag{22}$$

uniformly in S. Combining (20)-(22), we have with probability tending to 1 that

$$\max_{l\in\mathcal{S}_0-S} A_{1l}^2 \geq \frac{C_{B1}^2(1-\delta_{1n})^2}{p_0 C_{B2}(1+\delta_{1n})}\frac{n(\tau_{\min}(M))^2}{\tau_{\max}(M)}\frac{\min_{j\in\mathcal{S}_0}\|\beta_{0j}\|_{L_2}^4}{\|\boldsymbol{\beta}_0\|_{L_2}^2} \tag{23}$$

uniformly in S.

It follows from (17)-(19), (23), and Assumption D that there exists a positive sequence $\delta_{2n}\to 0$ such that

$$\max_{l\in S^c}\{n\widehat{\sigma}_S^2 - n\widehat{\sigma}_{S(l)}^2\} \geq (1-\delta_{2n})\frac{C_{B1}^2}{p_0 C_{B2}}\frac{n(\tau_{\min}(M))^2}{\tau_{\max}(M)}\frac{\min_{j\in\mathcal{S}_0}\|\beta_{0j}\|_{L_2}^4}{\|\boldsymbol{\beta}_0\|_{L_2}^2}$$

uniformly in S with probability tending to 1. Hence the proof of Theorem 3.1 is complete.

Proof of Corollary 3.1. We have

$$\mathrm{EBIC}(S) - \mathrm{EBIC}(S(l)) = n\log\Big(\frac{n\widehat{\sigma}_S^2}{n\widehat{\sigma}_{S(l)}^2}\Big) - L(\log n + 2\eta\log p)$$

and

$$\frac{n\widehat{\sigma}_S^2}{n\widehat{\sigma}_{S(l)}^2} = 1 + \frac{n\widehat{\sigma}_S^2 - n\widehat{\sigma}_{S(l)}^2}{n\widehat{\sigma}_{S(l)}^2} \geq 1 + \frac{n\widehat{\sigma}_S^2 - n\widehat{\sigma}_{S(l)}^2}{\sum_{i=1}^n (Y_i - \overline{Y})^2},$$

where $\overline{Y} = \sum_{i=1}^n Y_i$. Thus

$$\mathrm{EBIC}(S) - \mathrm{EBIC}(S(l)) \geq \frac{n\widehat{\sigma}_S^2 - n\widehat{\sigma}_{S(l)}^2}{n^{-1}\log 2\sum_{i=1}^n (Y_i-\overline{Y})^2} - L(\log n + 2\eta\log p).$$

Theorem 3.1 and the assumption of this corollary imply that we don't stop when

$$\mathcal{S}_0 \not\subset S \qquad \text{and} \qquad \#(S\cup\mathcal{S}_0) \leq M$$

with probability tending to 1. If $\mathcal{S}_0\not\subset S_1,\dots,\mathcal{S}_0\not\subset S_{k-1}$, and $\#(S_{k-1}\cup\mathcal{S}_0)\leq M$, we have

$$\frac{1}{n}\sum_{i=1}^n (Y_i-\overline{Y})^2 \geq \widehat{\sigma}_{S_1}^2 - \widehat{\sigma}_{S_k}^2$$

with probability tending to 1. Recall that $S_1 = \{0\}$. Therefore we have with probability tending to 1,

$$\mathrm{Var}(Y) > (k-1)D_M(1-\delta_{2n}) \tag{24}$$

uniformly in $\{S_j\}_{j=1}^{k-1}$ such that $\#(S_{k-1}\cup\mathcal{S}_0)\leq M$ and $\mathcal{S}_0\not\subset S_{k-1}$. Suppose that we have $\mathcal{S}_0\not\subset S_{T_M}$ ($k-1=T_M$). Then (24) implies that with probability tending to 1,

$$\frac{\mathrm{Var}(Y)}{D_M(1-\delta_{2n})} > T_M,$$

which contradicts the definition of T_M. Hence we have some $k-1$ such that $k-1 \leq T_M$ and $\mathcal{S}_0 \subset S_{k-1}$. Hence the proof of the corollary is complete.

Proof of Corollary 3.2. We have

$$\mathrm{EBIC}(S_k(l)) - \mathrm{EBIC}(S_k) = n \log \Big(1 - \frac{n\widehat{\sigma}^2_{S_k} - n\widehat{\sigma}^2_{S_k(l)}}{n\widehat{\sigma}^2_{S_k}} \Big) + L(\log n + 2\eta \log p). \tag{25}$$

As in the proof of Theorem 3.1, we can apply Proposition of [37] and obtain

$$\widehat{\sigma}^2_{S_k} = \mathrm{E}\{\epsilon^2\} + o_p(1) \tag{26}$$

uniformly in S_k. As in the proof of Theorem 3.1 again, we have uniformly in S_k,

$$\max_{l \in S_k^c} \{ n\widehat{\sigma}^2_{S_k} - n\widehat{\sigma}^2_{S_k(l)} \} = O_p(nL^{-4}) + O_p(L(\log n)^{d_m}). \tag{27}$$

Note that we used the fact that $\mathcal{S}_0 \subset S_k$. From (25)-(27), we have

$$\mathrm{EBIC}(S_k(l)) - \mathrm{EBIC}(S_k) = L(\log n + 2\eta \log p)(1 + o_p(1))$$

uniformly in S_k and $l \in S_k^c$. Hence the proof of the corollary is complete.

Proof of Theorem 3.2. Recalling (16) and (17) in the proof of Theorem 3.1, we have

$$\max_{l \in S^c \cap \mathcal{V}^c} \{ n\widehat{\sigma}^2_S - n\widehat{\sigma}^2_{S(l)} \} = \max_{l \in S^c \cap \mathcal{V}^c} |\widetilde{\mathbf{H}}_{lS} \boldsymbol{Y}|^2 \ \text{ and } \ |\widetilde{\mathbf{H}}_{lS} \boldsymbol{Y}| \leq A_{1l} + A_{2l} + A_{3l}.$$

In this proof, we have only to evaluate A_{1l} since we can handle A_{2l} and A_{3l} as in (18) and (19). First, we have

$$\begin{aligned} A_{1l} &= |\widetilde{\mathbf{H}}_{lS} \mathbf{Q}_S \boldsymbol{W}_{\mathcal{S}_0} \boldsymbol{\gamma}^*|^2 = \boldsymbol{\gamma}^{*T} \boldsymbol{W}_{\mathcal{S}_0}^T \mathbf{Q}_S \widetilde{\mathbf{H}}_{lS} \mathbf{Q}_S \boldsymbol{W}_{\mathcal{S}_0} \boldsymbol{\gamma}^* \\ &\leq \{ \lambda_{\min}(\widetilde{\boldsymbol{W}}_{lS}^T \widetilde{\boldsymbol{W}}_{lS}) \}^{-1} |\boldsymbol{W}_l^T \mathbf{Q}_S \boldsymbol{W}_{\mathcal{S}_0} \boldsymbol{\gamma}^*|^2 . \end{aligned} \tag{28}$$

