Submitted to the Annals of Statistics
arXiv: arXiv:0000.0000

SLICED INVERSE REGRESSION WITH VARIABLE SELECTION AND INTERACTION DETECTION

By Bo Jiang and Jun S. Liu

Harvard University

Variable selection methods play important roles in modeling high dimensional data and are keys to data-driven scientific discoveries. In this paper, we consider the problem of variable selection with interaction detection under the sliced inverse index modeling framework, in which the response is influenced by predictors through an unknown function of both linear combinations of predictors and interactions among them. Instead of building a predictive model of the response given combinations of predictors, we start by modeling the conditional distribution of predictors given responses. This inverse modeling perspective motivates us to propose a stepwise procedure based on likelihood-ratio tests that is effective and computationally efficient in detecting interaction with little assumptions on its parametric form. The proposed procedure is able to detect pairwise interactions among p predictors with a computational time of $O(p)$ instead of $O(p^2)$ under moderate conditions. Consistency of the procedure in variable selection under a diverging number of predictors and sample size is established. Its excellent empirical performance in comparison with some existing methods is demonstrated through simulation studies as well as real data examples.

1. Introduction. Recently there has been a significant surge of interest in analytically accurate, numerically robust, and algorithmically efficient variable selection methods, largely due to the tremendous advance in data collection techniques such as those in genetics, internet, and marketing. The importance of discovering truly influential factors from a large pool of possibilities is now widely recognized by both general scientists and quantitative modelers. Under linear regression models, various regularization methods have been proposed for simultaneously estimating regression coefficients and selecting predictors. Many promising algorithms, such as Lasso (Tibshirani, 1996; Zou, 2006; Friedman et al., 2007), LARS (Efron et al., 2004) and smoothly clipped absolute deviation (SCAD; Fan and Li (2001)), have been invented. When the number of the predictors is extremely large, Fan and Lv (2008) have proposed a sure independence screening (SIS) framework

AMS 2000 subject classifications: Primary 62J02; secondary 62H25, 62P10
Keywords and phrases: Sliced inverse regression, variable selection, sure independence screening

that first independently selects variables based on their correlations with the response and then applies variable selection methods in the second step.

1.1. *SIR with Variable Selection.* When the relationship between the response Y and predictors $\mathbf{X} = (X_1, X_2, \ldots, X_p)^T$ is beyond linear, the performance of the variable selection methods based on the linear model assumption can be severely compromised. In his seminal paper on dimension reduction, Li (1991) proposed a semi-parametric index model of the form

$$Y = f(\boldsymbol{\beta}_1^T \mathbf{X}, \boldsymbol{\beta}_2^T \mathbf{X}, \ldots, \boldsymbol{\beta}_q^T \mathbf{X}, \epsilon), \tag{1.1}$$

where f is an unknown link function and ϵ is a stochastic error independent of $\mathbf{X}$. A sliced inverse regression (SIR) method was developed by Li (1991) to estimate the so-called sufficient dimension reduction (SDR) directions $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_q$. For the ease of description, we standardize the predictors, $\mathbf{Z} = [\mathrm{Cov}(\mathbf{X})]^{-\frac{1}{2}} [\mathbf{X} - \mathbb{E}(\mathbf{X})]$, and rewrite model (1.1) as

$$Y = f\left(\boldsymbol{\eta}_1^T \mathbf{Z}, \boldsymbol{\eta}_2^T \mathbf{Z}, \ldots, \boldsymbol{\eta}_q^T \mathbf{Z}, \epsilon\right) \text{ with } \boldsymbol{\eta}_i = [\mathrm{Cov}(\mathbf{X})]^{\frac{1}{2}} \boldsymbol{\beta}_i. \tag{1.2}$$

The SIR algorithm was motivated by the following observation: under model (1.2), if $\mathbb{E}\left(\mathbf{b}^T \mathbf{Z} | \boldsymbol{\eta}_1^T \mathbf{Z}, \ldots, \boldsymbol{\eta}_q^T \mathbf{Z}\right)$ is linear in $\boldsymbol{\eta}_1^T \mathbf{Z}, \ldots, \boldsymbol{\eta}_q^T \mathbf{Z}$ for any vector $\mathbf{b} \in \mathbb{R}^p$, then $\mathbb{E}(\mathbf{Z}|Y)$ is contained in the linear subspace spanned by $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_q$ (Li, 1991). So if the first q largest eigenvalues of $\mathrm{Cov}(\mathbb{E}(\mathbf{Z}|Y))$ are all positive, SDR directions can be obtained by the corresponding eigenvectors.

Eigenvalues of $\mathrm{Cov}(\mathbb{E}(\mathbf{Z}|Y))$ also connects SIR with multiple linear regression (MLR). In MLR, the correlation squared, R^2, can be defined as

$$R^2 = \max_{\mathbf{b} \in \mathbb{R}^p} \left[\mathrm{Corr}\left(Y, \mathbf{b}^T \mathbf{Z}\right)\right]^2,$$

while in SIR, the largest eigenvalue of $\mathrm{Cov}(\mathbb{E}(\mathbf{Z}|Y))$, called the first *profile-*R^2, can be defined as

$$\lambda_1\left(\mathrm{Cov}\left(\mathbb{E}(\mathbf{Z}|Y)\right)\right) = \max_{\mathbf{b} \in \mathbb{R}^p} \max_{T} \left[\mathrm{Corr}\left(T(Y), \mathbf{b}^T \mathbf{Z}\right)\right]^2,$$

where maximum is taken over all bounded transformations $T(\cdot)$ and vectors $\mathbf{b} \in \mathbb{R}^p$ (Chen and Li, 1998). We can further define the kth profile-R^2, λ_k $(2 \le k \le q)$, as the kth largest eigenvalue of $\mathrm{Cov}(\mathbb{E}(\mathbf{Z}|Y))$ by restricting the vector $\mathbf{b}$ to be orthogonal to eigenvectors of the first $(k-1)$ profile-R^2.

Since the estimation of SDR directions does not automatically lead to variable selection, various methods have been developed to perform dimension reduction and variable selection simultaneously in the nonlinear setting.

For example, Li, Cook and Nachtsheim (2005) used a backward subset selection method based on χ^2-tests derived in Cook (2004), and Li (2007) developed sparse SIR (SSIR) algorithm to obtain shrinkage estimates of the SDR directions under L_1 norm. Motived by the F-test in stepwise regression and the connection between SIR and MLR, Zhong et al. (2012) proposed a stepwise variable selection procedure called correlation pursuit (COP) for index models under the SIR framework.

