Taylor & Francis
Taylor & Francis Group

Journal of Nonparametric Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/gnst20

A model-free conditional screening approach via sufficient dimension reduction

Lei Huo , Xuerong Meggie Wen & Zhou Yu

To cite this article: Lei Huo , Xuerong Meggie Wen & Zhou Yu (2020): A model-free conditional screening approach via sufficient dimension reduction, Journal of Nonparametric Statistics, DOI: 10.1080/10485252.2020.1834554

To link to this article: https://doi.org/10.1080/10485252.2020.1834554

Published online: 03 Nov 2020.

JOURNAL OF NONPARAMETRIC STATISTICS
https://doi.org/10.1080/10485252.2020.1834554

Taylor & Francis
Taylor & Francis Group

A model-free conditional screening approach via sufficient dimension reduction

Lei Huo[a], Xuerong Meggie Wen[a] and Zhou Yu[b]

[a]Department of Mathematics and Statistics, Missouri University of Science and Technology, Rolla, MO, USA; [b]School of Statistics, East China Normal University, Shanghai, People's Republic of China

ABSTRACT

Conditional variable screening arises when researchers have prior information regarding the importance of certain predictors. It is natural to consider feature screening methods conditioning on these known important predictors. Barut, E., Fan, J., and Verhasselt, A. [(2016), 'Conditional Sure Independence Screening', *Journal of the American Statistical Association*, 111, 1266–1277] proposed conditional sure independence screening (CSIS) to address this issue under the context of generalised linear models. While CSIS outperforms the marginal screening method when few of the factors are known to be important and/or significant correlations between some of the factors exist, unfortunately, CSIS is model based and might fail when the models are misspecified. We propose a model-free conditional screening method under the framework of sufficient dimension reduction for ultrahigh dimensional statistical problems. Numerical studies show our method easily beats CSIS for nonlinear models and performs comparable to CSIS for (generalised) linear models. Sure screening consistency property for our method is proved.

ARTICLE HISTORY
Received 22 August 2019
Accepted 27 September 2020

KEYWORDS
Conditional screening; trace pursuit; variable selection; sufficient dimension reduction

2010 MATHEMATICS SUBJECT CLASSIFICATION
62G05

1. Introduction

Researchers in many fields, such as economics and finance, need to analyse high-dimensional data, where the number of predictors p is frequently huge compared with the sample size n. Most traditional statistical methods failed when p is large. Also, with high-dimensional data, it is often reasonable to assume only a small number of predictors actually contribute to the response (sparsity assumption). Estimation accuracy and model interpretability can be greatly improved in the subsequent analysis by effectively identifying the few important predictors first. Hence, dimension reduction or feature selection is often conducted as the first step of data analysis. Fan and Lv (2008) proposed the sure independence screening (SIS), which is a feature screening procedure for linear models by ranking the marginal correlations between the response and each individual predictor. SIS has the so-called *sure screening property* (Fan and Lv 2008), in the sense that as $n \to \infty$, the important predictors are guaranteed to be retained in the model with probability tending

CONTACT Xuerong Meggie Wen wenx@mst.edu Department of Mathematics and Statistics, Missouri University of Science and Technology, Rolla 65409 MO, USA

to 1, even for ultra-high-dimensional predictor space, where p can diverge at an exponential rate of n. SIS was extended to generalised linear models in Fan and Song (2010). Fan, Feng, and Song (2011) proposed nonparametric independence screening (NIS) for nonparametric models with additive structure using nonparametric marginal ranking. Many other feature screening methodologies have been developed, such as Xue and Zou (2011), Wang (2012), Zhao and Li (2012), Chang, Tang, and Wu (2013) and Niu, Zhang, Liu, and Li (2020).

However, all the aforementioned procedures are model-based and might yield poor performance when the models are misspecified. Motivated by this fact, model-free feature screening procedures, which can identify the important predictors without specifying the model structure, were developed. To list a few: Zhu, Li, Li, and Zhu (2011) proposed a sure independent ranking and screening (SIRS) method. Lin, Sun, and Zhu (2013) proposed a nonparametric ranking feature screening (NRS) using the function-correlation between the response and the predictors. He, Wang, and Hong (2013) proposed quantile-adaptive model-free screening through the marginal quantile regression. Mai and Zou (2015) proposed the fused Kolmogorov filter approach, which performs feature screening for the data with many types of predictors and response. For discriminant analysis with high-dimensional data, model-free feature screening has been studied by Mai and Zou (2013), Cui, Li, and Zhong (2014), and Pan, Wang, and Li (2015).

Recently, under the paradigm of sufficient dimension reduction (Li 1991; Cook 1998), which aims to find the linear combinations of the predictors such that the response is independent with the original predictors given these linear combinations without requiring the knowledge of the model structure, Yu, Dong, and Zhu (2016) proposed a novel model-free feature screening method, the *forward trace pursuit (FTP)*. Although Yu et al. (2016) focused on conducting model-free feature screening via the three most well-known sufficient dimension reduction methods: sliced inverse regression (SIR) (Li 1991), sliced average variance estimation (SAVE) (Cook and Weisberg 1991), and directional regression (DR) (Li and Wang 2007), the general principle of FTP can be easily extended to other sufficient dimension reduction methods. The screening consistency property of forward regression in linear models is established in Wang (2009), which is extended to model-free setting via SIR-based forward trace pursuit in Yu et al. (2016).

The performance of all these feature screening procedures is heavily influenced by the correlations among the predictors, as mentioned in Fan and Lv (2008), Zhu et al. (2011), and Barut, Fan, and Verhasselt (2016). As Barut et al. (2016) pointed out, the correlations among predictors might cause false positives (where the unimportant predictors are mistakenly considered as important ones through the screening procedure), and/or false negatives (where the important predictors are screened out as the unimportant ones). Unfortunately, the correlations among predictors are unavoidable for high-dimensional data analysis (Hall and Li 1993; Fan and Lv 2008), since spurious correlations among predictors always exist as p diverges. To obtain the sure screening property, feature screening procedures usually need to impose some restrictions on the correlation structure among predictors.

One possible way to alleviate the above problem is to consider conditional screening method, since researchers in many applications have some prior information regarding the importance of certain predictors, such as the treatment effects in biological studies and

market risk factors in financial studies, it is natural to consider feature screening methods conditioning on these known important predictors.

For example, consider the leukaemia data studied by Golub et al. (1999), Barut et al. (2016) and others, where gene expression data from 72 patients with two types of acute leukaemia, acute lymphoblastic leukaemia (ALL) and acute myeloid leukaemia (AML) were collected. Gene expression levels were measured for 7129 genes. Golub et al. (1999) described that two genes, Zyxin and Transcriptional activator hSNF2b, had empirically high correlations for the difference between patients with AML and ALL. Barut et al. (2016) proposed a conditional screening method called *conditional sure independence screening (CSIS)* to conduct screening procedure in the presence of the known set of predictors. They applied CSIS to the aforementioned leukaemia data conditioning on the two known genes, and was able to select TCRD (T-cell receptor delta locus) which had not been detected before. Numerical studies also showed that, compared with SIS, CSIS makes it possible to identify those significant hidden predictors whose contributions might otherwise get cancelled out due to the correlations with other predictors. Also, when there are high correlations among significant predictors and insignificant ones, CSIS can help to reduce the number of false negatives.

Although CSIS improves the performance of the screening procedure by using prior information, however, it is still a model-based screening procedure for generalised linear models, it might fail when the model assumption is not satisfied. To address this issue, we propose a model-free conditional screening method via sufficient dimension reduction in this article. Conditioning on a few known important predictors which should always be included in the regression, we conduct feature screening procedure on the remaining predictors without assuming an underlying model between the response and predictors. Specifically, our method is based on the partial sufficient dimension reduction procedure proposed by Feng, Wen, Yu, and Zhu (2013). We employ the framework of partial sufficient dimension reduction while splitting the original predictors into two groups: those known important ones and the rest of the predictors which we will conduct feature screening on. Chang, Tang, and Wu (2016) proposed a nonparametric local independent feature screening method using the marginal empirical likelihood in conjunction with marginal kernel smoothing methods. They also developed an iterative version to deal with the situation that some predictors are marginally unrelated but jointly related to the response, which is different from our conditioning feature screening approach. Lu and Lin (2020) and Chu and Lin (2020) also studied model-free conditional feature screening methods via conditional distance correlation and empirical likelihood, respectively.

The rest of this paper is organised as the following. In Section 2, we briefly review partial sufficient dimension reduction. We then propose our model-free conditional screening method and discuss its properties in Section 3. Numerical studies and real data analysis are provided in Section 4. A brief discussion and conclusion are given in Section 5. We defer all proofs to the Appendix.

2. Partial sufficient dimension reduction

In this section, we give a brief introduction of the partial sufficient dimension reduction since our model-free conditional screening method is based on it. For a typical regression analysis with a response variable Y and a vector of random predictors $\mathbf{X} = (X_1, \ldots, X_p)^{\mathrm{T}} \in$

$\mathbb{R}^p$, we seek a parsimonious characterisation of the conditional distribution of $Y \mid \mathbf{X}$. Li (1991) and Cook (1998) proposed sufficient dimension reduction to reduce the dimension of $\mathbf{X}$ without loss of information on the original regression and without requiring a pre-specified parametric model. The basic idea is to replace $\mathbf{X}$ by a minimal set of linear combinations of $\mathbf{X}$ without loss of information on $Y \mid \mathbf{X}$. So we seek a $p \times d$ matrix $\boldsymbol{\eta}$, with $d \leq p$ such that

$$Y \perp\!\!\!\perp \mathbf{X} \mid \boldsymbol{\eta}^T \mathbf{X}, \tag{1}$$

where $\perp\!\!\!\perp$ indicates independence. When (1) holds, to study the relationship between $\mathbf{X}$ and Y, it is sufficient to focus only on the d reduced variables $\beta_i^T \mathbf{X}$.

Partial sufficient dimension reduction (Chiaromonte, Cook, and Li 2002; Feng et al. 2013) arises when one considers the predictive role of all predictors but limits dimension reduction to a subset of the predictors. Those predictors which dimension reduction is performed on are referred to as the predictors of primary interest, and the rest of predictors are referred to as the predictors of secondary interest. Partial dimension reduction is of practical use, since in many applications, some predictors play a particular role and must be shielded from the dimension reduction process. Considering the leukaemia data discussed in Section 1, the two predictors (genes), Zyxin and Transcriptional activator hSNF2b, are the predictors of 'secondary interest', since prior knowledge indicated that further dimension reduction should be conducted on other predictors (genes) while conditioning on these two predictors.

Let Y be a univariate random response, $\mathbf{X} = \{X_1, X_2, \ldots, X_p\} \in \mathbf{R}^p$ be a vector of continuous predictors of primary interest, and $\mathbf{W} = \{W_1, W_2, \ldots, W_q\} \in \mathbf{R}^q$ be a vector of predictors of secondary interest. The aim of partial sufficient dimension reduction is to find the partial central subspace $\mathcal{S}_{Y|\mathbf{X}}^{(\mathbf{W})}$, which is the intersection of all subspaces $\mathcal{S}$ such that

$$Y \perp\!\!\!\perp \mathbf{X} \mid (P_{\mathcal{S}}\mathbf{X}, \mathbf{W}), \tag{2}$$

where $\perp\!\!\!\perp$ stands for independence and $P_{\mathcal{S}}$ is the orthogonal projection on subspace $\mathcal{S}$. The concept of partial central subspace was first proposed by Chiaromonte et al. (2002) to deal with dimension reductions for regressions with a mixture of continuous and categorical predictors where the dimension reduction procedure focused on continuous predictors. Although it expands the scope of sufficient dimension reduction with practical applications, the method developed by Chiaromonte et al. (2002) is only limited to situations where $\mathbf{W}$ is categorical, and is difficult to be extended to cases with continuous $\mathbf{W}$. Hilafu and Wu (2017) proposed partial projective resampling dimension reduction (PPR-DR) to estimate the partial central subspace for any types of $\mathbf{W}$ by changing the role of $\mathbf{W}$ from predictor to the response variable. However, the subspace they estimated is larger than the partial central subspace when $\mathbf{W}$ is not independent with $\mathbf{X}$ given $P_{\mathcal{S}_{Y|\mathbf{X}}^{(\mathbf{W})}}\mathbf{X}$.