The latter half of Assumption Z implies that

$$\{ \lambda_{\min}(\widetilde{\boldsymbol{W}}_{lS}^T \widetilde{\boldsymbol{W}}_{lS}) \}^{-1} \leq \frac{L}{nC_m}(1 + o_p(1)) , \tag{29}$$

uniformly in S and $l \in \mathcal{V}^c$. Next we evaluate $|\boldsymbol{W}_l^T \mathbf{Q}_S \boldsymbol{W}_{\mathcal{S}_0} \boldsymbol{\gamma}^*|^2$. Recalling $\mathbf{Q}_S = I - \mathbf{H}_S$, we have

$$|\boldsymbol{W}_l^T \mathbf{Q}_S \boldsymbol{W}_{\mathcal{S}_0} \boldsymbol{\gamma}^*|^2 \leq 2|\boldsymbol{W}_l^T \boldsymbol{W}_{\mathcal{S}_0} \boldsymbol{\gamma}^*|^2 + 2|\boldsymbol{W}_l^T \mathbf{H}_S \boldsymbol{W}_{\mathcal{S}_0} \boldsymbol{\gamma}^*|^2. \tag{30}$$

As for the first term of the right-hand side of (30), it follows from the former half of Assumption Z that

$$|\boldsymbol{W}_l^T \boldsymbol{W}_{\mathcal{S}_0} \boldsymbol{\gamma}^*|^2 = L|\boldsymbol{\gamma}^*|^2 p_0 O_p\Big(\Big(\frac{n\kappa_n}{L}\Big)^2\Big) = O_p(\kappa_n^2 n^2 p_0) \tag{31}$$

uniformly in S and $l \in \mathcal{V}^c$. We exploited the local properties of B-spline bases. Finally we consider the second term in the right-hand side of (30). Notice that

$$|\boldsymbol{W}_l^T \mathbf{H}_S \boldsymbol{W}_{\mathcal{S}_0} \boldsymbol{\gamma}^*|^2 \leq \mathrm{tr}(\boldsymbol{W}_l^T \mathbf{H}_S \boldsymbol{W}_l) |\boldsymbol{\gamma}^{*T} \boldsymbol{W}_{\mathcal{S}_0}^T \mathbf{H}_S \boldsymbol{W}_{\mathcal{S}_0} \boldsymbol{\gamma}^*|$$

We have from Assumption Z that

$$\begin{aligned}\mathrm{tr}(\boldsymbol{W}_l^T \mathbf{H}_S \boldsymbol{W}_l) &= \mathrm{tr}\{\boldsymbol{W}_l^T \boldsymbol{W}_S (\boldsymbol{W}_S^T \boldsymbol{W}_S)^{-1} \boldsymbol{W}_S^T \boldsymbol{W}_l\} \\ &= O_p\Big(\frac{n\kappa_n^2 M}{\tau_{\min}^{\mathcal{V}}(M)(1-\delta_{1n})}\Big),\end{aligned}$$

uniformly in S and $l \in \mathcal{V}^c$. We exploited the local properties of B-spline bases again. We also have

$$|\boldsymbol{\gamma}^{*T} \boldsymbol{W}_{\mathcal{S}_0}^T \mathbf{H}_S \boldsymbol{W}_{\mathcal{S}_0} \boldsymbol{\gamma}^*| = O_p\Big(L\frac{n\tau_{\max}^{\mathcal{S}_0}}{L}\Big) = O_p(n\tau_{\max}^{\mathcal{S}_0}).$$

uniformly in S. Combining the above results, we obtain

$$|\boldsymbol{W}_l^T \mathbf{H}_S \boldsymbol{W}_{\mathcal{S}_0} \boldsymbol{\gamma}^*|^2 = O_p\Big(\frac{n\kappa_n^2 M}{\tau_{\min}^{\mathcal{V}}(M)} n\tau_{\max}^{\mathcal{S}_0}\Big), \tag{32}$$

uniformly in S and $l \in \mathcal{V}^c$. The desired results follow from (18)-(19) and (28)-(32).

Proof of Lemma 3.1. Let a_n increase to infinity. Then we have

$$\frac{1}{n}\boldsymbol{W}_S^T \boldsymbol{W}_S - \mathrm{E}\Big\{\frac{1}{n}\boldsymbol{W}_S^T \boldsymbol{W}_S\Big\} = O_p\Big(a_n\sqrt{\frac{\log n}{nL}}\Big) \quad \text{componentwise}, \tag{33}$$

uniformly in S, if

$$\max\{\log p, \log n\} = O(a_n^2 \log n). \tag{34}$$

This is just an application of the Bernstein inequality and we have only to consider every pair of covariates. Equation (33) implies that

$$|\lambda_{\min}(n^{-1}\boldsymbol{W}_S^T \boldsymbol{W}_S) - \lambda_{\min}(\mathrm{E}\{n^{-1}\boldsymbol{W}_S^T \boldsymbol{W}_S\})| = O_p\Big(Ma_n\sqrt{\frac{\log n}{nL}}\Big),$$

uniformly in S. Here we used the local properties of B-spline bases. If

$$Ma_n\sqrt{\frac{\log n}{nL}} = o\Big(\frac{\tau_{\min}(M)}{L}\Big) \tag{35}$$

we obtain the desired result. Equation (35) is equivalent to

$$a_n^2 = o\Big(\frac{n^{4/5}}{\log n}\Big(\frac{\tau_{\min}(M)}{M}\Big)^2\Big). \tag{36}$$

The assumption of the lemma and (36) imply the existence of $\{a_n\}$ satisfying (34) and (35). We can handle the maximum eigenvalues in the same way. Hence the proof of the lemma is complete.

Acknowledgements. The authors thank the associate editor and referees for their constructive comments, Dr Jialiang Li for providing the genetic dataset, and Dr Wei Dai for sharing the codes for implementation of the methods of [10].

References

[1] A. Antoniadis, I. Gijbels, and A. Verhasselt. Variable selection in varying-coefficient models using p-splines. *J. Comput. Graph. Stat.*, 21:638–661, 2012.

[2] P. J. Bickel, Y. A. Ritov, and A. B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *Ann. Statist.*, 37:1705–1732, 2009.

[3] E. Candes and T. Tao. The dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.*, 35:2313–2351, 2007.

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

[5] M.-Y. Cheng, T. Honda, J. Li, and H. Peng. Nonparametric independence screening and structure identification for ultra-high dimensional longitudinal data. *Ann. Statist.*, 42:1819–1849, 2014.

[6] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression (with discussions). *Ann. Statist.*, 32:407–499, 2004.

[7] J. Fan, Y. Feng, and R. Song. Nonparametric independence screening in sparse ultra-high-dimensional additive models. *J. Amer. Statist. Assoc.*, 106:544–557, 2011.

[8] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96:1348 – 1360, 2001.

[9] J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space. *J. Royal Statist. Soc. Ser. B*, 70:849–911, 2008.

[10] J. Fan, Y. Ma, and W. Dai. Nonparametric independence screening in sparse ultra-high dimensional varying coefficient models. *J. Amer. Statist. Assoc.*, 109:1270–1284, 2014.

[11] J. Fan and R. Song. Sure independence screening in generalized linear models with NP-dimensionality. *Ann. Statist.*, 38:3567–3604, 2010.

[12] J. Fan, L. Xue, and H. Zou. Strong oracle optimality of folded concave penalized estimation. *Ann. Statist.*, 42:819–849, 2014.

[13] J. Fan, C. Zhang, and J. Zhang. Generalized likelihood ratio statistics and wilks phenomenon. *Ann. Statist.*, 29:153–193, 2001.

[14] J. Fan and W. Zhang. Statistical methods with varying coefficient models. *Statistics and its Interface*, 1:179–195, 2008.