1.2. *Interaction Detection via Inverse Modeling.* The number of variables to be considered can be even larger with the inclusion of interaction terms between predictors. Consider the following simple example with a regression model for a univariate response variable Y and p independent normally distributed predictor variables $\mathbf{X} = (X_1, X_2, \ldots, X_p)^T$:

$$Y = X_1 X_2 + \epsilon, \tag{1.3}$$

where $\mathbf{X} \sim \mathrm{MVN}_p(\mathbf{0}, \mathbf{I}_p)$ and $\epsilon \sim N(0, 0.1)$. Since there are $\binom{p}{2}$ pairwise interactions, fitting regression models with 2-way interactions using variable selection methods, or even the sure independence screening procedure, is challenging when one has a moderate number of predictor variables, say $p = 1000$. Recently, there has been considerable effort in fitting interaction models in the statistical literatures. For example, Bien, Taylor and Tibshirani (2012) developed hierNet, an extension of Lasso to consider interactions in a model if one or both variables are marginally important (referred to as hierarchical interactions by the authors). Li, Zhong and Zhu (2012) proposed a sure independence screening procedure based on distance correlation (DC-SIS) that is shown to be capable of detecting important variables when interactions are presented.

Most of the aforementioned methods are derived from a *forward* modeling perspective, that is, a model for the conditional distribution of Y given $\mathbf{X}$. When predictor variables $\mathbf{X}$ can be treated as random, we obtain a different modeling perspective by "flipping" the roles of $\mathbf{X}$ and Y and putting the response Y behind the (conditioning) bar, which we call an *inverse* model. Indeed, this inverse modeling perspective has been taken by several researchers and has led to new developments in dimension reduction and variable selection methods. Cook (2007) proposed inverse regression models for dimension reduction, which have deep connections with the SIR method. Simon and Tibshirani (2012) proposed a permutation-based method for testing interactions by exploring the connection between the forward logistic model and the inverse normal mixture model when the response Y is binary. Another classical method derived from the inverse model perspective is the Naïve Bayes

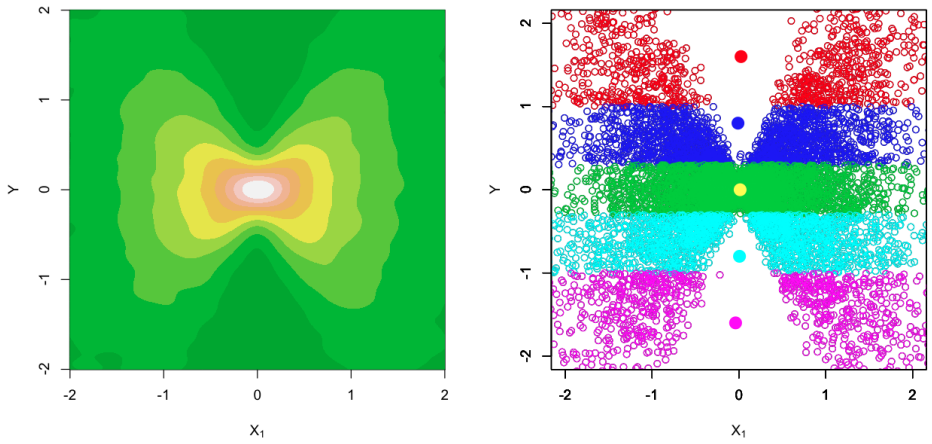

FIG 1. *Left panel: contour plot for the joint distribution of Y and X_1 in example (1.3). Right panel: conditional means and variances of X_1 given slices of Y. Slices are indicated by different colors and round dots marks conditional means of X_1 across slices. The conditional variances of X_1 within different slices (from top to bottom) are:* 2.29, 0.92, 0.41, 0.98 *and* 2.33

classifier for classifications with high dimensional features. Although Naïve Bayes classifier is limited by its strong independence assumption, it can be generalized by modeling the joint distribution of features. Murphy, Dean and Raftery (2010) proposed a variable selection method using Bayesian information criterion (BIC) for model-based discriminant analysis. Zhang and Liu (2007) proposed a Bayesian method called BEAM to detect epistatic interactions in genome-wide case-control studies, where Y is binary and $\mathbf{X}$ are discrete.

The inverse modeling perspective can also shed lights on interaction detections. Figure 1 shows the contour plot for the joint distribution of Y and X_1 from example (1.3). If we divide the response into five slices and calculate the mean and variance of X_1 within each slice (as shown in Figure 1), we can see that although X_1 is marginally uncorrelated with Y and the conditional means of X_1 are the same, the conditional variances of X_1 across slices are very different. As an illustrative example, we generated 200 observations from example (1.3) and divided the range of response into 5 slices with 40 observatoins in each slice. First, we calcualted the following test static:

$$n\widehat{D}_j^* = n\log\widehat{\sigma}_j^2 - \sum_{h=1}^{H} n_h \log\left[\widehat{\sigma}_j^{(h)}\right]^2, j = 1, 2, \ldots, p,$$

where $p = 1000$, $H = 5$, $n = 200$ and $n_h = 40$, $\left[\widehat{\sigma}_j^{(h)}\right]^2$ is the estimated variance of X_j in the hth slice, and $\widehat{\sigma}_j^2$ is the estimated variance of X_j using all the observations. We found that $\left(n\widehat{D}_1^*\right) = 62.48$ and $\left(n\widehat{D}_2^*\right) = 56.03$ are significant compared with the null distribution of irrelevant predictors, which is asymptotically $\chi^2(8)$ (we have $\max_{j\in\{3,4,\ldots,1000\}}\left(n\widehat{D}_j^*\right) = 28.46$ in this particular example). Even if we had not been able to select X_2 in the first round using marginal information, we can recover it via a conditional test given X_1 as follows:

$$n\widehat{D}_{j|\{1\}}^* = n\log\widehat{\sigma}_{j|\{1\}}^2 - \sum_{h=1}^{H} n_h \log\left[\widehat{\sigma}_{j|\{1\}}^{(h)}\right]^2, j = 2, 3, \ldots, p,$$

where $\left[\widehat{\sigma}_{j|\{1\}}^{(h)}\right]^2$ is the estimated variance by regressing X_j on X_1 in the hth slice, and $\widehat{\sigma}_{j|\{1\}}^2$ is the estimated variance by regressing X_j on X_1 using all the observations. We detected X_2 as an important predictor conditioning on X_1 because $\left(n\widehat{D}_{2|\{1\}}^*\right) = 148.83$ is highly significant compared with the null distribution, which is asymptotically $\chi^2(12)$ (here $\max_{j\in\{3,4,\ldots,1000\}}\left(n\widehat{D}_{j|\{1\}}^*\right) = 31.85$). Instead of screening all the $\binom{p}{2} = O(p^2)$ pairwise interaction terms, we discovered the importance of X_1 and X_2 by examining the conditional distributions of predictors given the sliced response, which only requires a computational complexity of $O(p)$. Motivated by this observation, we investigate an inverse modeling approach to tackle the problem of interaction detection in this paper. We will show that both $\left(n\widehat{D}_j^*\right)$ and $\left(n\widehat{D}_{j|\{1\}}^*\right)$ correspond to likelihood-ratio tests based on specific inverse models, and also describe a more general procedure to recursively select relevant predictors.