Feng et al. (2013) proposed partial discretisation–expectation estimation (PDEE) to estimate the partial central subspace $\mathcal{S}_{Y|\mathbf{X}}^{(\mathbf{W})}$ when $\mathbf{W}$ is continuous, which our model-free conditional screening method is based on. A brief review of PDEE is given below. First, the continuous $\mathbf{W}$ is discretised into a set of binary variables by defining $\mathbf{W}(\mathbf{T}) = (I_{\{W_1 \leq T_1\}}, I_{\{W_2 \leq T_2\}}, \ldots, I_{\{W_q \leq T_q\}})$, where $\mathbf{T} = \{T_1, T_2, \ldots, T_q\} \in \mathbf{R}^q$ is an independent copy of $\mathbf{W}$ with support of $\mathbf{R}_{\mathbf{T}}^q$, and $I_{\{W_i \leq T_i\}}$ is an indicator function taking value 1 for

$W_i \le T_i$, and 0 otherwise, for $i = 1, \ldots, q$. Then, let $\mathcal{S}_{Y|\mathbf{X}}^{\mathbf{W(t)}}$ be the partial central subspace of $Y \mid (\mathbf{X}, \mathbf{W(t)})$, for $\mathbf{T} = \mathbf{t} \in \mathbf{R}_{\mathbf{T}}^q$, Feng et al. (2013) showed that

$$\mathcal{S}_{Y|\mathbf{X}}^{(\mathbf{W})} = \bigcup_{\mathbf{t} \in \mathbf{R}_{\mathbf{T}}^q} \mathcal{S}_{Y|\mathbf{X}}^{\mathbf{W(t)}}. \tag{3}$$

Hence, an estimate of $\mathcal{S}_{Y|\mathbf{X}}^{(\mathbf{W})}$ can be obtained via those $\mathcal{S}_{Y|\mathbf{X}}^{\mathbf{W(t)}}$.

For simplicity, $(Y, \mathbf{X}) \mid \mathbf{W(t)}$ is denoted as $(\mathbf{X^t}, Y^{\mathbf{t}})$ for any fixed $\mathbf{t} \in \mathbf{R}_{\mathbf{T}}^q$. We can construct kernel matrices $\mathbf{M(t)}$ such that $\mathrm{Span}\{\mathbf{M(t)}\} = \mathcal{S}_{Y|\mathbf{X}}^{\mathbf{W(t)}}$ to infer about the partial central subspace $\mathcal{S}_{Y|\mathbf{X}}^{\mathbf{W(t)}}$. Note that (3) not only provides a general framework for estimating the partial central subspace, it can also be combined with many different sufficient dimension reduction methods by choosing different kernel matrices $\mathbf{M(t)}$. The following are the kernel matrices of the three most popular sufficient dimension reduction methods:

$$\begin{aligned}
\text{SIR:} \quad & \mathbf{M(t)} = c^{-1}\mathrm{Var}\{\mathrm{E}(\mathbf{X^t} \mid Y^{\mathbf{t}})\}\boldsymbol{\Sigma}_{\mathbf{t}}^{-1};\\
\text{SAVE:} \quad & \mathbf{M(t)} = \boldsymbol{\Sigma}_{\mathbf{t}}^{-1}\mathrm{E}\{\boldsymbol{\Sigma}_{\mathbf{t}} - \mathrm{Var}(\mathbf{X^t} \mid Y^{\mathbf{t}})\}^2\boldsymbol{\Sigma}_{\mathbf{t}}^{-1};\\
\text{DR:} \quad & \mathbf{M(t)} = \boldsymbol{\Sigma}_{\mathbf{t}}^{-1}\mathrm{E}\{2\boldsymbol{\Sigma}_{\mathbf{t}} - \mathrm{E}((\widetilde{\mathbf{X}}^{\mathbf{t}} - \mathbf{X^t})(\widetilde{\mathbf{X}}^{\mathbf{t}} - \mathbf{X^t})^T \mid Y^{\mathbf{t}}, \widetilde{Y}^{\mathbf{t}})\}^2\boldsymbol{\Sigma}_{\mathbf{t}}^{-1},
\end{aligned}$$

where $\boldsymbol{\Sigma}_{\mathbf{t}} = \mathrm{Var}(\mathbf{X^t})$, and $(\widetilde{Y}^{\mathbf{t}}, \widetilde{\mathbf{X}}^{\mathbf{t}})$ is an independent copy of $(Y^{\mathbf{t}}, \mathbf{X^t})$. Interested readers may refer to Li and Dong (2009) and Li, Kim, and Altman (2010) for further details.

The following conditions are commonly used in sufficient dimension reduction area to ensure that $\mathrm{Span}\{\mathbf{M(t)}\} = \mathcal{S}_{Y|\mathbf{X}}^{\mathbf{W(t)}}$ holds for the above choices of $\mathbf{M(t)}$.

Condition 2.1: For any $\mathbf{t} \in \mathbf{R}_{\mathbf{T}}^q$, we assume that

(a) $\mathrm{E}(\mathbf{X^t} \mid P_{\mathcal{S}_{Y|\mathbf{X}}^{\mathbf{W(t)}}}\mathbf{X^t})$ is linear combination of $P_{\mathcal{S}_{Y|\mathbf{X}}^{\mathbf{W(t)}}}\mathbf{X^t}$;
(b) $\mathrm{Var}(\mathbf{X^t} \mid P_{\mathcal{S}_{Y|\mathbf{X}}^{\mathbf{W(t)}}}\mathbf{X^t})$ is nonrandom.

Condition 2.1(a) is also called the linear conditional mean (LCM) assumption, while Condition 2.1(b) is the constant conditional variance (CCV) assumption. Both conditions hold for normally distributed $\mathbf{X}$. When $\mathbf{X}$ is not normally distributed, please refer to Cook and Nachtsheim (1994), Li and Dong (2009), Dong and Li (2010) for possible options. SIR (Li 1991) only requires Condition 2.1(a), while SAVE (Cook and Weisberg 1991) and DR (Li and Wang 2007) need both conditions.

Feng et al. (2013) showed that it suffices to take the expectation over the aforementioned random vector $\mathbf{T}$ (an independent copy of $\mathbf{W}$) to obtain the target matrix $\mathbf{M} = \mathrm{E}\{\mathbf{M(T)}\}$ such that $\mathrm{Span}\{\mathbf{M}\} = \mathcal{S}_{Y|\mathbf{X}}^{(\mathbf{W})}$.

3. Conditional screening through trace pursuit

For model-free conditional screening, we set $\mathbf{W}$ as the set of predictors which should be retained in the model based on the prior knowledge, and perform feature screening on $\mathbf{X}$

while conditioning on $\mathbf{W}$. We seek the smallest active index set $\mathcal{A}$ such that

$$Y \perp\!\!\!\perp \mathbf{X}_{\mathcal{A}^c} \mid (\mathbf{X}_{\mathcal{A}}, \mathbf{W}), \tag{4}$$

where $\mathcal{A}^c$ is the complement set of $\mathcal{A}$ with respective to the index set $\mathcal{I} = \{1, \ldots, p\}$. From (4), it is obvious that $\mathbf{X}_{\mathcal{A}}$ just includes all important predictors for predicting Y given $\mathbf{W}$. Without loss of generality, we may assume the active index set $\mathcal{A} = \{1, \ldots, K\}$ for ease of exposition. We can see that (4) is equivalent to $P_{\mathcal{H}} \mathcal{S}^{(\mathbf{W})}_{Y|\mathbf{X}} = \mathcal{O}_p$, where $\mathcal{H} = \text{Span}\{(\mathbf{0}_{(p-K)\times K}, \mathbf{I}_{p-K})^T\}$ is the subspace of the primary predictor space, corresponding to the coordinates of the inactive predictors, and $\mathcal{O}_p$ is the origin in $\mathbb{R}^p$.

Cook (2004) first considered variable selection via a testing hypothesis approach by testing $Y \perp\!\!\!\perp \mathbf{X}_{\mathcal{A}^c} \mid \mathbf{X}_{\mathcal{A}}$, when the predictors are treated indiscriminately. Under the context of the regression of Y versus $\mathbf{X}$, Cook (2004) proposed a test for the testing hypothesis of $Y \perp\!\!\!\perp \mathbf{X}_{\mathcal{A}^c} \mid \mathbf{X}_{\mathcal{A}}$ based on a generalised least square rederivation of the SIR estimator for $\mathcal{S}_{Y|\mathbf{X}}$. Shao, Cook, and Weisberg (2007) and many others also investigated the same testing problem based on other estimators of $\mathcal{S}_{Y|\mathbf{X}}$. Zhong, Zhang, Zhu, and Liu (2012) and Jiang and Liu (2013) tackled the problem when $n < p$ via sliced inverse regression (SIR) method. However, both methods require the estimation of the rank of $\mathcal{S}_{Y|\mathbf{X}}$ (the so-called order determination), which is a very challenging problem when $n < p$. The trace pursuit approach proposed by Yu et al. (2016) successfully circumvents the need for order determination to conduct model-free variable selection via sufficient dimension reduction approach for $n < p$. In this article, we will conduct conditional variable screening via testing approach (4) from the partial sufficient dimension reduction perspective. We give a detailed discussion of our method using SIR (Li 1991), though we can extend our approach to other sufficient dimension reduction methods such as SAVE (Cook and Weisberg 1991) and DR (Li and Wang 2007) by using different kernel matrices $\mathbf{M}$.

Let $\boldsymbol{\mu}_{\mathbf{t}} = \text{E}(\mathbf{X}^t)$, $\mathbf{Z}^t = \boldsymbol{\Sigma}_{\mathbf{t}}^{-1/2}(\mathbf{X}^t - \boldsymbol{\mu}_t)$ and denote the $\mathbf{Z}$-scaled central space as $\mathcal{S}^{\mathbf{W(t)}}_{Y|\mathbf{Z}}$. By the so-called invariance property Cook (1998), $\mathcal{S}^{\mathbf{W(t)}}_{Y|\mathbf{X}} = \boldsymbol{\Sigma}_{\mathbf{t}}^{-1/2} \mathcal{S}^{\mathbf{W(t)}}_{Y|\mathbf{Z}}$. We will work with the $\mathbf{Z}$-scaled central spaces first in the following discussions. For any given $\mathbf{t} \in \mathbf{R}^q_{\mathbf{T}}$, partition the range of $Y^{\mathbf{t}}$ into $H_{\mathbf{t}}$ nonoverlapping slices $J^{\mathbf{t}}_1, \ldots, J^{\mathbf{t}}_{H_{\mathbf{t}}}$. Let $p_{h_{\mathbf{t}}} = \Pr(Y^{\mathbf{t}} \in J^{\mathbf{t}}_{h_{\mathbf{t}}})$, $\mathbf{U}_{h_{\mathbf{t}}} = \text{E}(\mathbf{X}^t \mid \mathbf{Y}^t \in J^{\mathbf{t}}_{h_{\mathbf{t}}}) - \boldsymbol{\mu}_{\mathbf{t}}$, then the SIR-based $\mathbf{Z}$-scaled kernel matrix $\mathbf{M} = \text{E}\{\mathbf{M}(t)\} = \text{E}\{\boldsymbol{\Sigma}_{\mathbf{t}}^{-1/2}(\sum_{h_{\mathbf{t}}=1}^{H_{\mathbf{t}}} p_{h_{\mathbf{t}}} \mathbf{U}_{h_{\mathbf{t}}} \mathbf{U}_{h_{\mathbf{t}}}^{\top}) \boldsymbol{\Sigma}_{\mathbf{t}}^{-1/2}\}$. Note that for easy of exposition, with a slight abuse of notation, we keep using the same notation $\mathbf{M}$, for $\mathbf{Z}$-scaled kernel matrices as the $\mathbf{X}$-scaled ones, which were previously discussed in Section 2. For any index set $\mathcal{F}$, we denote $\mathbf{X}^{\mathbf{t}}_{\mathcal{F}} = \{X^{\mathbf{t}}_i, i \in \mathcal{F}\}$, $\boldsymbol{\mu}_{\mathcal{F},\mathbf{t}} = \text{E}(\mathbf{X}^t_{\mathcal{F}})$, $\mathbf{U}_{\mathcal{F},h_{\mathbf{t}}} = \text{E}(\mathbf{X}^t_{\mathcal{F}} \mid \mathbf{Y}^t \in J^{\mathbf{t}}_{h_{\mathbf{t}}}) - \boldsymbol{\mu}_{\mathcal{F},\mathbf{t}}$ and $\boldsymbol{\Sigma}_{\mathcal{F},\mathbf{t}} = \text{Var}(\mathbf{X}^{\mathbf{t}}_{\mathcal{F}})$. Moreover, we define $\mathbf{M}_{\mathcal{F}}(\mathbf{t}) = \boldsymbol{\Sigma}_{\mathcal{F},\mathbf{t}}^{-1/2}(\sum_{h_{\mathbf{t}}=1}^{H_{\mathbf{t}}} p_{h_{\mathbf{t}}} \mathbf{U}_{\mathcal{F},h_{\mathbf{t}}} \mathbf{U}_{\mathcal{F},h_{\mathbf{t}}}^{\top}) \boldsymbol{\Sigma}_{\mathcal{F},\mathbf{t}}^{-1/2}$ and $\mathbf{M}_{\mathcal{F}} = \text{E}(\mathbf{M}_{\mathcal{F}}(\mathbf{t}))$, then we have the following proposition.