[15] N. Hao and H. H. Zhang. Interaction screening for ultra-high dimensional data. *J. Amer. Statist. Assoc.*, 109:1285–1301, 2014.

[16] W. Y. Hwang, H. H. Zhang, and S. Ghosal. First: Combining forward iterative selection and shrinkage in high dimensional sparse linear regression. *Statistics and its Interface*, 2:341–348, 2009.

[17] E. R. Lee, H. Noh, and B. U. Park. Model selection via bayesian information criterion for quantile regression models. *J. Amer. Statist. Assoc.*, 109:216–229, 2014.

[18] H. Lian. Variable selection for high-dimensional generalized varying-coefficient models. *Statist. Sinica*, 22:1563–1588, 2012.

[19] H. Lian. Semiparametric bayesian information criterion for model selection in ultra-high dimensional additive models. *J. Multivar. Statist. Anal.*, 123:304–310, 2014.

[20] J. Liu, R. Li, and R. Wu. Feature selection for varying coefficient models with ultrahigh dimensional covariates. *J. Amer. Statist. Assoc.*, 109:266–274, 2014.

[21] LJ Vant Veer LJ, H. Dai, and et al. MJ van de Vijver. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530–536, 2002.

[22] S. Luo and Z. Chen. Sequential lasso cum ebic for feature selection with ultra-high dimensional feature space. *J. Amer. Statist. Assoc.*, 109:1229–1240, 2014.

[23] L. Meier, S. van de Geer, and P. Bühlmann. The group lasso for logistic regression. *J. Royal Statist. Soc. Ser. B*, 70:53–71, 2008.

[24] H. S. Noh and B. U. Park. Sparse variable coefficient models for longitudinal data. *Statist. Sinica*, 20:1183–1202, 2010.

[25] P. Radchenko and G. M. James. Improved variable selection with forward-lasso adaptive shrinkage. *Ann. Appl. Statis.*, 5:427–448, 2011.

[26] L. L. Schumaker. *Spline Functions: Basic Theory 3rd ed.* Cambridge University Press, Cambridge, 2007.

[27] R. Song, F. Yi, and H. Zou. On varying-coefficient independence screening for high-dimensional varying-coefficient models. *Statistica Sinica*, 24:1735–1752, 2014.

[28] Y. Tang, H. J. Wang, Z. Zhu, and X. Song. A unified variable selection approach for varying coefficient models. *Statist. Sinica*, 22:601–628, 2012.

[29] R. J. Tibshirani. Regression shrinkage and selection via the Lasso. *J. Royal Statist. Soc. Ser. B*, 58:267–288, 1996.

[30] H. Wang. Forward regression for ultra-high dimensional variable screening. *JASA*, 104:1512–1524, 2009.

[31] L. Wang, H. Li, and J. Z. Huang. Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *J. Amer. Statist. Assoc.*, 103:1556–1569, 2008.

[32] F. Wei, J. Huang, and H. Li. Variable selection and estimation in high-dimensional varying-coefficient models. *Statist. Sinica*, 21:1515–1540, 2011.

[33] Y. Xia, W. Zhang, and H. Tong. Efficient estimation for semivarying-coefficient models. *Biometrika*, 91:661–681, 2004.

[34] L. Xue and A. Qu. Variable selection in high-dimensional varying-coefficient models with global optimality. *J. Machine Learning Research*, 13:1973–1998, 2012.

[35] T. Yu, J. Li, and S. Ma. Adjusting confounders in ranking biomarkers: a model-based roc approach. *Brief Bioinform,*, 13:513–523, 2012.

[36] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *J. Royal Statist. Soc. Ser. B*, 68:49–67, 2006.

[37] C. H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, 38:894–942, 2010.

[38] W. Zhang, S. Lee, and X. Song. Local polynomial fitting in semivarying coefficient model. *J. Multivar. Statist. Anal.*, 82:166–168, 2002.

[39] H. Zou. The adaptive Lasso and its oracle properties. *J. Amer. Statist. Assoc.*, 101:1418–1429, 2006.

Supplement to “Forward variable selection for sparse ultra-high dimensional varying coefficient models”

by Ming-Yen Cheng, Toshio Honda, and Jin-Ting Zhang

S.1 Selection Consistency

In this section we state selection consistency of the forward selection procedure. The proofs are given in Section S.2. Here, we deal with the ultra-high dimensional case where

$$\log p = O(n^{1-c_p}/L), \tag{S.1}$$

where c_p is a positive constant. In addition, recalling $\mathcal{S}_0$ is the index set of the true variables and p_0 is the number of true covariates, i.e. $p_0 \equiv \#\mathcal{S}_0$, we consider the case that

$$p_0 = O((\log n)^{c_S}) \tag{S.2}$$

for some positive constant c_S. Condition (S.2) on p_0 is imposed for simplicity of presentation; it can be relaxed at the expense of restricting slightly the order of the dimension p specified in (S.1). Note that we treat the EBIC and the BIC ($\eta = 0$) in a unified way. The proofs are given in Section S.2.

First we state some assumptions. These assumptions, along with Assumption E given in Section 3, are needed for the results in this supplement. The following assumption on the index variable T in the varying coefficient model (1) is a standard one when we employ spline estimation.

Assumption T. The index variable T has density function $f_T(t)$ such that $C_{T1} < f_T(t) < C_{T2}$ uniformly in $t \in [0,1]$, for some positive constants C_{T1} and C_{T2}.

We define some more notation before we state our assumptions on the covariates. Recall that $\boldsymbol{X}_S$ consists of $\{X_j\}_{j\in S}$ and then $\boldsymbol{X}_S$ is a $\#S$-dimensional random vector. For a symmetric matrix $\mathbf{A}$, recall that $\lambda_{\max}(\mathbf{A})$ and $\lambda_{\min}(\mathbf{A})$ are the maximum and minimum eigenvalues respectively, and we define $|\mathbf{A}|$ as

$$|\mathbf{A}| = \sup_{|\mathbf{x}|=1} |\mathbf{A}\mathbf{x}| = \max\{|\lambda_{\max}(\mathbf{A})|, |\lambda_{\min}(\mathbf{A})|\}.$$

When g is a function of some random variable(s), we define the L_2 norm of g by $\|g\| = [\mathrm{E}\{g^2\}]^{1/2}$.

Assumption X.

X(1) There is a positive constant C_{X1} such that $|X_j| \leq C_{X1}$, $j = 1, \dots, p$.

X(2) Uniformly in $S \subsetneq \mathcal{S}_0$ and $l \in S^c$,

$$C_{X2} \leq \lambda_{\min}(\mathrm{E}\{\boldsymbol{X}_{S(l)}\boldsymbol{X}_{S(l)}^T|T\}) \leq \lambda_{\max}(\mathrm{E}\{\boldsymbol{X}_{S(l)}\boldsymbol{X}_{S(l)}^T|T\}) \leq C_{X3}$$

for some positive constants C_{X2} and C_{X3}.

We use assumption X(2) when we evaluate eigenvalues of the matrix $\mathrm{E}\{n^{-1}\boldsymbol{W}_{S(l)}^T\boldsymbol{W}_{S(l)}\}$. We can relax Assumption X(1) slightly by replacing C_{X1} with $C_{X1}(\log n)^{c_X}$ for some positive constant c_X. These are standard assumptions in the variable selection literature.