The rest of the paper is organized as follows. In Section 2, we introduce an inverse model for the conditional distribution of $\mathbf{X}$ given slices of Y. SIR method is presented as a maximum likelihood estimate of the model. A likelihood-ratio test statistic for selecting relevant predictors is derived in Section 2.1, which is shown to be asymptotically equivalent to the COP statistics of Zhong et al. (2012). We further augment the model to detect interactions between predictors in Section 2.2. A sure independence screening criterion based on the likelihood-ratio test statistic is proposed in Section 2.3. By cross-stitching independence screening and likelihood-ratio tests, an iterative stepwise procedure that we referred to as SIRI is developed in Section 3. Various implementation issues including the choices of slicing schemes and thresholds are also discussed. Simulations and real data examples are reported in Section 4 and 5. Additional remarks in Section 6 conclude the paper. Proofs of the theorems are provided in Appendix A.

2. Sliced Inverse Regression Model with Interaction Detection. Let $Y \in \mathbb{R}$ be a univariate response variable and $\mathbf{X} = (X_1, X_2, \ldots, X_p)^T \in \mathbb{R}^p$ be a vector of p continuous predictor variables. $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ are independent observations of $(\mathbf{X}, Y)$. For discrete response, we can naturally group $\{y_i\}_{i=1}^n$ into a finite number of classes. For continuous response, the range of $\{y_i\}_{i=1}^n$ can be divided into H disjoint intervals, referred to as slices, which are denoted as $S_1, S_2, \ldots, S_H$. We define a function $S(Y)$ as the slice membership of response Y, that is, $S(Y) = h$ if $Y \in S_h$. For a fixed slicing scheme, we denote $n_h = |S_h| = s_h n$ where $\sum_{h=1}^H s_h = 1$.

To take an inverse perspective on SIR, we start with a seemingly different model. We assume that the distribution of predictors given the sliced response is multivariate normal:

$$\mathbf{X}|Y \in S_h \sim \text{MVN}\left(\mu_h, \Sigma\right), 1 \leq h \leq H, \tag{2.1}$$

where $\mu_h \in \mu + \mathbb{V}^q$ belongs to a q-dimensional affine space, $\mathbb{V}^q$ is a q-dimensional subspace ($q < p$) and $\mu \in \mathbb{R}^p$. Alternatively, we can write $\mu_h = \mu + \Gamma\gamma_h$, where $\gamma_h \in \mathbb{R}^q$ and Γ is a p by q matrix whose columns form a basis of the subspace $\mathbb{V}^q$. Although this representation is only unique up to a orthogonal transformation on the bases Γ, the subspace $\mathbb{V}^q$ is unique and identifiable. The following proposition proved by Szretter and Yohai (2009) links the inverse model (2.1) with SIR.

PROPOSITION 1. *The maximum likelihood estimate (MLE) of the subspace $\mathbb{V}^q$ in model (2.1) coincides with the subspace spanned by SDR directions estimated from the SIR algorithm.*

According to Proposition 1, we could have re-derived the SIR algorithm from an inverse model. Next, we provide a view of the COP method for selecting variables with marginal effects from the inverse model (2.1). This will lay the ground work for the augmented model to select variables with interaction effects in Section 2.2.

2.1. *Likelihood Ratio Tests for Selecting Variables with Marginal Effects.* For the purpose of variable selection, we partition predictors into two subsets: a set of relevant predictors indexed by $\mathcal{A}$ and a set of redundant predictors indexed by $\mathcal{A}^c$, and assume the following model:

$$\begin{aligned} \mathbf{X}_{\mathcal{A}}|Y \in S_h &\sim \text{MVN}\left(\mu_h \in \mu + \mathbb{V}^q, \Sigma\right), \\ \mathbf{X}_{\mathcal{A}^c}|\mathbf{X}_{\mathcal{A}}, Y \in S_h &\sim \text{MVN}\left(\alpha + \boldsymbol{\beta}^T\mathbf{X}_{\mathcal{A}}, \Sigma_0\right). \end{aligned} \tag{2.2}$$

That is, we assume that the conditional distribution of relevant predictors follows the inverse model (2.1) of SIR and has a common covariance matrix

in different slices. Given the current set of selected predictors indexed by $\mathcal{C}$ with dimension d and another predictor indexed by $j \notin \mathcal{C}$, we propose the following hypotheses:

$$H_0 : \mathcal{A} = \mathcal{C} \text{ v.s. } H_1 : \mathcal{A} = \mathcal{C} \cup \{j\}.$$

Let $P_{\mathcal{M}_0}(\cdot)$ and $P_{\mathcal{M}_1}(\cdot)$ denote the probability distributions by plugging in the MLE of model parameters under H_0 and under H_1, respectively. Then, the likelihood-ratio test statistic can be written as

$$\begin{aligned} L_{j|\mathcal{C}} &= \frac{P_{\mathcal{M}_1}(\mathbf{X}|Y)}{P_{\mathcal{M}_0}(\mathbf{X}|Y)} = \frac{P_{\mathcal{M}_1}\left(\mathbf{X}_{[\mathcal{C}\cup\{j\}]^c}|X_j, \mathbf{X}_{\mathcal{C}}, Y\right) P_{\mathcal{M}_1}(X_j|\mathbf{X}_{\mathcal{C}}, Y)}{P_{\mathcal{M}_0}\left(\mathbf{X}_{[\mathcal{C}\cup\{j\}]^c}|X_j, \mathbf{X}_{\mathcal{C}}, Y\right) P_{\mathcal{M}_0}(X_j|\mathbf{X}_{\mathcal{C}}, Y)} \\ &= \frac{P_{\mathcal{M}_1}(X_j|\mathbf{X}_{\mathcal{C}}, Y)}{P_{\mathcal{M}_0}(X_j|\mathbf{X}_{\mathcal{C}}, Y)}, \end{aligned}$$

where the last equality follows from

$$P_{\mathcal{M}_1}\left(\mathbf{X}_{[\mathcal{C}\cup\{j\}]^c}|X_j, \mathbf{X}_{\mathcal{C}}, Y\right) = P_{\mathcal{M}_0}\left(\mathbf{X}_{[\mathcal{C}\cup\{j\}]^c}|X_j, \mathbf{X}_{\mathcal{C}}, Y\right)$$

according to model (2.2). The scaled log-likelihood-ratio test statistic is given by