Proposition 3.1: *Suppose Condition* 2.1 *holds, then for any index set* $\mathcal{F}$ *such that* $\mathcal{A} \subseteq \mathcal{F} \subseteq \mathcal{I}$, *we have* $\text{tr}(\mathbf{M}_{\mathcal{A}}) = \text{tr}(\mathbf{M}_{\mathcal{F}}) = \text{tr}(\mathbf{M}_{\mathcal{I}})$.

Proposition 3.1 shows that $\text{tr}(\mathbf{M}_{\mathcal{F}})$ can be used to capture the strength of the relationship between Y and $\mathbf{X}$ given $\mathbf{W}$. If $\mathcal{A}$ is a subset of $\mathcal{F}$, then the kernel matrix $\mathbf{M}_{\mathcal{F}}$ has the same trace as $\mathbf{M}_{\mathcal{A}}$.

Denote $\mathcal{F} \cup j$ as the index set consisting of j and all the indices in $\mathcal{F}$. Suppose we already have the index set $\mathcal{F}$ selected in the model, and $\mathcal{F}$ does not contain $\mathcal{A}$, based on the following theorem, we can use the difference between $\text{tr}(\mathbf{M}_{\mathcal{F}\cup j})$ and $\text{tr}(\mathbf{M}_{\mathcal{F}})$ to measure the contribution of the additional X_j to Y given $(\mathbf{X}_{\mathcal{F}}, \mathbf{W})$.

Theorem 3.1: *For any* $\mathbf{t} \in \mathbf{R}_{\mathbf{T}}^q$, *suppose that we have*

$$\text{E}(X_j^{\mathbf{t}} \mid \mathbf{X}_{\mathcal{F}}^{\mathbf{t}}) \text{ is a linear function of } \mathbf{X}_{\mathcal{F}}^{\mathbf{t}}, \text{ for any } j \notin \mathcal{F} \text{ and } \mathcal{F} \subseteq \mathcal{I}. \tag{5}$$

Then

- *If* $\mathcal{A} \subseteq \mathcal{F}$, *then* $\text{tr}(\mathbf{M}_{\mathcal{F}\cup j}) - \text{tr}(\mathbf{M}_{\mathcal{F}}) = 0$.
- *If* $\mathcal{A} \not\subseteq \mathcal{F}$, *then* $\text{tr}(\mathbf{M}_{\mathcal{F}\cup j}) - \text{tr}(\mathbf{M}_{\mathcal{F}}) = \text{E}_{\mathbf{T}}(\sum_{h_{\mathbf{t}}=1}^{H_{\mathbf{t}}} p_{h_{\mathbf{t}}}(\boldsymbol{\gamma}_{j|\mathcal{F},h_{\mathbf{t}}}^{\mathbf{t}})^2)$, *where* $\mu_{j|\mathcal{F}}^{\mathbf{t}} = \text{E}(\boldsymbol{\gamma}_{j|\mathcal{F}}|\mathbf{T}=\mathbf{t})$ *and* $\boldsymbol{\gamma}_{j|\mathcal{F},h_{\mathbf{t}}}^{\mathbf{t}} = \text{E}(\boldsymbol{\gamma}_{j|\mathcal{F}}|Y \in J_{h_{\mathbf{t}}}, \mathbf{T}=\mathbf{t}) - \mu_{j|\mathcal{F}}^{\mathbf{t}}$ *with* $X_{j|\mathcal{F}} = \mathbf{X}_j - \text{E}(X_j|\mathbf{X}_{\mathcal{F}})$, $\sigma_{j|\mathcal{F}}^2 = \text{Var}(X_{j|\mathcal{F}})$, *and* $\boldsymbol{\gamma}_{j|\mathcal{F}} = X_{j|\mathcal{F}}/\sigma_{j|\mathcal{F}}$.

Condition 5 is parallel to Condition 2.1(a). When $\mathbf{X}^{\mathbf{t}}$ follows an elliptical contour distribution for any $\mathbf{t}$, both conditions are satisfied. The first part of Theorem 3.1 shows that the trace difference between $\mathbf{M}_{\mathcal{F}\cup j}$ and $\mathbf{M}_{\mathcal{F}}$ is 0, when the active set $\mathcal{A}$ is already included in the set $\mathcal{F}$. The second part provides a formula to calculate the trace difference, when the set $\mathcal{F}$ does not include all the active predictors.

For the derivation of the asymptotic consistency of our method, we hence assume that

$$\boldsymbol{\Sigma}_{\mathbf{t}} = \boldsymbol{\Sigma}, \quad \text{for any } \mathbf{t} \in \mathbf{R}_{\mathbf{T}}^q \tag{6}$$

Although simulation studies suggest that our method still performs reasonably well in applications where this 'homogeneous variance condition' does not hold.

Suppose that $d = \dim(\mathcal{S}_{Y|\mathbf{Z}}^{(\mathbf{W})}) = \dim(\mathcal{S}_{Y|\mathbf{X}}^{(\mathbf{W})})$, and let $\lambda_1 \geq, \ldots, \geq \lambda_d$ be the nonzero eigenvalues for $\mathbf{M}$ and $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_d$ be the corresponding eigenvectors. Denote $\boldsymbol{\beta}_i = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\eta}_i = (\beta_{i,1}, \ldots, \beta_{i,p})^\top$, for $i = 1, \ldots, d$. Under Condition 2.1, we have $\text{Span}\{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d\} = \mathcal{S}_{Y|\mathbf{X}}^{(\mathbf{W})}$. Furthermore, we define $\beta_{\min}^2 = \min_{j\in\mathcal{A}} \sum_{i=1}^d \beta_{i,j}^2$, where $\lambda_{\min}$ and $\lambda_{\max}$ are the smallest and the largest eigenvalues of $\boldsymbol{\Sigma}$, respectively.

Proposition 3.2: *Suppose that condition* 5 *in Theorem* 3.1 *holds, for any* $\mathcal{F}$ *which* $\mathcal{A} \not\subseteq \mathcal{F}$ *we have*

$$\max_{j\in\mathcal{A}\cap\mathcal{F}^c} \left(\text{tr}(\mathbf{M}_{\mathcal{F}\cup j}) - \text{tr}(\mathbf{M}_{\mathcal{F}})\right) \geq \lambda_d \lambda_{\max}^{-1} \lambda_{\min} \beta_{\min}^2. \tag{7}$$

Under the sufficient dimension reduction framework, we know $Y \perp\!\!\!\perp \mathbf{X}|(\boldsymbol{\beta}_1^T\mathbf{X}, \ldots, \boldsymbol{\beta}_d^T\mathbf{X}, \mathbf{W})$. Since $\mathcal{A}$ is the smallest active index set such that $Y \perp\!\!\!\perp \mathbf{X}|(\mathbf{X}_{\mathcal{A}}, \mathbf{W})$, then $\sum_{i=1}^d \beta_{i,j}^2 > 0$ for any $j \in \mathcal{A}$. Hence, for any $\mathcal{F}$ which does not include all the active predictors, the maximum difference between $\mathbf{M}_{\mathcal{F}\cup j}$ and $\mathbf{M}_{\mathcal{F}}$ over $j \in \mathcal{F}^c \cap \mathcal{A}$ is larger than 0 based on the result in Proposition 3.2.

Let $(\mathbf{X}_i, Y_i, \mathbf{W}_i)$, $i = 1, \ldots, n$ be simple random sample of size n. Follow Feng et al. (2013), for easy of implementation, we choose l_n different $\mathbf{t}$'s of which l_n is of order

$O(n)$ and use $n_{\mathbf{t}}$ to denote the subsample size for a given $\mathbf{t}$. Then, we can rewrite the sample as $(\mathbf{X}_i^{\mathbf{t}}, Y_i^{\mathbf{t}})$, $i = 1, \ldots, n_{\mathbf{t}}$ for a given $\mathbf{t}$. Let $\widehat{\mathbf{M}}(\mathbf{t})$ be the sample estimate of $\mathbf{M}(\mathbf{t})$, then we can estimate $\mathbf{M}$ using $\widehat{\mathbf{M}} = \frac{1}{l_n}\sum_{i=1}^{l_n}\widehat{\mathbf{M}}(\mathbf{t}_i)$. Regarding the choice of $\mathbf{t}$, theoretically speaking, we only need $l_n = O_n$ different values of $\mathbf{t}$ to obtain a $\sqrt{n}$ consistent estimator of $\mathbf{M}$. One easy way is to choose $l_n = n$, and $\mathbf{t}_i = \mathbf{W}_i$. However, when q, the dimension of $\mathbf{W}$ is large, for many $\mathbf{W}_i$, the set of contaminated points $\{(\mathbf{X}_i, Y_i)\}$ associated with the supercube $\{\mathbf{W}_j : I(\mathbf{W}_j \le \mathbf{W}_i)\}$ are very few and then the partial central subspace $\mathrm{Span}\{M(\mathbf{W}_i)\}$ cannot be estimated well. Hence, $\frac{1}{n}\sum_{i=1}^{n} M_n(\mathbf{W}_i)$ cannot provide a good estimator of the partial central subspace $\mathrm{Span}\{M\}$. To deal with this issue, we follow Feng et al. (2013) and use

$$\frac{1}{qn}\sum_{k=1}^{q}\sum_{i=1}^{n}\mathbf{M}_n(\mathbf{W}_{ik}^{\infty})$$

as an estimator for $\mathbf{M}$, where $\mathbf{W}_{ik}^{\infty}$ is the column vector of which only the kth component is the same as that of $\mathbf{W}_i$ and the other components are the maximum values of the corresponding components of all $\mathbf{W}_i$'s. Please refer to Feng et al. (2013) for further details.

Follow the SIR-based forward trace pursuit algorithm in Yu et al. (2016), the screening procedure starts with an empty index set $\mathcal{F}_0$, then at each step, the index which maximises the difference between the traces of successive kernel matrices to the working set is added, until we acquire a working index set with n indices. Hence, we obtain a sequence of n nested working index sets $\mathcal{F}_1, \ldots, \mathcal{F}_n$. In order to select a model from this sequence of nested working index sets, we use the modified BIC criterion defined in Yu et al. (2016):

$$\mathrm{BIC}(\mathcal{F}) = -\log\left\{\mathrm{tr}(\widehat{\mathbf{M}}_{\mathcal{F}})\right\} + n^{-1}|\mathcal{F}|(\log n + 2\log p),$$

where $|\mathcal{F}|$ denotes the cardinality of set $\mathcal{F}$.

To obtain the sure screening property of conditional forward trace pursuit based on SIR, we need the following conditions.

Condition 3.1: (a) There exist some constants $\alpha_0 > 0$ and $0 < b_0 < 1/2$ such that

$$\min_{\mathcal{F}:\, \mathcal{A}\not\subseteq\mathcal{F}}\ \max_{j\in\mathcal{A}\cap\mathcal{F}^c}\left(\mathrm{tr}(\mathbf{M}_{\mathcal{F}\cup j}) - \mathrm{tr}(\mathbf{M}_{\mathcal{F}})\right) \ge a_0 n^{-b_0}. \tag{8}$$

(b) $\mathbf{X}$ and $\mathbf{X}^{\mathbf{t}}$ follows multi-normal distributions for any $\mathbf{t} \in \mathbf{R}_{\mathbf{T}}^{q}$.
(c) There exist $c_1 > 0$ and $c_2 > 0$ such that $c_1 < \lambda_{\min} < \lambda_{\max} < c_2$.
(d) There exist constants a_1, b_1 and b_2 such that $\log(p) \le a_1 n^{\theta_1}$, $|\mathcal{A}| \le a_1 n^{b_2}$ and $2b_0 + b_1 + 2b_2 < 1$.
(e) There exists constant b_3 such that $l_n = O(n^{b_3})$ and $n_{\mathbf{t}_m} = O(n^{1-b_3})$ for any $\mathbf{t}_m$ among the l_n points where $0.5(1 - c_3) < b_3 < 1 - c_3$.