We need some assumptions on the coefficient functions of the marginal models given in (4) in order to establish selection consistency of the proposed forward procedure, especially Assumption C(1) given below. Here, C_1, C_2, ... are generic positive constants and their values may change from line to line. Recall that $\mathcal{S}_0$ is the index set of the true variables in model (1).

Assumption C

C(1) For some large positive constant C_{B1},

$$\max_{l \in \mathcal{S}_0 - S} \|\overline{\beta}_l\|_{L_2} / \max_{l \in \mathcal{S}_0^c} \|\overline{\beta}_l\|_{L_2} > C_{B1}$$

uniformly in $S \subsetneq \mathcal{S}_0$. Note that C_{B1} should depend on the other assumptions on the covariates, specifically Assumptions X and T given in Section 3.

C(2) Set $\bar{\kappa}_n = \inf_{S \subsetneq \mathcal{S}_0} \max_{l \in \mathcal{S}_0 - S} \|\overline{\beta}_l\|_{L_2}$. We assume

$$\frac{n\bar{\kappa}_n^2}{L \max\{\log p, \log n\}} > n^{c_\beta} \quad \text{and} \quad \bar{\kappa}_n > \frac{L}{n^{1-c_\beta}}$$

for some small positive constant c_β. In addition, if $\eta = 0$ in (10) i.e. if BIC is used, we require further that

$$\frac{L \log n}{\log p} \to \infty \,.$$

C(3) $\bar{\kappa}_n L^2 \to \infty$ and $\bar{\kappa}_n = O(1)$, where $\bar{\kappa}_n$ is defined in Assumption C(2).

C(4) $\overline{\beta}_j$ is twice continuously differentiable for every $j \in S(l)$ for any $S \subseteq \mathcal{S}_0$ and $l \in S^c$.

C(5) There are positive constants C_{B2} and C_{B3} such that $\sum_{j\in S(l)} \|\overline{\beta}_j\|_\infty < C_{B2}$ and $\sum_{j\in S(l)} \|\overline{\beta}_j''\|_\infty < C_{B3}$ uniformly in $S \subset \mathcal{S}_0$ and $l \in S^c$.

Assumptions C(1) and C(2) are the sparsity assumptions. An assumption similar to Assumption C(1) is imposed in Luo and Chen (2014) and such assumptions are inevitable in establishing selection consistency of forward procedures. These assumptions ensure that the chosen index l^*, given in (9), will be coming from $\mathcal{S}_0 - S$ instead of $\mathcal{S}_0^c$. The first condition in Assumption C(2) is related to the convergence rate of $\widehat{\gamma}_l$, and it ensures that the signals are large enough to be detected. If $C_1 < \bar{\kappa}_n < C_2$ for some positive constants C_1 and C_2, this condition is simply $\log p < n^{1-c_\beta}/L$ for some small positive constant c_β, which is fulfilled by assumption (S.1) on p. When these assumptions fail to hold, our method may choose some covariates from $\mathcal{S}_0^c$. However, we can use the proposed method as a screening approach and remove the selected irreverent variables using variable selection methods such as group SCAD (Cheng, et al., 2014). The last condition in Assumption C(2) is to ensure that, when the BIC is used as the stopping criterion, our method can deal with ultra-high dimensional cases. For example, if L is taken of the optimal order $n^{1/5}$ then p can be taken as $p = \exp(n^c)$ for any $0 < c \le 1/5$. Assumptions C(3)-C(5) on the coefficient functions $\overline{\beta}_j$ in the extended marginal model (4) are needed in order to approximate them by the B-spline basis. Note that, in Assumptions C(4)-C(5), $\overline{\beta}_j \equiv \beta_{0j}$ for all $j \in \mathcal{S}_0$ and $\overline{\beta}_j \equiv 0$ for all $j \in \mathcal{S}_0^c$ when $S = \mathcal{S}_0$.

Theorem S.1.1 given below suggests that the forward selection procedure using criterion (9) can pick up all the relevant covariates in the varying coefficient model (1) when C_{B1} in Assumption C(1) is large enough.

Theorem S.1.1 *Assume that Assumptions T, X, C, and E hold, and define l^* as in (9) for any $S \subsetneq \mathcal{S}_0$. Then, with probability tending to 1, there is a positive constant C_L such that*

$$\frac{\|\overline{\beta}_{l^*}\|_{L_2}}{\max_{l\in\mathcal{S}_0 - S}\|\overline{\beta}_l\|_{L_2}} > C_L$$

uniformly in $S \subsetneq \mathcal{S}_0$, and thus we have $l^* \in \mathcal{S}_0 - S$ for any $S \subsetneq \mathcal{S}_0$ when C_{B1} in Assumption C(1) is larger than $1/C_L$.

Theorem S.1.2 given next implies that the proposed forward procedure will not stop until all of the relevant variables indexed by $\mathcal{S}_0$ have been selected, and it does stop when all the true covariates in model (1) have been selected.

Theorem S.1.2 *Assume that Assumptions T, X, B, and C hold. Then we have the following results.*

(i) For l^* as in Theorem S.1.1, we have

$$EBIC(S(l^*)) < EBIC(S)$$

uniformly in $S \subsetneq \mathcal{S}_0$, with probability tending to 1.

(ii) We have

$$EBIC(\mathcal{S}_0(l)) > EBIC(\mathcal{S}_0)$$

uniformly in $l \in \mathcal{S}_0^c$, with probability tending to 1.

S.2 Proofs of Theorem S.1.1 and Theorem S.1.2

First, based on the B-spline basis, we can approximate the varying coefficient model (2) by the following approximate regression model:

$$Y_i = \sum_{j=0}^{p} \boldsymbol{\gamma}_{0j}^T \boldsymbol{W}_{ij} + \epsilon_i', \ i = 1, \ldots, n, \tag{S.3}$$

where $\boldsymbol{\gamma}_{0j} \in \mathbb{R}^L$ and $\boldsymbol{\gamma}_{0j}^T \mathbf{B}(t) \approx \beta_{0j}(t)$, $j = 0, 1, \ldots, p$. We define some notation related to the approximate regression models (S.3) and (5). Let

$$\begin{aligned} D_{lSn} &= n^{-1} \boldsymbol{W}_{S(l)}^T \boldsymbol{W}_{S(l)} \quad \text{and} \quad D_{lS} = \mathrm{E}\{D_{lS_n}\}, \\ d_{lS_n} &= n^{-1} \boldsymbol{W}_{S(l)}^T \boldsymbol{Y} \quad \text{and} \quad d_{lS} = \mathrm{E}\{d_{lS_n}\}, \quad \text{and} \\ \Delta_{lSn} &= D_{lSn}^{-1} d_{lSn} - D_{lS}^{-1} d_{lS}\,. \end{aligned}$$

Then, the parameter vector $\overline{\boldsymbol{\gamma}}_l$ in model (5) can be expressed as $\overline{\boldsymbol{\gamma}}_l = (\mathbf{0}_L, \ldots, \mathbf{0}_L, \mathbf{I}_L) D_{lS}^{-1} d_{lS}$, where $\mathbf{0}_L$ denotes the $L \times L$ zero matrix and $\mathbf{I}_L$ is the L-dimensional identity matrix.

Before we prove Theorems S.1.1 and S.1.2, we present Lemmas S.2.1-S.2.3 whose proofs are deferred to Section S.2.1. We verify these lemmas at the end of this section. In Lemma S.2.1 we evaluate the minimum and maximum eigenvalues of some matrices.