$$\widehat{D}_{j|\mathcal{C}} = \frac{2}{n}\log\left(L_{j|\mathcal{C}}\right) = \sum_{k=1}^{q} \log\left(1 + \frac{\widehat{\lambda}_k^{d+1} - \widehat{\lambda}_k^{d}}{1 - \widehat{\lambda}_k^{d+1}}\right), \tag{2.3}$$

where $\widehat{\lambda}_k^d$ and $\widehat{\lambda}_k^{d+1}$ are estimates of the kth profile-R^2 based on $\mathbf{x}_{\mathcal{C}}$ and $\mathbf{x}_{\mathcal{C}\cup\{j\}}$, respectively. Under the null hypothesis, $\frac{\widehat{\lambda}_k^{d+1} - \widehat{\lambda}_k^{d}}{1 - \widehat{\lambda}_k^{d+1}} \xrightarrow{P} 0$. Since $\log(1 + t) \approx t$ when t is small, we have

$$\left(n\widehat{D}_{j|\mathcal{C}}\right) \xrightarrow{P} n\sum_{k=1}^{q} \frac{\widehat{\lambda}_k^{d+1} - \widehat{\lambda}_k^{d}}{1 - \widehat{\lambda}_k^{d+1}} \xrightarrow{d} \chi^2(q).$$

Coincidentally, from an inverse model, we re-discovered the COP statistics proposed by Zhong et al. (2012), which are defined as

$$\mathrm{COP}_k^{d+1} = n\frac{\widehat{\lambda}_k^{d+1} - \widehat{\lambda}_k^{d}}{1 - \widehat{\lambda}_k^{d+1}}, k = 1, 2, \ldots, q, \text{ and } \mathrm{COP}_{1:q}^{d+1} = \sum_{k=1}^{q} \mathrm{COP}_k^{d+1}.$$

For all the predictors indexed by $j \in \mathcal{C}^c$, we can also obtain the asymptotic joint distribution of $\left(n\widehat{D}_{j|\mathcal{C}}\right)$ under the null hypothesis:

$$\left(n\widehat{D}_{j|\mathcal{C}}\right)_{j\in\mathcal{C}^c} \xrightarrow{d} \left(\sum_{k=1}^{K} z_{kj}^2\right)_{j\in\mathcal{C}^c} \tag{2.4}$$

where $\mathbf{z}_k = (z_{kj})_{j\in\mathcal{C}^c} \sim \text{MVN}\left(\mathbf{0}, \left[\text{Corr}\left(X_i, X_j|\mathbf{X}_\mathcal{C}\right)\right]_{i,j\in\mathcal{C}^c}\right)$ and $\mathbf{z}_k$'s are independent.

Furthermore, as $n \to \infty$,

$$\begin{aligned}\widehat{D}_{j|\mathcal{C}} &\xrightarrow{a.s.} D_{j|\mathcal{C}} \\ &= \log\left(1 + \frac{\text{Var}(M_j) - \text{Cov}\left(M_j, \mathbf{X}_\mathcal{C}\right)\left[\text{Cov}\left(\mathbf{X}_\mathcal{C}\right)\right]^{-1}\text{Cov}\left(M_j, \mathbf{X}_\mathcal{C}\right)^T}{V_j}\right)\end{aligned}$$

where $M_j = \mathbb{E}\left(X_j|\mathbf{X}_\mathcal{C}, S(Y)\right)$, $V_j = \text{Var}\left(X_j|\mathbf{X}_\mathcal{C}, S(Y)\right)$ and $S(Y) = h$ when $Y \in S_h$ $(1 \le h \le H)$. Note that V_j does not depend on Y under the assumption in model (2.2). By Cauchy-Schwarz inequality and normality assumption,

$$D_{j|\mathcal{C}} = 0 \text{ iff } \mathbb{E}\left(X_j|\mathbf{X}_\mathcal{C}, Y \in S_h\right) = \mathbb{E}\left(X_j|\mathbf{X}_\mathcal{C}\right), 1 \le h \le H.$$

That is, the test statistic $\widehat{D}_{j|\mathcal{C}}$ almost surely converges to zero if the conditional mean of X_j is independent of slice membership $S(Y)$. Detailed proofs on properties of likelihood-ratio test statistic are delegated to Appendix A.1.

Given thresholds $\nu_a > \nu_d$ and the current set of selected predictors indexed by $\mathcal{C}$, we can select relevant variables by iterating the following steps until no new addition or deletion occurs:

- Addition step: find j_a such that $\widehat{D}_{j_a|\mathcal{C}} = \max_{j\in\mathcal{C}^c}\widehat{D}_{j|\mathcal{C}}$; let $\mathcal{C} = \mathcal{C}+\{j_a\}$ if $\widehat{D}_{j_a|\mathcal{C}} > \nu_a$.
- Deletion step: find j_d such that $\widehat{D}_{j_d|\mathcal{C}-\{j_d\}} = \min_{j\in\mathcal{C}}\widehat{D}_{j|\mathcal{C}-\{j\}}$; let $\mathcal{C} = \mathcal{C} - \{j_d\}$ if $\widehat{D}_{j_d|\mathcal{C}-\{j_d\}} < \nu_d$.

Under model (2.2), for relevant predictors indexed by $j \in \mathcal{A}$, we have

$$X_j|\mathbf{X}_{\mathcal{A}-\{j\}}, Y \in S_h \sim \text{N}\left(\alpha_j^{(h)} + \boldsymbol{\beta}_j^T\mathbf{X}_{\mathcal{A}-\{j\}}, \sigma^2_{j|\mathcal{A}-\{j\}}\right), 1 \le h \le H. \tag{2.5}$$

Let $\alpha_j(Y) = \sum_{h=1}^H \alpha_j^{(h)}\mathbb{I}\left(Y \in S_h\right)$. We introduce the following concept to study the marginal effect of relevant predictors.

Definition 1 (Marginally Detectable). *We say a predictor indexed by j is marginally detectable if there exist constants $\kappa \ge 0$ and $\xi > 0$ such that*

$$Var\left(\alpha_j(Y)\right) \ge \xi n^{-\kappa}. \tag{2.6}$$

Under the following conditions, the stepwise procedure proposed above is consistent for marginally detectable predictors by choosing the thresholds appropriately.

CONDITION 1. *There exist* $0 < \tau_{min} < \tau_{max} < \infty$ *such that*

$$\tau_{min} \leq \lambda_{min}\left(\mathrm{Cov}\left(\mathbf{X}|Y \in S_h\right)\right) < \lambda_{max}\left(\mathrm{Cov}\left(\mathbf{X}|Y \in S_h\right)\right) \leq \tau_{max},$$

and that

$$\lambda_{max}\left(\mathrm{Cov}\left(\mathbf{X}\right)\right) \leq \tau_{max},$$

where $\lambda_{min}(\cdot)$ *and* $\lambda_{max}(\cdot)$ *denote the smallest and largest eigenvalue of a positive definite matrix.*

CONDITION 2. $p = O(n^{\rho})$ *as* $n \to \infty$ *with* $\rho > 0$ *and* $2\rho + 2\kappa < 1$, *where* κ *is the same constant as in (2.6).*

The following theorem is proved in Appendix A.2.