Motivated by the conclusion in Proposition 3.2, we assume that Condition 3.1(a) holds. Condition 3.1(b,c) are common for variable screening of high-dimensional data. Assuming Condition 3.1(b,c), Wang (2009) studied the sure screening property of forward linear regression. Condition 3.1(d) allows the dimension p and the number of important predictors to go to infinity as sample size n goes to infinity. We assume Condition 3.1(e) to

guarantee that it is not too sparse for each subsample and $\widehat{\mathbf{M}}(\mathbf{t}_m)$ is $\sqrt{n}$ consistent estimator of $\mathbf{M}(\mathbf{t}_m)$ for $m = 1, \ldots, l_n$.

Theorem 3.2: *Assume Condition* 2.1 *and Condition* 3.1 *hold, then we have*

$$\Pr(\mathcal{A} \subset \mathcal{F}_{\hat{m}}) \to 1,$$

as $n \to \infty$ *and* $p \to \infty$, *where* $\hat{m} = \operatorname{argmin}_{1 \le k \le n} BIC(\mathcal{F}_k)$.

Theorem 3.2 shows that our conditional forward trace pursuit method based on SIR has the desired sure screening property.

4. Numerical studies

4.1. Simulation studies

In this part, we compare the screening performance of our conditional forward trace pursuit (CFTP) method with CSIS Barut et al. (2016). Based on 100 repetitions, we evaluate the performance using the true model coverage rate (CR, the rate of all the significant predictors being selected), the average model size (MS), the average false positive rate (FP), and the average false-negative rate (FN). For CSIS, we use random decoupling, which was discussed in Barut et al. (2016), to select the threshold parameters and determine the model size for Model I-VI; while for Model VII-IX, $[n/\log(n)]$ is used as the model size since those provided by random decoupling method would be too small.

The following models are considered.

$$\begin{aligned}
\text{(I)}\quad & Y = 3W_1 + 3W_2 + 3W_3 + 3W_4 + 3W_5 - 7.5X_1 + \epsilon,\\
\text{(II)}\quad & Y = (3W_1 + 3W_2 + 3W_3 + 3W_4 + 3W_5 - 7.5X_1 + \epsilon)^2,\\
\text{(III)}\quad & Y = \exp(3W_1 + 3W_2 + 3W_3 + 3W_4 + 3W_5 - 7.5X_1) + \epsilon,\\
\text{(IV)}\quad & Y = 5W + 2X_p + \epsilon,\\
\text{(V)}\quad & Y = (5W + 2X_p)^2 + \epsilon,\\
\text{(VI)}\quad & Y = \exp(5W + 2X_p) + (5W + 2X_p)^3 + \epsilon,\\
\text{(VII)}\quad & Y = 8W_1 - 6W_2 + 5W_3 + (X_1 + X_p)^2 + \epsilon,\\
\text{(VIII)}\quad & Y = 2W_1 - 1.5W_2 + \exp(X_{p-1}) + 2X_p^4 + \epsilon,\\
\text{(IX)}\quad & Y = \operatorname{sign}(W_1 - W_2)\exp(X_1 + X_2 + X_{p-1} + X_p) + \epsilon.
\end{aligned}$$

We set the sample size $n = 400$ for all models. The random error ϵ follows a standard normal distribution $N(0, 1)$ and is independent with $\mathbf{W}$ and $\mathbf{X}$. For Model I, II and III, we generate $[\mathbf{W}^\top, \mathbf{X}^\top]^\top$ from $N(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma} = 0.5\mathbf{I_{q+p}} + \mathbf{0.5J_{q+p}}$, $q = 5$, $p + q = 2000$. We use $\mathbf{I_p}$ to denote the p-dimensional identity matrix, and $\mathbf{J}_p$ is the $p \times p$ square matrix of all ones. Model I was also studied in Barut et al. (2016) to show that the conditional screening can recover the hidden significant predictors since $\operatorname{Cov}(Y, X_1) = 0$ under the setting in this model. For Model IV, V and VI, $[\mathbf{W}, \mathbf{X}]$ are also generated from multivariate normal distribution with zero mean vector. In these three models, we set $q = 1, p + q = 2000$,

$X_1, \ldots, X_{p-1}$ and W are all correlated with each other with correlation coefficient of 0.8, while X_p is independent with all of them. Under this setting, we have $\text{Cov}(Y, X_i) = 4$ for $i = 1, \ldots, p-1$, and $\text{Cov}(Y, X_p) = 2$ for Model IV. Barut et al. (2016) discussed a similar model and show that conditional screening can reduce the false negative rate. In Model VII, W_i, $i = 1, 2, 3$, are independently generated from $U[0, 1]$, and $\mathbf{X}$ follows $N(\mathbf{0}, \mathbf{\Sigma})$ with elements $\sigma_{i,j} = \rho^{|i-j|}$ for $i,\ j = 1, \ldots, p$ and $p = 2000$. For Model VIII, $[\mathbf{W}^\top, \mathbf{X}^\top]^\top$ are generated from $N(\mathbf{0}, \mathbf{\Sigma})$, where $\sigma_{i,j} = \rho^{|i-j|}$, $q = 2, p + q = 2000$. In Model IX, $\text{Var}(\mathbf{X}|W)$ is dependent on $\mathbf{W}$, which violates the homogeneous variance assumption. Here, W_1 and W_2 are independently generated from $U[0, 1]$, and $\mathbf{X}$ is generated from $N(\mathbf{0}, \mathbf{\Sigma})$. As in Model VII, we set $\sigma_{i,j} = \rho^{|i-j|}$ for $i,\ j = 1, \ldots, p$ and $p = 2000$. However, in this model, we consider $\rho = \rho^\star$ which takes two different values depending on the difference between W_1 and W_2: $\rho^\star = 0$ if $W_1 - W_2 > 0$, and $\rho^\star = 0.5$ otherwise.

Table 1 compares the performance of our method with CSIS for Model I–III. As expected, the SAVE based method does not perform well as it could not deal with linear trends well Cook and Forzani (2009). However, both SIR and DR based conditional forward trace pursuit methods outperform CSIS: the true model coverage rates provided by our methods are 1, which means that our method can always include all the significant predictors; the false positive rate and false-negative rate are also much smaller than those of CSIS; the average model sizes are also much smaller than those of CSIS. For example, for Model III, CR and FN from CSIS are 0.18 and 0.82, respectively, comparing with 1 (the closer to one the better) and 0 (the smaller the better) from our method.

Table 2 gives simulation results for Model IV–VI. Still, CSIS is outperformed by our SIR and DR based methods. Our methods can provide screening results with much smaller model sizes, similar or better coverage rates, smaller false positive rates and/or smaller false negative rates for all three models. The nonlinear model structure does not affect the performance of our screening method, however it adversely affects the performance of CSIS greatly for Model V and VI. Results for Model VII and VIII are given on Table 3. Model VII has a quadratic structure in the mean function where SAVE is expected to perform well, which agrees with the simulation results. For Model VIII, DR based method dominates all the other methods. Simulation results for Model IX with different correlation structures are shown on Table 4. We discussed before, when $\rho = \rho*$, the homogeneous variance assumption is violated. As we can see, both SIR and DR based methods still outperform CSIS. Though DR based method does not perform as well as SIR based method

Table 1. Results for Model I, II and III.

Model	Method	CR	MS	FP	FN
I	CFTP-SIR	1	8.3	0.0037	0
	CFTP-SAVE	0	31.8	0.0159	1
	CFTP-DR	1	32.3	0.0157	0
	CSIS	1	859	0.4303	0
II	CFTP-SIR	1	13	0.0065	0
	CFTP-SAVE	0	30	0.0150	1
	CFTP-DR	1	33	0.0165	0
	CSIS	0.1	16.5	0.0082	0.9
III	CFTP-SIR	1	11.2	0.005	0
	CFTP-SAVE	0	32.3	0.016	1
	CFTP-DR	1	33.3	0.016	0
	CSIS	0.18	10.5	0.052	0.82

Table 2. Results for Model IV, V and VI.

Model	Method	CR	MS	FP	FN
IV	CFTP-SIR	1	9.75	0.0044	0
	CFTP-SAVE	0	27	0.0135	1
	CFTP-DR	0.97	28.6	0.0138	0.03
	CSIS	1	221	0.1106	0
V	CFTP-SIR	1	9.2	0.0041	0
	CFTP-SAVE	0.04	27.1	0.0135	0.96
	CFTP-DR	1	28.1	0.0136	0
	CSIS	0.01	223.94	0.1121	0.99
VI	CFTP-SIR	1	9.1	0.0041	0
	CFTP-SAVE	0	27.1	0.0135	1
	CFTP-DR	1	28.3	0.0137	0
	CSIS	0.13	209.05	0.1046	0.87

Table 3. Results for Model VII and VIII.

		$\rho = 0$				$\rho = 0.5$			
Model	Method	CR	MS	FP	FN	CR	MS	FP	FN
VII	CFTP-SIR	0	15.5	0.0078	1	0	15.15	0.0076	0.975
	CFTP-SAVE	1	30.6	0.0143	0	1	30.3	0.0142	0
	CFTP-DR	0.94	33.6	0.0159	0.060	1	33.6	0.0158	0
	CSIS	0	67	0.0333	0.885	0	67	0.0333	0.865
VIII	CFTP-SIR	0.03	11.9	0.0050	0.475	0.10	12.36	0.0056	0.450
	CFTP-SAVE	0.20	27.5	0.0132	0.400	0.08	27.3	0.0132	0.465
	CFTP-DR	1	32.4	0.0152	0	1	32.2	0.0151	0
	CSIS	0	67	0.0332	0.810	0	67	0.0332	0.790

Table 4. Results for Model IX.

ρ	Method	CR	MS	FP	FN
$\rho = \rho^\star$	CFTP-SIR	1	10.3	0.0032	0
	CFTP-SAVE	0	30.9	0.0155	1
	CFTP-DR	0.83	33.2	0.0148	0.075
	CSIS	0.21	67	0.0325	0.455

since the constant variance condition required for DR does not hold for this model. The false negative rates for SIR based method, DR based method, and CSIS are 0, 0.075, and 0.455 respectively; while the coverage rates for the three methods are 1, 0.83 and 0.21, respectively. CSIS mistakenly screens out some of the significant predictors frequently. All our simulation results suggest that DR based conditional forward trace pursuit method is the most robust screening method, while SIR based conditional forward trace pursuit conditional forward trace pursuit method provides the best screening performance for most of the time. We suggest to use SIR based screening method first, and use DR based method as a complement.

We also considered the following model where W is not related to the response Y.

$$\text{(X)}\ Y = 5 * \text{sign}(X_{2000})\exp(X_1) + \epsilon,$$

where $\mathbf{X} = (X_1, \ldots, X_{2000})$ follows multivariate standard normal distribution, W and ϵ are both univariate standard normal random variables. The sample size n is also set as 400. Simulation results are given on Table 5. As expected, SAVE based method did not perform

Table 5. Results for Model X.

Method	CR	MS	FP	FN
CFTP-SIR	0.95	7.46	5.51	0.05
CFTP-SAVE	0	26.34	26.33	1.99
CFTP-DR	0.92	25.72	23.81	0.09
CSIS	0.03	1.09	0.06	0.97

well as in Model I–III. The true model coverage rates provided by our conditional forward trace pursuit method based on SIR and DR are 0.95 and 0.92, respectively, comparing with 0.03 from CSIS. Also the false negative rates for SIR based method, DR based method, and CSIS are 0.05, 0.09, and 0.97, respectively, which agrees with the conclusion we draw from previous simulation studies.

4.2. Real data analysis

In this section, we consider the aforementioned leukaemia data set, which was first studied by Golub et al. (1999) and has become a benchmark in many gene expression studies. The dataset consists of 72 samples and gene expression level of 7129 genes in two types of acute leukaemia, acute lymphoblastic leukaemia (ALL) and acute myeloid leukaemia (AML). There are 38 (27 ALL and 11 AML) training samples and 34 (20 ALL and 14 AML) testing samples. Our goal is to select related genes and classify future patients to the two leukaemia types based on those genes.

We standardised the gene expression dataset by centring and scaling each array with mean 0 and standard deviation 1. The proposed conditional screening method and CSIS are performed based on the following three different choices of $\mathbf{W}$.

- $\mathbf{W}_1 = \{\text{X95735, D26156}\}$;
- $\mathbf{W}_2 = \{\text{X95735, M27783}\}$;
- $\mathbf{W}_3 = \{\text{X95735, MD88422}\}$.