Lemma S.2.1 *Assume that Assumptions T, X, and E hold. Then, with probability tending to 1, there are positive constants M_{11}, M_{12}, M_{13}, and M_{14} such that*

$$L^{-1}M_{11} \leq \lambda_{\min}(D_{lSn}) \leq \lambda_{\max}(D_{lSn}) \leq L^{-1}M_{12}$$

and

$$L^{-1}M_{13} \leq \lambda_{\min}(n^{-1}\widetilde{\boldsymbol{W}}_{lS}^T\widetilde{\boldsymbol{W}}_{lS}) \leq \lambda_{\max}(n^{-1}\widetilde{\boldsymbol{W}}_{lS}^T\widetilde{\boldsymbol{W}}_{lS}) \leq L^{-1}M_{14}$$

uniformly in $S \subsetneq \mathcal{S}_0$ and $l \in S^c$.

Lemma S.2.2 is about the relationship between $\overline{\beta}_l$ and $\overline{\boldsymbol{\gamma}}_l$ in the extended marginal models (4) and (S.3).

Lemma S.2.2 *Assume that Assumptions T, X, and C(4)-(5) hold. Then there are positive constants M_{21} and M_{22} such that*

$$M_{21}\sqrt{L}\big(\|\overline{\beta}_l\|_{L_2} - O(L^{-2})\big) \leq |\overline{\boldsymbol{\gamma}}_l| \leq M_{22}\sqrt{L}\big(\|\overline{\beta}_l\|_{L_2} + O(L^{-2})\big)$$

uniformly in $S \subsetneq \mathcal{S}_0$ and $l \in S^c$.

We use Lemma S.2.3 to evaluate the estimation error for $\overline{\boldsymbol{\gamma}}_j$, $j \in S(l)$, in model (S.3).

Lemma S.2.3 *Assume that Assumptions T, X, and C(4)-(5) hold. Then, for any $\delta > 0$, there are positive constants M_{31}, M_{32}, M_{33}, and M_{34} such that*

$$|\Delta_{lSn}| \leq M_{31}L^{3/2}p_0^{3/2}\delta/n$$

uniformly in $S \subsetneq \mathcal{S}_0$ and $l \in S^c$, with probability

$$1 - M_{32}\, p_0^2\, L \exp\Big\{ - \frac{\delta^2}{M_{33}nL^{-1} + M_{34}\delta} + \log p + p_0 \log 2\Big\}.$$

Now we prove Theorems S.1.1 and S.1.2 by employing Lemmas S.2.1-S.2.3.

Proof of Theorem S.1.1. Note that $S \subsetneq \mathcal{S}_0$ and $l \in S^c$. We can write

$$\widehat{\boldsymbol{\gamma}}_l = \overline{\boldsymbol{\gamma}}_l + (\mathbf{0}_L, \ldots, \mathbf{0}_L, \mathbf{I}_L)\Delta_{lSn}. \tag{S.4}$$

Lemma S.2.1 implies we should deal with Δ_{lSn} on the right-hand side of (S.4) when we evaluate $\widehat{\sigma}_S^2 - \widehat{\sigma}_{S(l)}^2$ given in equation (8). For this purpose, Assumption C(2) suggests that we should take δ in Lemma S.2.3 as $\delta = n^{1-c_\beta/4}\bar{\kappa}_n/L$ tending to ∞. Recall the definition of $\bar{\kappa}_n$ in Assumption C(2). Then we have from Assumption C(2) that

$$\frac{\sqrt{L}\bar{\kappa}_n}{L^{3/2}p_0^{3/2}\delta/n} = \frac{n^{c_\beta/4}}{p_0^{3/2}} \to \infty \tag{S.5}$$

and

$$\begin{aligned} & p_0^2 L \exp\Big\{-\frac{1}{2M_{33}}\frac{\delta^2}{nL^{-1}} + \log p + p_0 \log 2\Big\} \qquad\qquad \text{(S.6)} \\ & = p_0^2 L \exp\Big\{-(2M_{33})^{-1} n^{1-c_\beta/2}\bar{\kappa}_n^2 L^{-1} + \log p + p_0 \log 2\Big\} \\ & < p_0^2 L \exp\Big\{-(2M_{33})^{-1} n^{c_\beta/2}\log p + \log p + p_0 \log 2\Big\} \to 0. \end{aligned}$$

It follows from (S.5), (S.6), and Lemma S.2.3 that, with probability tending to 1, $(\mathbf{0}_L, \dots, \mathbf{0}_L, \mathbf{I}_L)\Delta_{lSn}$ is negligible compared to $\overline{\boldsymbol{\gamma}}_l$ on the right-hand side of (S.4). Therefore Lemmas S.2.1 and S.2.2 and Assumption C(3) imply that we should focus on $\sqrt{L}\|\overline{\beta}_l\|_{L_2}$ in evaluating $\widehat{\sigma}^2_{S(l)}$ in (8). Hence the desired result follows from Assumption C(1). □

Proof of Theorem S.1.2. To prove result (i), we evaluate

$$\mathrm{EBIC}(S) - \mathrm{EBIC}(S(l)) = n \log\Big(\frac{n\widehat{\sigma}^2_S}{n\widehat{\sigma}^2_{S(l)}}\Big) - L(\log n + 2\eta \log p).$$

Since

$$n\widehat{\sigma}^2_S - n\widehat{\sigma}^2_{S(l)} = (\widetilde{\boldsymbol{W}}_{lS}^T\widetilde{\boldsymbol{Y}}_S)^T(\widetilde{\boldsymbol{W}}_{lS}^T\widetilde{\boldsymbol{W}}_{lS})^{-1}(\widetilde{\boldsymbol{W}}_{lS}^T\widetilde{\boldsymbol{Y}}_S) = \widehat{\boldsymbol{\gamma}}_l^T\widetilde{\boldsymbol{W}}_{lS}^T\widetilde{\boldsymbol{W}}_{lS}\widehat{\boldsymbol{\gamma}}_l,$$

we have

$$\frac{n\widehat{\sigma}^2_S}{n\widehat{\sigma}^2_{S(l)}} \geq 1 + (n^{-1}\boldsymbol{Y}^T\boldsymbol{Y})^{-1}\widehat{\boldsymbol{\gamma}}_l^T\Big(\frac{1}{n}\widetilde{\boldsymbol{W}}_{lS}^T\widetilde{\boldsymbol{W}}_{lS}\Big)\widehat{\boldsymbol{\gamma}}_l. \tag{S.7}$$

Then Lemma S.2.1 and (S.7) imply that we have for some positive C,

$$\mathrm{EBIC}(S) - \mathrm{EBIC}(S(l)) \geq CnL^{-1}|\widehat{\boldsymbol{\gamma}}_l|^2 - L(\log n + 2\eta \log p) \tag{S.8}$$

uniformly in $S \subsetneq \mathcal{S}_0$ and $l \in S^c$, with probability tending to 1. Here we use the fact that $L^{-1}|\widehat{\boldsymbol{\gamma}}_l|^2$ is uniformly bounded with probability tending to 1. Then as in the proof of Theorem S.1.1, we should consider $\sqrt{L}\|\overline{\beta}_l\|_{L_2}$ in evaluating the right-hand side of (S.8). Since Assumption C(2) implies that

$$\frac{nL^{-1}(\sqrt{L}\bar{\kappa}_n)^2}{L(\log n + 2\eta \log p)} = \frac{n\bar{\kappa}_n^2}{L(\log n + 2\eta \log p)} \to \infty,$$

we have from (S.8) that

$$\mathrm{EBIC}(S) - \mathrm{EBIC}(S(l)) > 0$$

uniformly in $S \subsetneq \mathcal{S}_0$ and $l \in S^c$ satisfying $\|\overline{\beta}_l\|_{L_2} / \max_{j \in \mathcal{S}_0 - S} \|\overline{\beta}_j\|_{L_2} > C_L$, with probability tending to 1. Hence the proof of result (i) is complete.