THEOREM 1. *If all the relevant predictors indexed by* $\mathcal{A}$ *are marginally detectable with constant* κ, *then under model (2.2), Condition 1 and Condition 2, there exists constant* $c > 0$ *such that*

$$\Pr\left(\min_{\mathcal{C}:\mathcal{C}^c\cap\mathcal{A}\neq\emptyset}\max_{j\in\mathcal{C}^c}\widehat{D}_{j|\mathcal{C}} \geq cn^{-\kappa}\right) \to 1$$

and

$$\Pr\left(\max_{\mathcal{C}:\mathcal{C}^c\cap\mathcal{A}=\emptyset}\max_{j\in\mathcal{C}^c}\widehat{D}_{j|\mathcal{C}} < \frac{c}{2}n^{-\kappa}\right) \to 1,$$

as $n \to \infty$.

Thus, if we choose the threshold $\nu_a = cn^{-\kappa}$ and $\nu_d = (c/2)n^{-\kappa}$ with c and κ defined above, then the addition step will not stop selecting variables until all the relevant predictors have been included, and once all the relevant predictors have been included, all the redundant variables will be removed from the set of selected variables.

2.2. *Interaction Detection Under the Augmented Model.* Let us revisit the example (1.3) with the stepwise procedure based on the likelihood-ratio test statistic proposed in the previous section. As illustrated in Figure 1, we have $\mathbb{E}\left(X_1|Y \in S_h\right) = \mathbb{E}\left(X_2|Y \in S_h\right) = 0$ for $1 \leq h \leq H$. In the first addition step with $\mathcal{C} = \emptyset$, the stepwise procedure fails to capture either X_1 or X_2 since $D_{1|\mathcal{C}=\emptyset} = D_{2|\mathcal{C}=\emptyset} = 0$. In order to detect predictors with interactions, such as X_1 and X_2 in this example, we augment model (2.2) to a more general form,

$$\begin{aligned} \mathbf{X}_{\mathcal{A}}|Y \in S_h &\sim \mathrm{MVN}\left(\mu_h, \Sigma_h\right), \\ \mathbf{X}_{\mathcal{A}^c}|\mathbf{X}_{\mathcal{A}}, Y \in S_h &\sim \mathrm{MVN}\left(\alpha + \boldsymbol{\beta}^T\mathbf{X}_{\mathcal{A}}, \Sigma_0\right), \end{aligned} \tag{2.7}$$

which differs from model (2.2) in its allowing for slice-dependent means and covariance matrices for relevant predictors.

Under model (2.7), a predictor indexed by $j \in \mathcal{A}$ is conditionally irrelevant if the conditional distribution of X_j given $X_{\mathcal{A}-\{j\}}$ and $Y \in S_h$ does not depend on slice S_h. If there exists a conditionally irrelevant predictor indexed by $j \in \mathcal{A}$, then we can always redefine the index set of relevant predictors to be $\mathcal{A} - \{j\}$ in model (2.7). To guarantee identifiability, variables indexed by $\mathcal{A}$ in model (2.7) have to be minimally relevant, that is, $\mathcal{A}$ does not contain any conditionally irrelevant predictor. In Appendix A.3, we proved the following proposition:

PROPOSITION 2. *Minimally relevant predictors indexed by $\mathcal{A}$ in model (2.7) is unique under Condition* 1.

By following the same hypothesis testing framework for variable selection in the previous section, we can derive the scaled log-likelihood-ratio test statistic under the augmented model (2.7):

$$\widehat{D}^*_{j|\mathcal{C}} = \log \widehat{\sigma}^2_{j|\mathcal{C}} - \sum_{h=1}^{H} \frac{n_h}{n} \log \left[\widehat{\sigma}^{(h)}_{j|\mathcal{C}}\right]^2, \tag{2.8}$$

where $\mathcal{C}$ indexes currently selected predictors and $j \in \mathcal{C}^c$, $\left[\widehat{\sigma}^{(h)}_{j|\mathcal{C}}\right]^2$ is the estimated variance by regressing X_j on $\mathbf{X}_{\mathcal{C}}$ in slice S_h, and $\widehat{\sigma}^2_{j|\mathcal{C}}$ is the estimated variance by regressing X_j on $\mathbf{X}_{\mathcal{C}}$ using all the observations. The augmented test statistic $\left(n\widehat{D}^*_{j|\mathcal{C}}\right)$ was used to select relevant predictors in the illustrative example of Section 1.2.

Under the assumption that $\mathcal{A} \subset \mathcal{C}$ with $|\mathcal{C}| = d$, we can derive the exact and asymptotic distribution of $\left(n\widehat{D}^*_{j|\mathcal{C}}\right)$:

$$\begin{aligned} n\widehat{D}^*_{j|\mathcal{C}} &\sim n \log\left(1 + \frac{Q_0}{\sum_{h=1}^{H} Q_h}\right) - \sum_{h=1}^{H} \frac{n_h}{n} \log\left(\frac{Q_h/n_h}{\sum_{h=1}^{H} Q_h/n}\right) \\ &\xrightarrow{d} \chi^2((H-1)(d+2)), \end{aligned}$$

where $Q_0 \sim \chi^2((H-1)(d+1))$ and $Q_h \sim \chi^2(n_h - (d+1))$ $(1 \le h \le H)$ are mutually independent according to *Cochran's theorem*. For all the predictors indexed by $j \in \mathcal{C}^c$ given predictors indexed by $\mathcal{C}$, we can also obtain the asymptotic joint distribution of $\left(n\widehat{D}^*_{j|\mathcal{C}}\right)$ under the assumption that $\mathcal{A} \subset \mathcal{C}$:

$$\left(n\widehat{D}^*_{j|\mathcal{C}}\right)_{j\in\mathcal{C}^c} \xrightarrow{d} \left(\sum_{i=1}^{(H-1)(d+1)} z^2_{ij} + \sum_{i=1}^{H-1} \tilde{z}^2_{ij}\right)_{j\in\mathcal{C}^c}, \tag{2.9}$$

where $\mathbf{z}_i$'s and $\widetilde{\mathbf{z}}_i$'s are mutually independent with

$$\mathbf{z}_i = (z_{ij})_{j\in\mathcal{C}^c} \sim \text{MVN}\left(\mathbf{0}, \left[\text{Corr}\left(X_j, X_k|\mathbf{X}_\mathcal{C}\right)\right]_{j,k\in\mathcal{C}^c}\right),$$

and

$$\widetilde{\mathbf{z}}_i = (\widetilde{z}_{ij})_{j\in\mathcal{C}^c} \sim \text{MVN}\left(\mathbf{0}, \left[\text{Corr}^2\left(X_j, X_k|\mathbf{X}_\mathcal{C}\right)\right]_{j,k\in\mathcal{C}^c}\right).$$