The genes X95735 (Zyxin) and D26156 (Transcriptional activator hSNF2b) in $\mathbf{W}_1$ have empirically high correlations for the difference between patients with AML and ALL and were used in Barut et al. (2016). The genes X95735 and M27783 (ELA2 Elastatse 2, neutrophil) in $\mathbf{W}_2$ are the two top-ranked genes from marginal screening SIS. For $\mathbf{W}_3$, the genes X95735 and MD88422 (CYSTATIN A) were identified in Hong, Wang, and He (2016). To compare with CSIS, we first perform our conditional forward trace pursuit method to select genes based on the training samples given $\mathbf{W}_i$, $i = 1, 2, 3$, respectively. Next, we establish a classification rule through the logistic model based on the genes being selected and apply this rule to the testing samples. The results are shown on Table 6.

Conditioning on {X95735, D26156} ($\mathbf{W}_1$), we identified another gene Z32765 (GB DEF = CD36 gene exon 15) using SIR-based conditional trace pursuit method. Armesilla, Calvo, and Vega (1996) showed that Gene CD36 was associated with acute myeloid leukaemia. The classification rule based on these three genes can achieve 0/38 training error rate and 1/34 testing error rate.

Table 6. Results for Model V.

	$\mathbf{W}_1$		$\mathbf{W}_2$		$\mathbf{W}_3$	
Method	Train Err	Test Err	Train Err	Test Err	Train Err	Test Err
CSIS	0/38	2/34	1/38	5/34	0/38	2/34
CFTP-SIR	0/38	1/34	0/38	5/34	0/38	3/34
CFTP-SAVE	0/38	3/34	0/38	5/34	0/38	3/34
CFTP-DR	0/38	3/34	0/38	5/34	0/38	3/34

5. Conclusions

In this paper, we proposed a model-free conditional screening method to fully utilise the prior information regarding the importance of certain predictors. Comparing to CSIS developed by Barut et al. (2016), our method outperforms CSIS when the model structure is nonlinear and is comparable to CSIS for generalised linear model. Numerical studies suggest that our methods can provide screening results with much smaller model sizes, similar or better coverage rates, smaller false-positive rates, and/or smaller false-negative rates for nonlinear models.

Acknowledgements

The authors are grateful to the Associate Editor and two anonymous referees for their constructive comments and suggestions. Zhou Yu's research was partially supported by the National Scientific Foundation of China 11971170, and the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

Zhou Yu's research was partially supported by the National Scientific Foundation of China [grant number 11971170].

References

Armesilla, A.L., Calvo, D., and Vega, M.A. (1996), 'Structural and Functional Characterization of the Human CD36 Gene Promoter', *Journal of Biological Chemistry*, 271, 7781–7787.

Barut, E., Fan, J., and Verhasselt, A (2016), 'Conditional Sure Independence Screening', *Journal of the American Statistical Association*, 111, 1266–1277.

Chang, J., Tang, C., and Wu, Y. (2013), 'Marginal Empirical Likelihood and Sure Independence Feature Screening', *The Annals of Statistics*, 41, 1693–2262.

Chang, J., Tang, C., and Wu, Y. (2016), 'Local Independence Feature Screening for Nonparametric and Semiparametric Models by Marginal Empirical Likelihood', *The Annals of Statistics*, 44, 515–539.

Chiaromonte, F., Cook, R.D., and Li, B. (2002), 'Sufficient Dimension Reduction in Regressions with Categorical Predictors', *The Annals of Statistics*, 30, 475–497.

Chu, Y., and Lin, L. (2020), 'Conditional SIRS for Nonparametric and Semiparametric Models by Marginal Empirical Likelihood', *Statistical Papers*, 61, 1589–1606.

Cook, R.D. (1998), *Regression Graphics*, New York, NY: Wiley.

Cook, R.D. (2004), 'Testing Predictor Contributions in Sufficient Dimension Reduction', *The Annals of Statistics*, 32, 1062–1092.
Cook, R.D., and Forzani, B. (2009), 'Likelihood-Based Sufficient Dimension Reduction', *Journal of the American Statistical Association*, 104, 197–208.
Cook, R.D., and Nachtsheim, C. (1994), 'Reweighting to Achieve Elliptically Contoured Covariates in Regression', *Journal of the American Statistical Association*, 89, 592–599.
Cook, R.D., and Weisberg, S. (1991), 'Discussion of "sliced Inverse Regression for Dimension Reduction"', *Journal of the American Statistical Association*, 86, 328–332.
Cui, H., Li, R., and Zhong, W. (2014), 'Model-Free Feature Screening for Ultrahigh Dimensional Discriminant Analysis', *Journal of the American Statistical Association*, 110, 630–641.
Dong, Y., and Li, B. (2010), 'Dimension Reduction for Non-Elliptically Distributed Predictors: Second-Order Methods', *Biometrika*, 97, 279–294.
Fan, J., and Lv, J. (2008), 'Sure Independence Screening for Ultrahigh Dimensional Feature Space (with Discussion)', *Journal of the Royal Statistical: Society Series B*, 70, 849–911.
Fan, J., and Song, R. (2010), 'Sure Independence Screening in Generalized Linear Models with NP-dimensionality', *The Annals of Statistics*, 38, 3567–3604.
Fan, J., Feng, Y., and Song, R. (2011), 'Nonparametric Independence Screening in Sparse Ultrahigh-Dimensional Additive Models', *Journal of the American Statistical Association*, 106, 544–557.
Feng, Z., Wen, X., Yu, Z., and Zhu, L.-X. (2013), 'On Partial Sufficient Dimension Reduction With Applications to Partially Linear Multi-Index Models', *Journal of the American Statistical Association*, 108, 237–246.
Golub, T., Slonim, D., Tamyo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., and Caligiuri, M. (1999), 'Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring', *Science (New York, N.Y.)*, 286, 531–536.
Hall, P., and Li, K.C. (1993), 'On Almost Linearity of Low Dimensional Projection From High Dimensional Data', *The Annals of Statistics*, 21, 867–889.
He, X., Wang, L., and Hong, H. (2013), 'Quantile-Adaptive Model-Free Variable Screening for High-Dimensional Heterogeneous Data', *The Annals of Statistics*, 41, 342–369.
Hilafu, H., and Wu, W (2017), 'Partial Projective Resampling Method for Dimension Reduction: With Applications to Partially Linear Models', *Computational Statistics and Data Analysis*, 109, 1–14.
Hong, H.G., Wang, L., and He, X. (2016), 'A Data-Driven Approach to Conditional Screening of High-Dimensional Variables', *Stat*, 5, 200–212.
Jiang, B., and Liu, J.S. (2013), 'Sliced Inverse Regression With Variable Selection and Interaction Detection', Manuscript.
Li, K.C. (1991), 'Sliced Inverse Regression for Dimension Reduction (with Discussion)', *Journal of the American Statistical Association*, 86, 316–327.
Li, B., and Dong, Y. (2009), 'Dimension Reduction for Non-Elliptically Distributed Predictors', *The Annals of Statistics*, 37, 1272–1298.
Li, B., and Wang, S. (2007), 'On Directional Regression for Dimension Reduction', *Journal of the American Statistical Association*, 102, 997–1008.
Li, B., Kim, M.K., and Altman, N. (2010), 'On Dimension Folding of Matrix Or Array Valued Statistical Objects', *The Annals of Statistics*, 38, 1097–1121.
Lin, L., Sun, J., and Zhu, L.-X. (2013), 'Nonparametric Feature Screening', *Computational Statistics and Data Analysis*, 67, 162–174.
Lu, J., and Lin, L. (2020), 'Model-Free Conditional Screening Via Conditional Distance Correlation', *Statistical Papers*, 61, 225–244.
Mai, Q., and Zou, H. (2013), 'The Kolmogorov Filter for Variable Screening in High-Dimensional Binary Classification', *Biometrika*, 100, 229–234.
Mai, Q., and Zou, H. (2015), 'The Fused Kolmogorov Filter: A Nonparametric Model-Free Screening Method', *The Annals of Statistics*, 43, 1471–1497.
Niu, Y., Zhang, R., Liu, J., and Li, H. (2020), 'Group Screening for Ultra-High-Dimensional Feature Under Linear Model', *Statistical Theory and Related Fields* , 4, 43–54.

Pan, R., Wang, H., and Li, R. (2015), 'Ultrahigh Dimensional Multi-Class Linear Discriminant Analysis by Pairwise Sure Independence Screening', *Journal of the American Statistical Association.*, 111, 169–179.
Shao, Y., Cook, R.D., and Weisberg, S. (2007), 'Marginal Tests with Sliced Average Variance Estimation', *Biometrika*, 94, 285–296.
Wang, H. (2009), 'Forward Regression for Ultra-High Dimensional Variable Screening', *Journal of the American Statistical Association*, 104, 1512–1524.
Wang, H. (2012), 'Factor Profiled Sure Independence Screening', *Biometrika*, 99, 15–28.
Xue, L., and Zou, H. (2011), 'Sure Independence Screening and Compressed Random Sensing', *Biometrika*, 98, 371–380.
Yu, Z., Dong, Y., and Zhu, L.-X. (2016), 'Trace Pursuit: a General Framework for Model-Free Variable Selection', *Journal of the American Statistical Association*, 111, 813–821.
Zhao, S.D., and Li, Y. (2012), 'Sure Screening for Estimating Equations in Ultra-High Dimensions', Manuscript.
Zhong, W., Zhang, T., Zhu, M., and Liu, J.S. (2012), 'Correlation Pursuit: Forward Stepwise Variable Selection for Index Models', *Journal of Royal Statistical Society: Series B*, 74, 849–870.
Zhu, L., Li, L., Li, R., and Zhu, L.-X. (2011), 'Model-Free Feature Screening for Ultrahigh Dimensional Data', *Journal of American Statistical Association*, 106, 1464–1475.

Appendix

Proof of Proposition 3.1: For any given $\mathbf{t}$, we denote $\lambda_1^{\mathbf{t}} \geq \cdots \geq \lambda_{d_{\mathbf{t}}}^{\mathbf{t}}$ as the nonzero eigenvalues for $\mathbf{M}(\mathbf{t})$ and $\boldsymbol{\eta}_1(\mathbf{t}), \ldots, \boldsymbol{\eta}_{d_{\mathbf{t}}}(\mathbf{t})$ as the corresponding eigenvectors. Let $\boldsymbol{\beta}_i(\mathbf{t}) = \boldsymbol{\Sigma}_{\mathbf{t}}^{-1/2}\boldsymbol{\eta}_i(\mathbf{t}) = (\beta_{i,1}(\mathbf{t}), \ldots, \beta_{i,p}(\mathbf{t}))^\top$ for $i = 1, \ldots, d_{\mathbf{t}}$. Since $Y \perp\!\!\!\perp \mathbf{X}_{\mathcal{A}^c} | (\mathbf{X}_{\mathcal{A}}, W)$, then we have $\beta_{i,j}(\mathbf{t}) = 0$, for any $j \in \mathcal{A}^c$. Recall that $\mathcal{A} = \{1, \ldots, K\}$. Define $\boldsymbol{\beta}_{\mathcal{A},i}(\mathbf{t}) = (\beta_{i,1}, \ldots, \beta_{i,K})^\top$ and $\boldsymbol{\beta}_{\mathcal{A}^c,i}(\mathbf{t}) = (\beta_{i,K+1}, \ldots, \beta_{i,p})^\top$, then $\boldsymbol{\beta}_{i,\mathcal{A}^c}(\mathbf{t}) = \mathbf{0}$.