To prove result (ii), we evaluate

$$\begin{aligned} & \mathrm{EBIC}(\mathcal{S}_0(l)) - \mathrm{EBIC}(\mathcal{S}_0) \\ = \;& n \log \Big\{ 1 - \frac{\boldsymbol{Y}^T \widetilde{\boldsymbol{W}}_{l\mathcal{S}_0} (\widetilde{\boldsymbol{W}}_{l\mathcal{S}_0}^T \widetilde{\boldsymbol{W}}_{l\mathcal{S}_0})^{-1} \widetilde{\boldsymbol{W}}_{l\mathcal{S}_0}^T \boldsymbol{Y}}{n \widehat{\sigma}_{\mathcal{S}_0}^2} \Big\} + L(\log n + 2\eta \log p) \end{aligned} \tag{S.9}$$

for $l \in \mathcal{S}_0^c$. It is easy to prove that $\widehat{\sigma}_{\mathcal{S}_0}^2$ converges to $\mathrm{E}\{\epsilon^2\}$ in probability and the details are omitted. We denote $\widetilde{\boldsymbol{W}}_{l\mathcal{S}_0} (\widetilde{\boldsymbol{W}}_{l\mathcal{S}_0}^T \widetilde{\boldsymbol{W}}_{l\mathcal{S}_0})^{-1} \widetilde{\boldsymbol{W}}_{l\mathcal{S}_0}^T$ by $\widetilde{\mathbf{P}}_{l\mathcal{S}_0}$, which is an orthogonal projection matrix. Thus, from (S.9) we have for some positive C,

$$\mathrm{EBIC}(\mathcal{S}_0(l)) - \mathrm{EBIC}(\mathcal{S}_0) \geq -\frac{C}{\mathrm{E}\{\epsilon^2\}} \boldsymbol{Y}^T \widetilde{\mathbf{P}}_{l\mathcal{S}_0} \boldsymbol{Y} + L(\log n + 2\eta \log p) \tag{S.10}$$

uniformly in $l \in \mathcal{S}_0^c$, with probability tending to 1.

Now we evaluate $\boldsymbol{Y}^T \widetilde{\mathbf{P}}_{l\mathcal{S}_0} \boldsymbol{Y}$ on the right-hand side of (S.10). From the definition of $\widetilde{\boldsymbol{W}}_{l\mathcal{S}_0}$, we have

$$\boldsymbol{Y}^T \widetilde{\mathbf{P}}_{l\mathcal{S}_0} \boldsymbol{Y} = (\boldsymbol{Y} - \boldsymbol{W}_{\mathcal{S}_0} \boldsymbol{\gamma}_{\mathcal{S}_0})^T \widetilde{\mathbf{P}}_{l\mathcal{S}_0} (\boldsymbol{Y} - \boldsymbol{W}_{\mathcal{S}_0} \boldsymbol{\gamma}_{\mathcal{S}_0})$$

for any $\boldsymbol{\gamma}_{\mathcal{S}_0} \in \mathbb{R}^{L\#\mathcal{S}_0}$. Therefore we obtain

$$\boldsymbol{Y}^T \widetilde{\mathbf{P}}_{l\mathcal{S}_0} \boldsymbol{Y} \leq \boldsymbol{\epsilon}^T \widetilde{\mathbf{P}}_{l\mathcal{S}_0} \boldsymbol{\epsilon} + |\boldsymbol{b}|^2$$

where $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$ and $\boldsymbol{b}$ is some n-dimensional vector of spline approximation errors satisfying $|\boldsymbol{b}|^2 = O(nL^{-4})$ uniformly in $l \in \mathcal{S}_0^c$. By applying Proposition 3 of Zhang (2010), we obtain

$$\mathrm{P}\Big(\frac{\boldsymbol{\epsilon}^T \widetilde{\mathbf{P}}_{l\mathcal{S}_0} \boldsymbol{\epsilon}}{LC_{E2}} \geq \frac{1+x}{\{1 - 2/(e^{x/2}\sqrt{1+x} - 1)\}_+^2} \Big) \leq \exp(-Lx/2)(1+x)^{L/2}, \tag{S.11}$$

where $\{x\}_+ = \max\{0, x\}$. We take $x = \log(p^{2\eta} n) a_n / 2$ with a_n tending to 0 sufficiently slowly. Then from the above inequality, we have $\boldsymbol{\epsilon}^T \widetilde{\mathbf{P}}_{l\mathcal{S}_0} \boldsymbol{\epsilon} = o_p(L \log(p^{2\eta} n))$ uniformly in $l \in \mathcal{S}_0^c$. Thus we have

$$\boldsymbol{Y}^T \widetilde{\mathbf{P}}_{l\mathcal{S}_0} \boldsymbol{Y} = O(nL^{-4}) + o_p(L \log(p^{2\eta} n)) \tag{S.12}$$

uniformly in $l \in \mathcal{S}_0^c$. Hence the desired result follows from (S.10), (S.12), and the assumption that $L = c_L n^{\bar{\kappa}_L}$ with $\bar{\kappa}_L \geq 1/5$. Note that, here we use the condition that $L \log n / \log p \to \infty$ when $\eta = 0$, which is stated in Assumption C(2). □

S.2.1 Proofs of Lemmas S.2.1 – S.2.3

We use the following inequalities in the proofs of Lemmas S.2.1-S.2.2.

$$\frac{C_{S1}}{L} \leq \lambda_{\min}(\mathrm{E}\{\mathbf{B}(T)\mathbf{B}(T)^T\}) \leq \lambda_{\max}(\mathrm{E}\{\mathbf{B}(T)\mathbf{B}(T)^T\}) \leq \frac{C_{S2}}{L}, \tag{S.13}$$

where C_{S1} and C_{S2} are positive constants independent of L. See Huang, et al. (2008) for the proof of (S.13).