When the sample size $n \to \infty$,

$$\begin{aligned} & \widehat{D}^*_{j|\mathcal{C}} \xrightarrow{a.s.} D^*_{j|\mathcal{C}} \\ = \quad & \log\left(1 + \frac{\text{Var}(M_j) - \text{Cov}\left(M_j, \mathbf{X}_\mathcal{C}\right)\left[\text{Cov}\left(\mathbf{X}_\mathcal{C}\right)\right]^{-1}\text{Cov}\left(M_j, \mathbf{X}_\mathcal{C}\right)^T}{\mathbb{E}\left(V_j\right)}\right) \\ & + \log\mathbb{E}\left(V_j\right) - \mathbb{E}\log\left(V_j\right) \end{aligned}$$

where $M_j = \mathbb{E}\left(X_j|\mathbf{X}_\mathcal{C}, S(Y)\right)$, $V_j = \text{Var}\left(X_j|\mathbf{X}_\mathcal{C}, S(Y)\right)$ and $S(Y) = h$ when $Y \in S_h$ $(1 \le h \le H)$. According to Cauchy-Schwarz inequality and Jensen's inequality,

$$\begin{aligned} D^*_{j|\mathcal{C}} = 0 \quad \text{iff} \quad & \mathbb{E}\left(X_j|\mathbf{X}_\mathcal{C}, Y \in S_h\right) = \mathbb{E}\left(X_j|\mathbf{X}_\mathcal{C}\right) \text{ and} \\ & \text{Var}\left(X_j|\mathbf{X}_\mathcal{C}, Y \in S_h\right) = \text{Var}\left(X_j|\mathbf{X}_\mathcal{C}\right), \end{aligned}$$

for $1 \le h \le H$. That is, the augmented test statistic $\widehat{D}^*_{j|\mathcal{C}}$ almost surely converges to zero if both the conditional mean and the conditional variance of X_j is independent of slice membership $S(Y)$. Proofs on properties of the augmented likelihood-ratio test statistic are delegated to Appendix A.4.

A forward-addition backward-deletion algorithm similar to the stepwise procedure proposed in Section 2.1 can be used with the augmented likelihood-ratio test statistic $\widehat{D}^*_{j|\mathcal{C}}$. To investigate the detectability of the augmented likelihood-ratio test, we introduce the following concepts.

DEFINITION 2 (Conditionally Detectable). *We say a collection of predictors indexed by $\mathcal{C}_2$ is conditionally detectable given predictors indexed by $\mathcal{C}_1$ if $\mathcal{C}_2 \cap \mathcal{C}_1 = \emptyset$, and for any set $\mathcal{C}$ satisfying $\mathcal{C}_1 \subset \mathcal{C}$ and $\mathcal{C}_2 \not\subset \mathcal{C}$, there exist constants $\kappa \ge 0$, $\xi_1, \xi_2 > 0$ such that either*

(2.10)
$$\max_{j\in\mathcal{C}^c\cap\mathcal{C}_1}\left[\frac{Var\left(M_j\right) - Cov\left(M_j, \mathbf{X}_\mathcal{C}\right)\left[Cov\left(\mathbf{X}_\mathcal{C}\right)\right]^{-1}Cov\left(M_j, \mathbf{X}_\mathcal{C}\right)^T}{\mathbb{E}\left(V_j\right)}\right] \ge \xi_1 n^{-\kappa},$$

or

$$\max_{j\in\mathcal{C}^c\cap\mathcal{C}_1}\left[\log\left(\mathbb{E}V_j\right) - \mathbb{E}\log\left(V_j\right)\right] \ge \xi_2 n^{-\kappa}$$

where $M_j = \mathbb{E}\left(X_j|\mathbf{X}_\mathcal{C}, S(Y)\right)$, $V_j = Var\left(X_j|\mathbf{X}_\mathcal{C}, S(Y)\right)$.

In other words, if the current selection $\mathcal{C}$ contains $\mathcal{C}_1$, then there always exist detectable predictors conditioning on currently selected variables until we include all the predictors indexed by $\mathcal{C}_2$. A relevant predictor X_j indexed by $j \notin \mathcal{C}_2$ is *not* conditionally detectable given $\mathcal{C}_1$ either because it is highly correlated with some other predictors, or its effect can only be detected when conditioning on predictors that have not been included in $\mathcal{C}_1$. Based on Definition 2, we define stepwise detectable recursively as following.

DEFINITION 3 (Stepwise Detectable). *A collection of predictors indexed by $\mathcal{T}_0$ is said to be 0-level detectable if $\mathbf{X}_{\mathcal{T}_0}$ is conditionally detectable given an empty set, and a collection of predictors indexed by $\mathcal{T}_m$ is said to be m-level detectable ($m \geq 1$) if $\mathbf{X}_{\mathcal{T}_m}$ is conditionally detectable given predictors indexed by $\cup_{i=1}^{m-1}\mathcal{T}_i$. Finally, a predictor indexed by j is said to be stepwise detectable if $j \in \cup_{i=1}^{\infty}\mathcal{T}_i$.*

According to Lemma 1 in Appendix A.2, given the same constant κ, there exists ξ_1 such that the set of marginally detectable predictors defined in Definition 1 is contained in $\mathcal{T}_0$, the set of 0-level detectable predictors. As a result, the definition of stepwise detectable expand the concept of marginally detectable. In Appendix A.5, we will show that by appropriately choosing thresholds, the stepwise procedure will keep adding predictors until all the stepwise detectable predictors have been included.

THEOREM 2. *If all the relevant predictors indexed by $\mathcal{A}$ are stepwise detectable with constant κ, then under model (2.2), Condition 1 and Condition 2, there exists constant $c^* > 0$ such that as $n \to \infty$,*

$$\Pr\left(\min_{\mathcal{C}:\mathcal{C}^c\cap\mathcal{A}\neq\emptyset}\max_{j\in\mathcal{C}^c}\widehat{D}^*_{j|\mathcal{C}} \geq c^* n^{-\kappa}\right) \to 1,$$

and

$$\Pr\left(\max_{\mathcal{C}:\mathcal{C}^c\cap\mathcal{A}=\emptyset}\max_{j\in\mathcal{C}^c}\widehat{D}^*_{j|\mathcal{C}} < \frac{c^*}{2} n^{-\kappa}\right) \to 1.$$

Thus, by appropriately choosing the thresholds, the stepwise procedure based on $\widehat{D}^*_{j|\mathcal{C}}$ is consistent in identifying stepwise detectable predictors.