Note that $\mathbf{M}(\mathbf{t}) = \sum_{i=1}^{d_{\mathbf{t}}} \lambda_i^{\mathbf{t}} \boldsymbol{\eta}_i(\mathbf{t})\boldsymbol{\eta}_i(\mathbf{t})^\top = \boldsymbol{\Sigma}_{\mathbf{t}}^{1/2}(\sum_{i=1}^{d_{\mathbf{t}}} \lambda_i^{\mathbf{t}} \boldsymbol{\beta}_i(\mathbf{t})\boldsymbol{\beta}_i(\mathbf{t})^\top)\boldsymbol{\Sigma}_{\mathbf{t}}^{1/2}$, then we have

$$\operatorname{tr}(\mathbf{M}(\mathbf{t})) = \operatorname{tr}\left\{\boldsymbol{\Sigma}_{\mathbf{t}}\left(\sum_{i=1}^{d_{\mathbf{t}}} \lambda_i^{\mathbf{t}} \boldsymbol{\beta}_i(\mathbf{t})\boldsymbol{\beta}_i(\mathbf{t})^\top\right)\right\} = \operatorname{tr}\left\{\boldsymbol{\Sigma}_{\mathbf{t},\mathcal{A}}\left(\sum_{i=1}^{d_{\mathbf{t}}} \lambda_i^{\mathbf{t}} \boldsymbol{\beta}_{\mathcal{A},i}(\mathbf{t})\boldsymbol{\beta}_{\mathcal{A},i}(\mathbf{t})^\top\right)\right\}. \tag{A1}$$

Since $\mathbf{M}_{\mathcal{A}}(\mathbf{t}) = \operatorname{Var}\{\mathrm{E}(\mathbf{Z}_{\mathcal{A}}^{\mathbf{t}} | Y^{\mathbf{t}} \in J_{h_{\mathbf{t}}}^{\mathbf{t}})\} = \boldsymbol{\Sigma}_{\mathcal{A},\mathbf{t}}^{-1/2}\operatorname{Var}\{\mathrm{E}(\mathbf{X}_{\mathcal{A}}^{\mathbf{t}} | Y^{\mathbf{t}} \in J_{h_{\mathbf{t}}}^{\mathbf{t}})\}\boldsymbol{\Sigma}_{\mathcal{A},\mathbf{t}}^{-1/2}$, we have

$$\operatorname{tr}(\mathbf{M}_{\mathcal{A}}(\mathbf{t})) = \operatorname{tr}\left\{\boldsymbol{\Sigma}_{\mathcal{A},\mathbf{t}}^{-1}\operatorname{Var}\{\mathrm{E}(\mathbf{X}_{\mathcal{A}}^{\mathbf{t}} | Y^{\mathbf{t}} \in J_{h_{\mathbf{t}}}^{\mathbf{t}})\}\right\}. \tag{A2}$$

Note that

$$\begin{aligned}\operatorname{Var}\{\mathrm{E}(\mathbf{X}^{\mathbf{t}} | Y^{\mathbf{t}} \in J_{h_{\mathbf{t}}}^{\mathbf{t}})\} &= \boldsymbol{\Sigma}_{\mathbf{t}}^{1/2}\mathbf{M}(\mathbf{t})\boldsymbol{\Sigma}_{\mathbf{t}}^{1/2} = \boldsymbol{\Sigma}_{\mathbf{t}}\left(\sum_{i=1}^{d_{\mathbf{t}}} \lambda_i^{\mathbf{t}} \boldsymbol{\beta}_i(\mathbf{t})\boldsymbol{\beta}_i(\mathbf{t})^{\mathbf{t}}\right)\boldsymbol{\Sigma}_{\mathbf{t}} \\ &= \begin{pmatrix} \boldsymbol{\Sigma}_{\mathcal{A},\mathbf{t}} & \boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}^c,\mathbf{t}} \\ \boldsymbol{\Sigma}_{\mathcal{A}^c\mathcal{A},\mathbf{t}} & \boldsymbol{\Sigma}_{\mathcal{A}^c,\mathbf{t}} \end{pmatrix}\begin{pmatrix} \sum_{i=1}^{d_{\mathbf{t}}} \lambda_i^{\mathbf{t}} \boldsymbol{\beta}_{\mathcal{A},i}(\mathbf{t})\boldsymbol{\beta}_{\mathcal{A},i}(\mathbf{t})^\top & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} \boldsymbol{\Sigma}_{\mathcal{A},\mathbf{t}} & \boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}^c,\mathbf{t}} \\ \boldsymbol{\Sigma}_{\mathcal{A}^c\mathcal{A},\mathbf{t}} & \boldsymbol{\Sigma}_{\mathcal{A}^c,\mathbf{t}} \end{pmatrix}. \end{aligned} \tag{A3}$$

From A3, it is obvious that $\operatorname{Var}\{\mathrm{E}(\mathbf{X}_{\mathcal{A}}^{\mathbf{t}} | Y^{\mathbf{t}} \in J_{h_{\mathbf{t}}}^{\mathbf{t}})\} = \boldsymbol{\Sigma}_{\mathcal{A},\mathbf{t}}(\sum_{i=1}^{d_{\mathbf{t}}} \lambda_i^{\mathbf{t}} \boldsymbol{\beta}_{\mathcal{A},i}(\mathbf{t})\boldsymbol{\beta}_{\mathcal{A},i}(\mathbf{t})^{\mathbf{t}})\boldsymbol{\Sigma}_{\mathcal{A},\mathbf{t}}$. Combined with A1 and A2, we have $\operatorname{tr}(\mathbf{M}_{\mathcal{A}}(\mathbf{t})) = \operatorname{tr}(\mathbf{M}_{\mathcal{I}}(\mathbf{t}))$. Similarly, we have $\operatorname{tr}(\mathbf{M}_{\mathcal{A}}(\mathbf{t})) = \operatorname{tr}(\mathbf{M}_{\mathcal{F}}(\mathbf{t}))$ for any $\mathcal{F}$ such that $\mathcal{A} \subseteq \mathcal{F}$. Then the conclusion follows. ■

Proof of Theorem 3.1: From Proposition 3.1, we know that $\operatorname{tr}(\mathbf{M}_{\mathcal{A}}) = \operatorname{tr}(\mathbf{M}_{\mathcal{F}})$ for any $\mathcal{F}$ such that $\mathcal{A} \subseteq \mathcal{F}$. Then the first part of Theorem 3.1 follows.

For any fixed $\mathbf{t}$, if Condition 5 holds, we have $\mathrm{E}(x_j^{\mathbf{t}}|\mathbf{X}_{\mathcal{F}}^{\mathbf{t}}) = \mathrm{Cov}(\mathbf{X}_{\mathcal{F}}^{\mathbf{t}}, X_j^{\mathbf{t}})\boldsymbol{\Sigma}_{\mathcal{F},\mathbf{t}}^{-1}\mathbf{X}_{\mathcal{F}}^{\mathbf{t}}$. Let $|\mathcal{F}|$ denote as the cardinality of $\mathcal{F}$, then we construct two $(|\mathcal{F}|+1)\times(|\mathcal{F}|+1)$ matrices $\mathbf{A}_{\mathbf{t}}$ and $\mathbf{C}_{\mathbf{t}}$ as

$$\mathbf{A}_{\mathbf{t}} = \begin{pmatrix} \mathbf{I}_{|\mathcal{F}|} & \mathbf{0} \\ \mathrm{Cov}(\mathbf{X}_{\mathcal{F}}^{\mathbf{t}}, X_j^{\mathbf{t}})\boldsymbol{\Sigma}_{\mathcal{F},\mathbf{t}}^{-1} & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{C}_{\mathbf{t}} = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathcal{F},\mathbf{t}} & \mathbf{0} \\ \mathbf{0} & \sigma^2_{j|\mathcal{F},\mathbf{t}} \end{pmatrix},$$

where $\sigma^2_{j|\mathcal{F},\mathbf{t}} = \mathrm{Var}(X^{\mathbf{t}}_{j|\mathcal{F}})$ with $X^{\mathbf{t}}_{j|\mathcal{F}} = \mathbf{X}^{\mathbf{t}}_j - \mathrm{E}(X^{\mathbf{t}}_j|\mathbf{X}^{\mathbf{t}}_{\mathcal{F}})$. Then we have that

$$\mathbf{A}_{\mathbf{t}}\mathbf{X}^{\mathbf{t}}_{\mathcal{F}\cup j} = \begin{pmatrix} \mathbf{X}^{\mathbf{t}}_{\mathcal{F}} \\ X^{\mathbf{t}}_{j|\mathcal{F}} \end{pmatrix} \quad \text{and} \quad \mathbf{A}_{\mathbf{t}}\mathbf{U}_{\mathcal{F}\cup j,h_{\mathbf{t}}} = \begin{pmatrix} \mathbf{U}_{\mathcal{F},h_{\mathbf{t}}} \\ \mathrm{E}(X^{\mathbf{t}}_{j|\mathcal{F}}|\mathbf{Y}^t \in J^{\mathbf{t}}_{h_{\mathbf{t}}}) - \mathrm{E}(X^{\mathbf{t}}_{j|\mathcal{F}}) \end{pmatrix}.$$

From the definition of $X^{\mathbf{t}}_{j|\mathcal{F}}$, it is obvious that $\mathrm{Cov}(X^{\mathbf{t}}_{j|\mathcal{F}}, \mathbf{X}^{\mathbf{t}}_{\mathcal{F}}) = \mathbf{0}$. Then we have $\mathrm{Var}(\mathbf{A}_{\mathbf{t}}\mathbf{X}^{\mathbf{t}}_{\mathcal{F}\cup j}) = \mathbf{A}_{\mathbf{t}}\boldsymbol{\Sigma}_{\mathcal{F}\cup j,\mathbf{t}}\mathbf{A}_{\mathbf{t}}^{\top} = \mathbf{C}_{\mathbf{t}}$. Therefore, we have $\boldsymbol{\Sigma}^{-1}_{\mathcal{F}\cup j,\mathbf{t}} = \mathbf{A}_{\mathbf{t}}\mathbf{C}_{\mathbf{t}}^{-1}\mathbf{A}_{\mathbf{t}}^{\top}$. Then we can rewrite $\mathrm{tr}(\mathbf{M}_{\mathcal{F}\cup j}(\mathbf{t}))$ as

$$\begin{aligned}
\mathrm{tr}(\mathbf{M}_{\mathcal{F}\cup j}(\mathbf{t})) &= \mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1/2}_{\mathcal{F}\cup j,\mathbf{t}}\left(\sum_{h_{\mathbf{t}}=1}^{H_{\mathbf{t}}} p_{h_{\mathbf{t}}}\mathbf{U}_{\mathcal{F}\cup j,h_{\mathbf{t}}}\mathbf{U}^{\top}_{\mathcal{F}\cup j,h_{\mathbf{t}}}\right)\boldsymbol{\Sigma}^{-1/2}_{\mathcal{F}\cup j,\mathbf{t}}\right\} \\
&= \mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}_{\mathcal{F}\cup j,\mathbf{t}}\left(\sum_{h_{\mathbf{t}}=1}^{H_{\mathbf{t}}} p_{h_{\mathbf{t}}}\mathbf{U}_{\mathcal{F}\cup j,h_{\mathbf{t}}}\mathbf{U}^{\top}_{\mathcal{F}\cup j,h_{\mathbf{t}}}\right)\right\} \\
&= \mathrm{tr}\left\{\mathbf{C}_{\mathbf{t}}^{-1}\left(\sum_{h_{\mathbf{t}}=1}^{H_{\mathbf{t}}} p_{h_{\mathbf{t}}}(\mathbf{A}_{\mathbf{t}}\mathbf{U}_{\mathcal{F}\cup j,h_{\mathbf{t}}})(\mathbf{A}_{\mathbf{t}}\mathbf{U}^{\top}_{\mathcal{F}\cup j,h_{\mathbf{t}}})\right)\right\} \\
&= \mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}_{\mathcal{F},\mathbf{t}}\left(\sum_{h_{\mathbf{t}}=1}^{H_{\mathbf{t}}} p_{h_{\mathbf{t}}}\mathbf{U}_{\mathcal{F},h_{\mathbf{t}}}\mathbf{U}^{\top}_{\mathcal{F},h_{\mathbf{t}}}\right)\right\} + \sum_{h_{\mathbf{t}}=1}^{H_{\mathbf{t}}} p_{h_{\mathbf{t}}}(\boldsymbol{\gamma}^{\mathbf{t}}_{j|\mathcal{F},h_{\mathbf{t}}})^2.
\end{aligned}$$

Then we have $\mathrm{tr}(\mathbf{M}_{\mathcal{F}\cup j}) - \mathrm{tr}(\mathbf{M}_{\mathcal{F}}) = \mathrm{E}_{\mathbf{T}}\{\mathrm{tr}(\mathbf{M}_{\mathcal{F}\cup j}(t)) - \mathrm{tr}(\mathbf{M}_{\mathcal{F}}(\mathbf{t}))\} = \mathrm{E}_{\mathbf{T}}(\sum_{h_{\mathbf{t}}=1}^{H_{\mathbf{t}}} p_{h_{\mathbf{t}}}(\boldsymbol{\gamma}^{\mathbf{t}}_{j|\mathcal{F},h_{\mathbf{t}}})^2)$. ■