Proof of Lemma S.2.1. Write

$$n^{-1}D_{lSn} = n^{-1}\sum_{i=1}^{n}(\boldsymbol{X}_{iS(l)}\boldsymbol{X}_{iS(l)}^T) \otimes (\mathbf{B}(T_i)\mathbf{B}(T_i)^T), \tag{S.14}$$

where $\boldsymbol{X}_{iS(l)}$ is the ith sample version of $\boldsymbol{X}_{S(l)}$ and $\otimes$ is the kronecker product. Note that (S.13), (S.14), and Assumption X(2) imply that, for any $\delta > 0$,

$$\frac{C_1}{L} \leq \lambda_{\min}(D_{lS}) \leq \lambda_{\max}(D_{lS}) \leq \frac{C_2}{L}. \tag{S.15}$$

for some positive C_1 and C_2. In addition, by exploiting the band-diagonal property of D_{lSn} and D_{lS} and an exponential inequality, we can demonstrate that

$$|D_{lSn} - D_{lS}| \leq n^{-1}\delta p_0 \tag{S.16}$$

uniformly in $S \subsetneq \mathcal{S}_0$ and $l \in S^c$ with probability

$$1 - C_3 p_0^2 L \exp\{-\delta^2(C_4 n L^{-1} + C_5\delta)^{-1}\} \times p\exp(p_0 \log 2), \tag{S.17}$$

where C_3, C_4, and C_5 are positive constants independent of p_0, L, n, p, and δ. When we take $\delta = n^{1-c_\beta/4}L^{-1}$, the probability in (S.17) tends to 0 and the former result follows since $\delta p_0/n = p_0 n^{-c_\beta/4}/L = o(L^{-1})$. The latter result follows from the following relationship between D_{lSn}^{-1} and $n^{-1}\widetilde{\boldsymbol{W}}_{lS}^T\widetilde{\boldsymbol{W}}_{lS}$:

$$D_{lSn}^{-1} = \begin{pmatrix} * & * \\ * & \left(n^{-1}\widetilde{\boldsymbol{W}}_{lS}^T\widetilde{\boldsymbol{W}}_{lS}\right)^{-1} \end{pmatrix}.$$

□

Proof of Lemma S.2.2. Let $\{b_j\}_{j\in S(l)}$ be a set of square integrable functions on $[0,1]$. Then Assumption X(2) implies that

$$C_{X2}\sum_{j\in S(l)}\|b_j(T)\|^2 \le \|\sum_{j\in S(l)} X_j b_j(T)\|^2 \le C_{X3}\sum_{j\in S(l)}\|b_j(T)\|^2. \tag{S.18}$$

Besides, Assumption T implies

$$C_{T1}\|b\|_{L_2}^2 \le \|b(T)\|^2 \le C_{T2}\|b\|_{L_2}^2 \tag{S.19}$$

for any square integrable function b. In addition, due to Assumptions C(4) and C(5), we can choose some positive constant C_1 and a set of L-dimensional vectors $\{\tilde{\gamma}_j\}_{j\in S(l)}$ such that

$$\sum_{j\in S(l)}\|\overline{\beta}_j - \tilde{\gamma}_j^T\mathbf{B}\|_\infty \le C_1 L^{-2}, \tag{S.20}$$

where C_1 depends only on the assumptions. By exploiting (S.18)-(S.20), we obtain

$$\begin{aligned}
&C_{X2}\sum_{j\in S(l)}\|\overline{\beta}_j(T) - \overline{\gamma}_j^T\mathbf{B}(T)\|^2 \\
\le\ &\|\sum_{j\in S(l)}(\overline{\beta}_j(T) - \overline{\gamma}_j^T\mathbf{B}(T))X_j\|^2 \ \le\ \|\sum_{j\in S(l)}(\overline{\beta}_j(T) - \tilde{\gamma}_j^T\mathbf{B}(T))X_j\|^2 \\
\le\ &\sum_{j\in S(l)}\|\overline{\beta}_j(T) - \tilde{\gamma}_j^T\mathbf{B}(T)\|^2 \ \le\ C_{X3}\sum_{j\in S(l)}\|\overline{\beta}_j - \tilde{\gamma}_j^T\mathbf{B}\|_\infty^2 \ \le\ C_{X3}C_1^2L^{-4}.
\end{aligned}$$

Therefore, there is a positive constant C_2 such that

$$\|\overline{\beta}_j(T) - \overline{\gamma}_j^T\mathbf{B}(T)\| \le C_2L^{-2}.$$

This implies that

$$\|\overline{\beta}_j(T)\| - C_2L^{-2} \le \left\{\overline{\gamma}_j^T\mathrm{E}\{\mathbf{B}(T)\mathbf{B}(T)^T\}\overline{\gamma}_j\right\}^{1/2} \le \|\overline{\beta}_j(T)\| + C_2L^{-2}. \tag{S.21}$$

The desired result follows from (S.13) and (S.21). □

Proof of Lemma S.2.3. First we deal with $|d_{lS}|$ and $|d_{lSn} - d_{lS}|$. We have $|d_{lS}| \le C_1(p_0/L)^{1/2}$ from the definition of the B-spline basis. Also, as in the proof of Lemma 2 of Cheng, et al. (2014). we have

$$|d_{lSn} - d_{lS}| \le \delta(Lp_0)^{1/2}/n$$

uniformly in $S \subsetneq \mathcal{S}_0$ and $l \in S^c$ with probability

$$1 - C_2p_0L\exp\{-\delta^2(C_3nL^{-1} + C_4\delta)^{-1}\} \times p\exp(p_0\log 2),$$

where C_2, C_3, and C_4 are positive constants independent of p_0, L, n, p, and δ. By combining the above results, (S.16), and Lemma S.2.1, we obtain

$$|D_{lSn}^{-1}(d_{lSn} - d_{lS})| \leq C_5 L^{3/2} p_0^{1/2} \delta / n \tag{S.22}$$

and

$$|(D_{lSn}^{-1} - D_{lS}^{-1}) d_{lS}| \leq |D_{lS}^{-1}||D_{lSn} - D_{lS}||D_{lSn}^{-1}||d_{lS}| \leq C_5 L^{3/2} p_0^{3/2} \delta / n \tag{S.23}$$

uniformly in $S \subsetneq \mathcal{S}_0$ and $l \in S^c$, with probability given in the lemma. Note that C_5 is independent of p_0, L, n, and δ. Hence the desired result follows from (S.22) and (S.23). □

S.3 Additional simulation results

In this section we present more simulation results. In particular, we summarize results of the simulations considered in Examples 1 and 2 in Section 4.1 when the noise level is large ($\sigma = 2$). Comparing Tables S.1 and S.2 with Tables 2 and 3 given in the paper, we can see that when $\sigma = 2$ we still have similar conclusions as when $\sigma = 1$. In addition, when $\sigma = 2$ all the four considered approaches, fAIC, fBIC, fEBIC, and gNIS, perform worse than they do when $\sigma = 1$. This is expected because, as the signal-to-noise ratio decreases, less information is available from the data and it is harder to detect the true variables. Overall, we can conclude that the fBIC is more reliable than the other three.

References

[1] M.-Y. Cheng, T. Honda, J. Li and H. Peng, Nonparametric independence screening and structure identification for ultra-high dimensional longitudinal data, *Ann. Statist.* 42:1819-1849, 2012.

[2] S. Luo and Z. Chen, Sequential Lasso cum EBIC for feature selection with ultra-high dimensional feature space, *J. Amer. Statist. Assoc.* 109:1229-1240, 2014.

[3] J.Z. Huang, C.O. Wu, and L. Zhou, Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statistica Sinica* 14: 763-788, 2004.

Table S.1: Average numbers of true positive (TP) and false positive (FP) over 1000 repetitions and their robust standard deviations (in parentheses) for the gNIS, fAIC, fBIC and fEBIC approaches under the varying coefficient model defined in Example 1 with $\sigma = 2$ and $p = 1000$.

			$n = 200$		$n = 300$		$n = 400$	
$[t_1, t_2]$	SNR	Approach	TP	FP	TP	FP	TP	FP
$[0, 0]$	4.32	gNIS	3.77(0.00)	9.47(16.4)	3.91(0.00)	8.24(20.9)	4.00(0.00)	0.85(0.75)
		fAIC	4.00(0.00)	65.9(11.9)	4.00(0.00)	83.5(13.4)	4.00(0.00)	106(15.7)
		fBIC	4.00(0.00)	7.27(0.00)	4.00(0.00)	0.01(0.00)	4.00(0.00)	0.00(0.00)
		fEBIC	2.91(0.00)	0.00(0.00)	3.95(0.00)	0.00(0.00)	4.00(0.00)	0.00(0.00)
$[2, 0]$	0.91	gNIS	0.41(0.00)	0.43(0.00)	0.26(0.00)	0.00(0.00)	0.59(0.00)	0.00(0.00)
		fAIC	2.89(0.75)	69.3(13.4)	3.64(0.75)	85.8(16.8)	3.93(0.00)	107(19.4)
		fBIC	1.97(0.00)	1.98(0.75)	2.55(0.75)	0.08(0.00)	3.12(0.75)	0.01(0.00)
		fEBIC	0.93(0.00)	0.07(0.00)	0.99(0.00)	0.01(0.00)	1.07(0.00)	0.00(0.00)
$[2, 1]$	0.80	gNIS	0.09(0.00)	0.07(0.00)	0.22(0.00)	0.06(0.00)	0.27(0.00)	0.03(0.00)
		fAIC	2.89(1.50)	83.3(14.2)	3.62(0.75)	104(17.2)	3.90(0.00)	133(23.1)
		fBIC	1.93(0.00)	0.85(0.75)	2.36(0.75)	0.06(0.00)	2.90(0.75)	0.01(0.00)
		fEBIC	0.96(0.00)	0.04(0.00)	0.99(0.00)	0.01(0.00)	1.01(0.00)	0.00(0.00)
$[3, 1]$	0.69	gNIS	0.37(0.75)	0.27(0.00)	0.43(0.75)	0.14(0.00)	0.74(0.75)	0.10(0.00)
		fAIC	1.96(1.50)	84.1(15.6)	2.78(0.75)	105(19.40)	3.37(0.75)	131(24.63)
		fBIC	1.11(0.75)	0.80(0.75)	1.63(0.75)	0.33(0.75)	2.01(0.00)	0.12(0.00)
		fEBIC	0.78(0.00)	0.22(0.00)	0.94(0.00)	0.06(0.00)	0.99(0.00)	0.01(0.00)
$[4, 5]$	0.75	gNIS	0.04(0.00)	0.01(0.00)	0.04(0.00)	0.00(0.00)	0.12(0.00)	0.00(0.00)
		fAIC	1.39(0.75)	75.4(16.4)	2.09(1.12)	92.3(18.7)	2.72(0.75)	115.0(22)
		fBIC	0.64(0.75)	1.00(0.75)	1.05(0.00)	0.43(0.75)	1.52(0.75)	0.30(0.75)
		fEBIC	0.54(0.75)	0.46(0.75)	0.82(0.00)	0.18(0.00)	0.95(0.75)	0.05(0.00)
$[6, 8]$	0.80	gNIS	0.01(0.00)	0.02(0.00)	0.00(0.00)	0.00(0.00)	0.01(0.00)	0.00(0.00)
		fAIC	0.83(0.75)	73.9(16.4)	1.23(0.75)	90.3(20.9)	1.66(0.75)	110(21.64)
		fBIC	0.26(0.75)	1.22(0.00)	0.47(0.75)	0.57(0.75)	0.69(0.75)	0.43(0.75)
		fEBIC	0.25(0.75)	0.75(0.75)	0.46(0.75)	0.54(0.75)	0.64(0.75)	0.36(0.75)

Table S.2: The same caption as that of Table S.1 but now under the varying coefficient model defined in Example 2 with $\sigma = 2$ and $p = 1000$.

			$n = 200$		$n = 300$		$n = 400$	
$[t_1, t_2]$	SNR	Approach	TP	FP	TP	FP	TP	FP
$[0, 0]$	11.9	gNIS	7.23(0.75)	10.0(13.4)	7.31(0.75)	10.3(19.4)	7.61(0.00)	13.5(27.6)
		fAIC	8.00(0.00)	61.6(11.2)	8.00(0.00)	78.9(12.7)	8.00(0.00)	102(15.67)
		fBIC	8.00(0.00)	48.9(10.4)	8.00(0.00)	0.24(0.00)	8.00(0.00)	0.00(0.00)
		fEBIC	1.01(0.00)	0.00(0.00)	1.24(0.00)	0.00(0.00)	5.51(5.22)	0.00(0.00)
$[2, 0]$	2.36	gNIS	1.41(2.24)	2.49(0.00)	1.32(2.24)	0.10(0.00)	2.99(4.48)	0.08(0.00)
		fAIC	4.99(1.49)	67.7(14.2)	7.05(0.75)	83.9(15.7)	7.77(0.00)	105(20.15)
		fBIC	3.80(1.49)	15.8(3.73)	5.08(1.49)	0.37(0.00)	6.34(0.75)	0.04(0.00)
		fEBIC	0.85(0.00)	0.15(0.00)	0.97(0.00)	0.03(0.00)	1.03(0.00)	0.00(0.00)
$[2, 1]$	2.18	gNIS	0.50(0.00)	0.39(0.00)	1.03(0.00)	0.25(0.00)	2.94(5.22)	0.71(0.75)
		fAIC	5.06(1.49)	80.5(14.2)	7.07(0.75)	103(18.65)	7.79(0.00)	129(21.6)
		fBIC	3.25(1.49)	7.73(1.49)	4.82(1.49)	0.22(0.00)	6.06(1.49)	0.04(0.00)
		fEBIC	0.88(0.00)	0.12(0.00)	0.98(0.00)	0.02(0.00)	1.03(0.00)	0.00(0.00)
$[3, 1]$	1.91	gNIS	0.47(0.75)	0.42(0.00)	1.15(1.49)	0.17(0.00)	2.27(2.99)	0.30(0.00)
		fAIC	3.58(1.12)	83.2(15.7)	5.32(0.75)	105(20.89)	6.75(0.75)	131(23.88)
		fBIC	1.59(0.75)	1.87(1.11)	2.38(0.75)	0.38(0.75)	3.73(1.49)	0.17(0.00)
		fEBIC	0.74(0.75)	0.26(0.75)	0.92(0.00)	0.08(0.00)	0.98(0.00)	0.02(0.00)
$[4, 5]$	2.31	gNIS	0.11(0.00)	0.02(0.00)	0.15(0.00)	0.00(0.00)	0.43(0.75)	0.00(0.00)
		fAIC	2.56(0.75)	74.6(15.7)	4.11(1.49)	92.8(19.4)	5.44(0.75)	115(21.64)
		fBIC	1.01(0.75)	1.82(0.75)	1.55(0.75)	0.62(0.75)	2.30(0.75)	0.31(0.75)
		fEBIC	0.54(0.75)	0.46(0.75)	0.80(0.00)	0.20(0.00)	0.93(0.00)	0.07(0.00)
$[6, 8]$	2.60	gNIS	0.02(0.00)	0.03(0.00)	0.01(0.00)	0.00(0.00)	0.07(0.00)	0.00(0.00)
		fAIC	1.67(0.75)	73.4(14.9)	2.49(0.75)	90.9(19.4)	3.62(0.75)	113(25.37)
		fBIC	0.41(0.75)	1.39(0.00)	0.68(0.75)	0.83(0.75)	1.12(0.75)	0.79(0.75)
		fEBIC	0.32(0.75)	0.68(0.75)	0.53(0.75)	0.47(0.75)	0.70(0.75)	0.30(0.75)