Proof of Proposition 3.2: Denote $\boldsymbol{\Sigma}_{\mathcal{F}_1\mathcal{F}_2,\mathbf{t}} = \mathrm{Cov}(\mathbf{X}^{\mathbf{t}}_{\mathcal{F}_1}, \mathbf{X}^{\mathbf{t}}_{\mathcal{F}_2})$ and $\boldsymbol{\Sigma}_{\mathcal{F}_1\mathcal{F}_2,\mathbf{t}} = \mathrm{Cov}(\mathbf{X}_{\mathcal{F}_1}, \mathbf{X}_{\mathcal{F}_2})$ for any $\mathcal{F}_1, \mathcal{F}_2 \subseteq \mathcal{I}$. Since we suppose $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{\mathbf{t}}$, then we have that $\boldsymbol{\Sigma}_{\mathcal{F}_1\mathcal{F}_2,\mathbf{t}} = \boldsymbol{\Sigma}_{\mathcal{F}_1\mathcal{F}_2,\mathbf{t}}$. For any $j \in \mathcal{F}^c \cap \mathcal{A}$, we have

$$\begin{aligned}
\sigma^2_{j|\mathcal{F}}\big(\mathrm{tr}(\mathbf{M}_{\mathcal{F}\cup j}) - \mathrm{tr}(\mathbf{M}_{\mathcal{F}})\big) &= \mathrm{E}_{\mathbf{T}}\{\mathrm{Var}(\mathrm{E}(X_{j|\mathcal{F}}|\mathbf{Y}))\} \\
&= \big(-\boldsymbol{\Sigma}_{\mathcal{F}j}\boldsymbol{\Sigma}^{-1}_{\mathcal{F}}, 1\big)\mathbf{P}\mathrm{E}_{\mathbf{T}}\{\mathrm{Var}(\mathrm{E}(\mathbf{X}^t|\mathbf{Y}^t \in J^{\mathbf{t}}_{h_{\mathbf{t}}}))\}\mathbf{P}^{\top}\big(-\boldsymbol{\Sigma}_{\mathcal{F}j}\boldsymbol{\Sigma}^{-1}_{\mathcal{F}}, 1\big)^{\top}. \qquad \text{(A4)}
\end{aligned}$$

For simplicity, we suppose the first $|\mathcal{F}|+1$ elements of $\mathbf{X}$ is $(\mathbf{X}_{\mathcal{F}}, X_j)^{\top}$, then $\mathbf{P}$ in A4 can be denoted as $\mathbf{P} = (\mathbf{I}_{|\mathcal{F}|+1}, \mathbf{0}_{(|\mathcal{F}|+1)(p-|\mathcal{F}|-1)})$. Since $\mathbf{M} = \mathrm{E}_{\mathbf{T}}\{\mathrm{Var}(\mathrm{E}(\mathbf{Z}^{\mathbf{t}}|Y^{\mathbf{t}} \in J^{\mathbf{t}}_{h_{\mathbf{t}}}))\} = \sum_{i=1}^{d}\lambda_i\boldsymbol{\eta}_i\boldsymbol{\eta}_i^{\top}$, and $\boldsymbol{\beta}_i = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\eta}_i$, then we have

$$\mathrm{E}_{\mathbf{T}}\{\mathrm{Var}(\mathrm{E}(\mathbf{X}^t|\mathbf{Y}^t \in J^{\mathbf{t}}_{h_{\mathbf{t}}}))\} = \boldsymbol{\Sigma}^{1/2}\left(\sum_{i=1}^{d}\lambda_i\boldsymbol{\eta}_i\boldsymbol{\eta}_i^{\top}\right)\boldsymbol{\Sigma}^{1/2} = \boldsymbol{\Sigma}\left(\sum_{i=1}^{d}\lambda_i\boldsymbol{\beta}_i\boldsymbol{\beta}_i^{\top}\right)\boldsymbol{\Sigma}. \qquad \text{(A5)}$$

It follows that

$$\begin{aligned}
\big(-\boldsymbol{\Sigma}_{\mathcal{F}j}\boldsymbol{\Sigma}^{-1}_{\mathcal{F}}, 1\big)\mathbf{P}\boldsymbol{\Sigma}\boldsymbol{\beta}_i &= \big(-\boldsymbol{\Sigma}_{\mathcal{F}j}\boldsymbol{\Sigma}^{-1}_{\mathcal{F}}, 1\big)\mathbf{P}\boldsymbol{\Sigma}_{(\mathcal{F}\cup j)\mathcal{I}}\boldsymbol{\beta}_i \\
&= \big(\boldsymbol{\Sigma}_{j\mathcal{I}} - \boldsymbol{\Sigma}_{j\mathcal{F}}\boldsymbol{\Sigma}^{-1}_{\mathcal{F}}\boldsymbol{\Sigma}_{\mathcal{F}\mathcal{I}}\big)\boldsymbol{\beta}_i. \qquad \text{(A6)}
\end{aligned}$$

Let $\boldsymbol{\beta}_{i,\mathcal{F}} = \{\beta_{i,j}, j \in \mathcal{F}\}$. Since $(\boldsymbol{\Sigma}_{j\mathcal{I}} - \boldsymbol{\Sigma}_{j\mathcal{F}}\boldsymbol{\Sigma}^{-1}_{\mathcal{F}}\boldsymbol{\Sigma}_{\mathcal{F}\mathcal{F}})\boldsymbol{\beta}_i = 0$ and $\boldsymbol{\beta}_{i,\mathcal{F}^c\cap\mathcal{I}^c} = 0$, we have

$$\big(-\boldsymbol{\Sigma}_{\mathcal{F}j}\boldsymbol{\Sigma}^{-1}_{\mathcal{F}}, 1\big)\mathbf{P}\boldsymbol{\Sigma}\boldsymbol{\beta}_i = \big(\boldsymbol{\Sigma}_{j\mathcal{F}^c} - \boldsymbol{\Sigma}_{j\mathcal{F}}\boldsymbol{\Sigma}^{-1}_{\mathcal{F}}\boldsymbol{\Sigma}_{\mathcal{F}\mathcal{F}^c}\big)\boldsymbol{\beta}_{i,\mathcal{F}^c}$$

$$= \left(\boldsymbol{\Sigma}_{j(\mathcal{F}^c\cap\mathcal{A})} - \boldsymbol{\Sigma}_{j\mathcal{F}}\boldsymbol{\Sigma}_{\mathcal{F}}^{-1}\boldsymbol{\Sigma}_{\mathcal{F}(\mathcal{F}^c\cap\mathcal{A})}\right)\boldsymbol{\beta}_{i,\mathcal{F}^c\cap\mathcal{A}}, \tag{A7}$$

for any $i = 1, \ldots, d$. From A4, A5 and A7, it follows that

$$\sigma_{j|\mathcal{F}}^2\left(\mathrm{tr}(\mathbf{M}_{\mathcal{F}\cup j}) - \mathrm{tr}(\mathbf{M}_{\mathcal{F}})\right) = \sum_{i=1}^{d}\lambda_i\{\left(\boldsymbol{\Sigma}_{j(\mathcal{F}^c\cap\mathcal{A})} - \boldsymbol{\Sigma}_{j\mathcal{F}}\boldsymbol{\Sigma}_{\mathcal{F}}^{-1}\boldsymbol{\Sigma}_{\mathcal{F}(\mathcal{F}^c\cap\mathcal{A})}\right)\boldsymbol{\beta}_{i,\mathcal{F}^c\cap\mathcal{A}}\}^2.$$

Note that

$$\sum_{j\in\mathcal{F}^c}\{\left(\boldsymbol{\Sigma}_{j(\mathcal{F}^c\cap\mathcal{A})} - \boldsymbol{\Sigma}_{j\mathcal{F}}\boldsymbol{\Sigma}_{\mathcal{F}}^{-1}\boldsymbol{\Sigma}_{\mathcal{F}(\mathcal{F}^c\cap\mathcal{A})}\right)\boldsymbol{\beta}_{i,\mathcal{F}^c\cap\mathcal{A}}\}^2$$
$$= \boldsymbol{\beta}_{i,\mathcal{F}^c\cap\mathcal{A}}^{\top}\left(\boldsymbol{\Sigma}_{(\mathcal{F}^c\cap\mathcal{A})} - \boldsymbol{\Sigma}_{(\mathcal{F}^c\cap\mathcal{A})\mathcal{F}}\boldsymbol{\Sigma}_{\mathcal{F}}^{-1}\boldsymbol{\Sigma}_{\mathcal{F}(\mathcal{F}^c\cap\mathcal{A})}\right)^2\boldsymbol{\beta}_{i,\mathcal{F}^c\cap\mathcal{A}}$$

and

$$\lambda_{\min}\left(\boldsymbol{\Sigma}_{(\mathcal{F}^c\cap\mathcal{A})} - \boldsymbol{\Sigma}_{(\mathcal{F}^c\cap\mathcal{A})\mathcal{F}}\boldsymbol{\Sigma}_{\mathcal{F}}^{-1}\boldsymbol{\Sigma}_{\mathcal{F}(\mathcal{F}^c\cap\mathcal{A})}\right)$$
$$= \lambda_{\max}^{-1}\{\left(\boldsymbol{\Sigma}_{(\mathcal{F}^c\cap\mathcal{A})} - \boldsymbol{\Sigma}_{(\mathcal{F}^c\cap\mathcal{A})\mathcal{F}}\boldsymbol{\Sigma}_{\mathcal{F}}^{-1}\boldsymbol{\Sigma}_{\mathcal{F}(\mathcal{F}^c\cap\mathcal{A})}\right)^{-1}\}$$
$$\geq \lambda_{\max}^{-1}(\boldsymbol{\Sigma}^{-1}) = \lambda_{\min}.$$

Then we have that

$$\max_{j\in\mathcal{F}^c\cap\mathcal{A}}\sigma_{j|\mathcal{F}}^2\left(\mathrm{tr}(\mathbf{M}_{\mathcal{F}\cup j}) - \mathrm{tr}(\mathbf{M}_{\mathcal{F}})\right) \geq |\mathcal{F}^c\cap\mathcal{A}|^{-1}\sum_{j\in\mathcal{F}^c\cap\mathcal{A}}[\sigma_{j|\mathcal{F}}^2\left(\mathrm{tr}(\mathbf{M}_{\mathcal{F}\cup j}) - \mathrm{tr}(\mathbf{M}_{\mathcal{F}})\right)]$$

$$= |\mathcal{F}^c\cap\mathcal{A}|^{-1}\sum_{i=1}^{d}\lambda_i\boldsymbol{\beta}_{i,\mathcal{F}^c\cap\mathcal{A}}^{\top}\left(\boldsymbol{\Sigma}_{(\mathcal{F}^c\cap\mathcal{A})} - \boldsymbol{\Sigma}_{(\mathcal{F}^c\cap\mathcal{A})\mathcal{F}}\boldsymbol{\Sigma}_{\mathcal{F}}^{-1}\boldsymbol{\Sigma}_{\mathcal{F}(\mathcal{F}^c\cap\mathcal{A})}\right)^2\boldsymbol{\beta}_{i,\mathcal{F}^c\cap\mathcal{A}}$$

$$\geq |\mathcal{F}^c\cap\mathcal{A}|^{-1}\sum_{i=1}^{d}\lambda_i\lambda_{\min}^2\left(\boldsymbol{\Sigma}_{(\mathcal{F}^c\cap\mathcal{A})} - \boldsymbol{\Sigma}_{(\mathcal{F}^c\cap\mathcal{A})\mathcal{F}}\boldsymbol{\Sigma}_{\mathcal{F}}^{-1}\boldsymbol{\Sigma}_{\mathcal{F}(\mathcal{F}^c\cap\mathcal{A})}\right)^2\boldsymbol{\beta}_{i,\mathcal{F}^c\cap\mathcal{A}}^{\top}\boldsymbol{\beta}_{i,\mathcal{F}^c\cap\mathcal{A}}$$

$$\geq \lambda_d\lambda_{\min}|\mathcal{F}^c\cap\mathcal{A}|^{-1}\sum_{i=1}^{d}\boldsymbol{\beta}_{i,\mathcal{F}^c\cap\mathcal{A}}^{\top}\boldsymbol{\beta}_{i,\mathcal{F}^c\cap\mathcal{A}} \geq \lambda_d\lambda_{\max}^{-1}\lambda_{\min}\beta_{\min}^2. \qquad \blacksquare$$

To prove Theorem 3.2, we need the following lemmas.

Lemma A.1: *Let* $\widetilde{\mathbf{M}} = 1/l_n\sum_{m=1}^{l_n}\mathbf{M}(\mathbf{t}_m)$, $\psi_{h_{\mathbf{t}}} = p_{h_{\mathbf{t}}}^{-1/2}(I(Y^{\mathbf{t}}\in J_{h_{\mathbf{t}}}^{\mathbf{t}}))$ *and* $\zeta_{h_{\mathbf{t}}} = \boldsymbol{\Sigma}_{\mathbf{t}}^{-1}\mathrm{E}(\mathbf{X}^{\mathbf{t}}\psi_{h_{\mathbf{t}}})$, *then we have* $\mathrm{tr}(\widetilde{\mathbf{M}}) = (H-1) - \frac{1}{l_n}\sum_{m=1}^{l_n}\sum_{h_{\mathbf{t}_m}=1}^{H_{\mathbf{t}_m}}\mathrm{E}(\psi_{h_{\mathbf{t}_m}} - \zeta_{h_{\mathbf{t}_m}}^{\top}\mathbf{X}^{\mathbf{t}_m})^2$, *where* $H = 1/l_n\sum_{m=1}^{l_n}H_{\mathbf{t}_m}$.

Proof of Lemma A.1: For any $\mathbf{t}_m$, $m = 1, \ldots, l_n$ and $h_{\mathbf{t}_m}, h_{\mathbf{t}_m} = 1, \ldots, H_{\mathbf{t}_m}$,

$$\mathrm{E}(\psi_{h_{\mathbf{t}}} - \zeta_{h_{\mathbf{t}_m}}^{\top}\mathbf{X}^{\mathbf{t}_m})^2 = \mathrm{E}(\psi_{h_{\mathbf{t}}}^2) - 2\mathrm{E}(\psi_{h_{\mathbf{t}_m}}\zeta_{h_{\mathbf{t}_m}}^{\top}\mathbf{X}^{\mathbf{t}_m}) + \mathrm{E}((\zeta_{h_{\mathbf{t}_m}}^{\top}\mathbf{X}^{\mathbf{t}_m}\mathbf{X}^{\mathbf{t}_m\top}\zeta_{h_{\mathbf{t}_m}})$$
$$= \mathrm{E}(\psi_{h_{\mathbf{t}}}^2) - \zeta_{h_{\mathbf{t}_m}}^{\top}\boldsymbol{\Sigma}_{\mathbf{t}}\zeta_{h_{\mathbf{t}_m}} = (1-p_{h_{\mathbf{t}}}) - p_{h_{\mathbf{t}}}^{-1}\mathrm{E}\{\mathbf{Z}^{\mathbf{t}_m\top}I(Y^{\mathbf{t}}\in J_{h_{\mathbf{t}}}^{\mathbf{t}})\}\mathrm{E}\{\mathbf{Z}^{\mathbf{t}_m}I(Y^{\mathbf{t}}\in J_{h_{\mathbf{t}}}^{\mathbf{t}})\}.$$

Then we have that

$$\mathrm{tr}(\widetilde{\mathbf{M}}) = 1/l_n\sum_{m=1}^{l_n}\mathrm{tr}(\mathbf{M}(\mathbf{t}_m)) = 1/l_n\sum_{m=1}^{l_n}\sum_{h_{\mathbf{t}_m}=1}^{H_{\mathbf{t}_m}}p_{h_{\mathbf{t}}}^{-1}\mathrm{E}\{\mathbf{Z}^{\mathbf{t}_m\top}I(Y^{\mathbf{t}}\in J_{h_{\mathbf{t}}}^{\mathbf{t}})\}\mathrm{E}\{\mathbf{Z}^{\mathbf{t}_m}I(Y^{\mathbf{t}}\in J_{h_{\mathbf{t}}}^{\mathbf{t}})\}$$
$$= (H-1) - \frac{1}{l_n}\sum_{m=1}^{l_n}\sum_{h_{\mathbf{t}_m}=1}^{H_{\mathbf{t}_m}}\mathrm{E}(\psi_{h_{\mathbf{t}_m}} - \zeta_{h_{\mathbf{t}_m}}^{\top}\mathbf{X}^{\mathbf{t}_m})^2. \qquad \blacksquare$$

Lemma A.2: *Let* $D_{\mathcal{F}\cup j} = \mathrm{tr}(\mathbf{M}_{\mathcal{F}\cup j}) - \mathrm{tr}(\mathbf{M}_{\mathcal{F}})$ *and* $\widehat{D}_{\mathcal{F}\cup j} = \mathrm{tr}(\widehat{\mathbf{M}}_{\mathcal{F}\cup j}) - \mathrm{tr}(\widehat{\mathbf{M}}_{\mathcal{F}})$. *Suppose* $|\mathcal{F}| = O(n^{b_0+b_2})$ *and Condition* 3.1 *holds, there exists some constant* d_0 *such that* $\widehat{D}_{\mathcal{F}\cup j} - D_{\mathcal{F}\cup j} \le d_0|\mathcal{F}|\sqrt{\log p/n^{1-b_3}}$ *with probability tending to* 1.

Proof of Lemma A.2: Define $\widetilde{D}_{F\cup j} = \mathrm{tr}(\widetilde{\mathbf{M}}_{\mathcal{F}\cup j}) - \mathrm{tr}(\widetilde{\mathbf{M}}_{\mathcal{F}})$, then we have that

$$\widehat{D}_{\mathcal{F}\cup j} - D_{\mathcal{F}\cup j} = [\widehat{D}_{\mathcal{F}\cup j} - \widetilde{D}_{F\cup j}] + [\widetilde{D}_{F\cup j} - D_{\mathcal{F}\cup j}]. \tag{A8}$$

Form Lemma 7 in Yu et al. (2016), we know that $\widehat{\mathrm{Var}}\{\mathrm{E}(X_{j|\mathcal{F}}^{\mathbf{t}_m})\} - \mathrm{Var}\{\mathrm{E}(X_{j|\mathcal{F}}^{\mathbf{t}_m})\} = O(|\mathcal{F}|\sqrt{\log p/n^{1-b_3}})$ for any given $\mathbf{t}_m$, $m = 1, \ldots, l_n$. Furthermore, from the proof of Lemma 3 in Jiang and Liu (2013), we have $\hat{\sigma}^2_{j|\mathcal{F},\mathbf{t}} - \sigma^2_{j|\mathcal{F},\mathbf{t}} = O(|\mathcal{F}|\sqrt{\log p/n^{1-b_3}})$. Then we have that

$$\begin{aligned}
&\big\{\mathrm{tr}(\widehat{\mathbf{M}}_{\mathcal{F}\cup j}(\mathbf{t}_m)) - \mathrm{tr}(\widehat{\mathbf{M}}_{\mathcal{F}}(\mathbf{t}_m))\big\} - \big\{\mathrm{tr}(\widetilde{\mathbf{M}}_{\mathcal{F}\cup j}(\mathbf{t}_m)) - \mathrm{tr}(\widetilde{\mathbf{M}}_{\mathcal{F}}(\mathbf{t}_m))\big\} \\
&\quad = \hat{\sigma}^2_{j|\mathcal{F},\mathbf{t}}\widehat{\mathrm{Var}}\{\mathrm{E}(X_{j|\mathcal{F}}^{\mathbf{t}_m})\} - \sigma^2_{j|\mathcal{F},\mathbf{t}}\mathrm{Var}\{\mathrm{E}(X_{j|\mathcal{F}}^{\mathbf{t}_m})\} \\
&\quad = \big\{\hat{\sigma}^2_{j|\mathcal{F},\mathbf{t}}\widehat{\mathrm{Var}}\{\mathrm{E}(X_{j|\mathcal{F}}^{\mathbf{t}_m})\} - \hat{\sigma}^2_{j|\mathcal{F},\mathbf{t}}\mathrm{Var}\{\mathrm{E}(X_{j|\mathcal{F}}^{\mathbf{t}_m})\}\big\} \\
&\qquad - \big\{\hat{\sigma}^2_{j|\mathcal{F},\mathbf{t}}\mathrm{Var}\{\mathrm{E}(X_{j|\mathcal{F}}^{\mathbf{t}_m})\} - \sigma^2_{j|\mathcal{F},\mathbf{t}}\mathrm{Var}\{\mathrm{E}(X_{j|\mathcal{F}}^{\mathbf{t}_m})\}\big\} \\
&\quad = O\big(|\mathcal{F}|\sqrt{\log p/n^{1-b_3}}\big) + O\big(|\mathcal{F}|\sqrt{\log p/n^{1-b_3}}\big) = O\big(|\mathcal{F}|\sqrt{\log p/n^{1-b_3}}\big).
\end{aligned}$$

Hence, we have that

$$\begin{aligned}
&\widehat{D}_{\mathcal{F}\cup j} - \widetilde{D}_{F\cup j} \\
&\quad = \frac{1}{l_n}\sum_{m=1}^{l_n}\big[\big\{\mathrm{tr}(\widehat{\mathbf{M}}_{\mathcal{F}\cup j}(\mathbf{t}_m)) - \mathrm{tr}(\widehat{\mathbf{M}}_{\mathcal{F}}(\mathbf{t}_m))\big\} - \big\{\mathrm{tr}(\widetilde{\mathbf{M}}_{\mathcal{F}\cup j}(\mathbf{t}_m)) - \mathrm{tr}(\widetilde{\mathbf{M}}_{\mathcal{F}}(\mathbf{t}_m))\big\}\big] \\
&\quad = \frac{1}{l_n}\sum_{m=1}^{l_n} O\big(|\mathcal{F}|\sqrt{\log p/n^{1-b_3}}\big) = O\big(|\mathcal{F}|\sqrt{\log p/n^{1-b_3}}\big).
\end{aligned}$$

From this, it is obvious that there exists some constant d_0 such that $\widehat{D}_{\mathcal{F}\cup j} - D_{\mathcal{F}\cup j} \le d_0|\mathcal{F}|\sqrt{\log p/n^{1-b_3}}$ with probability tending to 1. ■

Proof of Theorem 3.2: Firstly, we prove that CFTP method can select in all $|\mathcal{A}|$ important predictors within $[2Ha_0^{-1}a_1 n^{b_0+b_2}]$ steps by showing that at least one important predictor in the model within $[2Ha_0^{-1}n^{b_0}]$ steps since $|\mathcal{A}| \le a_1 n^{b_2}$ under Condition 3.1. Without loss of generality, we just show that at least one important is selected in the model within the first $[2Ha_0^{-1}n^{b_0}]$.

Recall that $\mathcal{F}_k$ is the index set after kth step, we let $Q(k) = \mathrm{tr}(\widehat{\mathbf{M}}_{\mathcal{F}_k}) - \mathrm{tr}(\widehat{\mathbf{M}}_{\mathcal{F}_{k-1}})$. We assume that no important is selected in the model within the first $[2Ha_0^{-1}n^{b_0}]$ steps. From lemma A.2 and Condition 3.1, we have that

$$\begin{aligned}
Q(k) &\ge 2^{-1}\big(\mathrm{tr}(\mathbf{M}_{\mathcal{F}_k}) - \mathrm{tr}(\mathbf{M}_{\mathcal{F}_{k-1}}) - d_0|\mathcal{F}_k|\sqrt{\log p/n^{1-b_3}}\big) \\
&\ge 2^{-1}\big(a_0 n^{-b_0} - d_0 2Ha_0^{-1}a_1 n^{b_0+b_2}\sqrt{\log p/n^{1-b_3}}\big) \to 2^{-1}a_0 n^{-b_0}
\end{aligned}$$

if $\mathcal{F}_k \cap \mathcal{A} = \varnothing$ for any $k = 1, \ldots, [2Ha_0^{-1}n^{b_0}]$.
Hence, we have that

$$\sum_{k=1}^{[2Ha_0^{-1}n^{b_0}]} Q(k) \ge [2Ha_0^{-1}n^{b_0}] \times 2^{-1}a_0 n^{-b_0} \ge H.$$

However, from Lemma A.1, we know

$$\sum_{k=1}^{[2Ha_0^{-1}n^{b_0}]} Q(k) = \mathrm{tr}(\widehat{\mathbf{M}}_{\mathcal{F}_{[2Ha_0^{-1}n^{b_0}]}}) \leq H - 1.$$

Therefore, this implies at least one important predictor is selected in the model within the first $[2Ha_0^{-1}n^{b_0}]$ steps.Moreover, follow the proof of Theorem 5.2 in Yu et al. (2016) and the proof of Theorem 2 in Wang (2009), it is easy to prove that $\Pr(\mathcal{A} \subset \mathcal{F}_{\hat{m}}) \to 1$, and the details are omitted. ■