2.3. *A Sure Independence Screening Strategy.* When dimensionality p is extremely large (e.g., exceeding n^2), the performance of the stepwise procedure can be compromised. So we recommend adding an independence screening step to first reduce the dimensionality from ultra-high to moderately high. A natural choice of test statistic for the independence screening

procedure is $\widehat{D}^*_{j|\mathcal{C}}$ with $\mathcal{C} = \emptyset$, that is,

$$\widehat{D}^*_j = \log \widehat{\sigma}^2_j - \sum_{h=1}^{H} \frac{n_h}{n} \log \left[\widehat{\sigma}^{(h)}_j\right]^2,$$

where $\left[\widehat{\sigma}^{(h)}_j\right]^2$ is the estimated variance of X_j in slice S_h, and $\widehat{\sigma}^2_j$ is the estimated variance of X_j using all the observations. If we rank predictors according to $\{\widehat{D}^*_j, 1 \le j \le p\}$, then a sure independence screening (SIS) procedure that takes the first $O(n)$ predictors has a high probability (almost surely) of including the independently detectable predictors defined below.

DEFINITION 4 (Independently Detectable). *We say a predictor X_j is independently detectable if there exist constants $\kappa \ge 0$ and $\xi_1, \xi_2 > 0$ such that either*

$$\frac{Var(\mathbb{E}(X_j|S(Y)))}{\mathbb{E}(Var(X_j|S(Y)))} \ge \xi_1 n^{-\kappa}, \tag{2.11}$$

or

$$\log \mathbb{E}(Var(X_j|S(Y))) - \mathbb{E} \log [Var(X_j|S(Y))] \ge \xi_2 n^{-\kappa}.$$

Simply put, independently detectable predictors have either different means or different variances across slices. Therefore, in the example (1.3), both X_1 and X_2 are independently detectable because $\mathrm{Var}(X_1|Y \in S_h)$ and $\mathrm{Var}(X_2|Y \in S_h)$ $(1 \le h \le H)$ are different across slices.

In Theorem 3, we proved that the SIS procedure based on $\{\widehat{D}^*_j, 1 \le j \le p\}$ almost surely includes the independently detectable predictors under the following condition with ultra-high dimensionality of predictors.

CONDITION 3. $\log(p) = O(n^\gamma)$ *as* $n \to \infty$ *with* $0 < \gamma + 2\kappa < 1$, *where* κ *is the same constant as in (2.11). Furthermore, the number of the relevant predictors* $|\mathcal{A}| \le \xi_0 n^\eta$ *with* $\eta + \kappa < 1$ *and constant* $\xi_0 > 0$.

We proved the following theorem in Appendix A.6.

THEOREM 3. *Under Condition 1 and Condition 3, if all the relevant predictors indexed by $\mathcal{A}$ are independently detectable, then there exist $c > 0$ and $C > 0$ such that*

$$\Pr\left(\min_{j \in \mathcal{A}} \widehat{D}^*_j \ge c n^{-\kappa}\right) \to 1,$$

and

$$\Pr\left(\left|\{j : \widehat{D}^*_j \ge c n^{-\kappa}, 1 \le j \le p\}\right| \le C n^{\kappa+\eta}\right) \to 1.$$

According to Theorem 3, we can first use the SIS procedure to reduce the dimensionality from p to a scale below sample size, say $n/\log(n)$, and then apply the stepwise procedure proposed in the previous sections. Note that predictors that are marginally or stepwise detectable according to Definition 1 and Definition 3 are not necessarily independently detectable. Fan and Lv (2008) advocated an iterative procedure that alternate between a large-scale screening and a moderate-scale variable selection to enhance the performance, which will be discussed in the next section.

3. Implementation Issues: Cross-Stitching and Cross-Validation. The simple model (2.2) and the augmented model (2.7) compensate each other in terms of the bias-variance trade-off. Given finite observations, model (2.2) is simpler and more powerful when the response is driven by some linear combinations of covariates, while model (2.7) is useful in detecting more complex relationships such as heteroscedastic effects or interactions. Similarly, the SIS procedure introduced in the previous section can be used with a large number of predictors, but cannot pick up stepwise detectable predictors that have the same marginal distributions across slices. To find a balance between simplicity and detectability, we propose the following cross-stitching strategy:

- Step 0: start with the SIS procedure with currently selected predictors $\mathcal{C} = \emptyset$;
- Step 1: select predictors by using the stepwise procedure with addition and deletion steps based on $\widehat{D}_{j|\mathcal{C}}$ in (2.3) and add the selected predictors into $\mathcal{C}$;
- Step 2: select predictors by using the stepwise procedure with addition and deletion steps based on $\widehat{D}^*_{j|\mathcal{C}}$ in (2.8) and add the selected predictors into $\mathcal{C}$;
- Step 3: run the SIS procedure on the remaining predictors conditioning on the current selection $\mathcal{C}$, and iterate Step $1-3$ until no more predictors are selected.

We name the proposed procedure *Sliced Inverse Regression for Interaction Detection*, or SIRI for short. An flowchart of the SIRI procedure is illustrated in Figure 2.

In the addition step of the stepwise procedure, instead of selecting the variable from $j \in \mathcal{C}^c$ with the maximum value of $\widehat{D}_{j|\mathcal{C}}$ (or $\widehat{D}^*_{j|\mathcal{C}}$), we may also sequentially add variables with $\widehat{D}_{j|\mathcal{C}} > \nu_a$ (or $\widehat{D}^*_{j|\mathcal{C}} > \nu^*_a$). Specifically, given thresholds $\nu_a > \nu_d$ and the current set of selected predictors indexed by $\mathcal{C}$, we can modify each iteration of the original stepwise procedure as following:

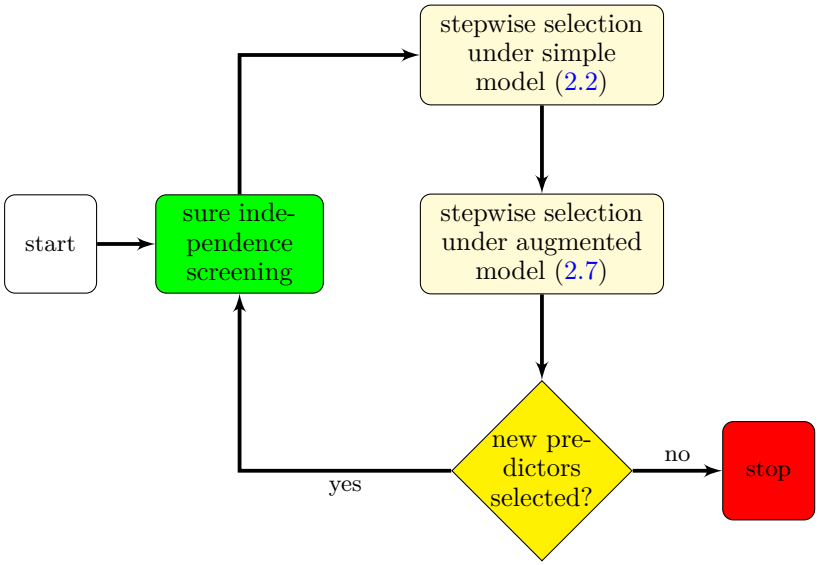

FIG 2. *Flowchart of SIRI*

- Modified addition step: for each variable $j \in \{1, \ldots, p\}$, let $\mathcal{C} = \mathcal{C} + \{j\}$ if $j \notin \mathcal{C}$ and $\widehat{D}_{j|\mathcal{C}} > \nu_a$.
- Deletion step: find j_d such that $\widehat{D}_{j_d|\mathcal{C}-\{j_d\}} = \min_{j\in\mathcal{C}} \widehat{D}_{j|\mathcal{C}-\{j\}}$; let $\mathcal{C} = \mathcal{C} - \{j_d\}$ if $\widehat{D}_{j_d|\mathcal{C}-\{j_d\}} < \nu_d$.

The stepwise procedure with the modified addition step may use fewer iterations to find all the relevant predictors and will not stop until all the relevant predictors have been included if we choose $\nu_a = cn^{-\kappa}$ in Theorem 1. However, in practice, the performance of the modified procedure depends on the ordering of the variables and is less stable than the original procedure. Since we are less concerned about the computational cost of SIRI, we implement the original addition step in the following study.

There are some implementation issues that we have not discussed so far. First, we need to choose a slicing scheme. If we assume there is a true slicing scheme from which data are generated, we showed in Appendix A.7 that the power of the stepwise procedure tends to increase with a larger number of slices, but there is no gain by further increasing the number of slices once the slicing is already more refined than the true slicing scheme. In practice, the true slicing scheme is usually unknown (except maybe in the case when the response is discrete). When a slicing scheme uses a larger number of slices, the number of observations in each slice decreases, which makes the estimation of parameters in the model less accurate and less stable. We

observed from intensive simulation studies that, with a reasonable number of observations in each slice (say 40 or more), a larger number of slices is preferred.

Second, we need to choose the number of dimensions q in model (2.2) and the thresholds in adding and deleting variables. Section 2.1 and 2.2 characterize the asymptotic distributions and behaviors of stepwise procedures, and provide some theoretical guidelines for choosing the thresholds. However, these theoretical results are not directly usable because: (1) the asymptotic distributions that we derived in (2.4) and (2.9) are for a single addition or deletion step; (2) the consistency results are valid in asymptotic sense and the rate of increase in dimension relative to sample size is usually unknown. In practice, we propose to use a K-fold cross-validation (CV) procedure for selecting thresholds and the number of dimensions q.

We consider two performance measures for K-fold cross-validation: classification error (CE) and mean absolute error (AE). Suppose there are n training samples and m testing samples. The ith observation ($i = 1, 2, \ldots, m$) in the testing set has response y_i and slice membership $S(y_j)$ (the slicing scheme is fixed based on training samples). Let $p_j^{(h)} = \Pr_{\mathcal{M}}(S(y_j) = h | \mathbf{X} = \mathbf{x}_j)$ be the estimated probability that the observation j is from slice S_h, where $\mathcal{M}$ denotes the maximum likelihood estimate of model parameters. The classification error is defined as

$$\text{CE} = \frac{1}{m} \sum_{j=1}^{m} \mathbb{I}\left[S(y_j) \neq \underset{h}{\text{argmax}} \left(p_j^{(h)} \right) \right].$$

We denote the average response of training samples in slice S_h as

$$\bar{y}^{(h)} = \frac{\sum_{i=1}^{n} \mathbb{I}\left[S(y_i) = h\right] y_i}{\sum_{i=1}^{n} \mathbb{I}\left[S(y_i) = h\right]}, h = 1, 2, \ldots, H,$$

The absolute error is defined as

$$\text{AE} = \frac{1}{m} \sum_{j=1}^{m} \left| y_j - \sum_{h=1}^{H} p_j^{(h)} \bar{y}^{(h)} \right|.$$

CE is a more relevant performance measure when the response is categorical or there is a non-smooth functional relationship (e.g., rational functions) between the response and predictors, and AE is a better measure when there is a monotonic and smooth functional relationship between the response and predictors. There are other measures that have compromise features between these two measures, such as median absolute deviation, which will not be explored here. We will use CE and AE as performance measures through out simulation studies and name the corresponding methods SIRI-AE and SIRI-CE, respectively.

4. Simulation Studies. In order to facilitate fair comparisons with other existing methods that are motivated from the forward model perspective, the examples presented here are all generated under forward models, which differs from the inverse model assumptions of SIRI. The setting of the simulation also demonstrates the robustness of SIRI when some of its model assumptions are violated, especially the normal distribution assumption on relevant predictor variables within each slice.

We start with the comparison of independence screening methods in reducing the ultra-high dimensionality while retaining the relevant predictors. Then, we evaluate different variable selection methods under a variety of forward models including linear model, single- and multi-index models and models with different types of interactions.

4.1. *Independence Screening Performance.* We first compare the variable screening performance of SIRI with iterative sure independence screening (ISIS) based on correlation learning proposed by Fan and Lv (2008) and sure independence screening based on distance correlation (DC-SIS) proposed by Li, Zhong and Zhu (2012). We evaluate the performance using the proportion that relevant predictors are placed among the top $[n/\log(n)]$ predictors ranked by the corresponding method, with larger values indicating better performance in variable screening.

In the simulation, the predictor variables $\mathbf{X} = (X_1, X_2, \ldots, X_p)^T$ were generated from a p-variate normal distribution with mean 0 and covariances $\operatorname{Cov}(X_i, X_j) = \rho^{|i-j|}$ for $1 \le i, j \le p$. We generate the response variable from the following three scenarios:

$$\begin{aligned}
&\text{Scenario 0.1}: \quad Y = X_2 - \rho X_1 + 0.2X_{100} + \sigma\epsilon,\\
&\text{Scenario 0.2}: \quad Y = X_1X_2 + \sigma e^{2|X_{100}|}\epsilon,\\
&\text{Scenario 0.3}: \quad Y = \frac{X_{100}}{X_1 + X_2} + \sigma\epsilon,
\end{aligned}$$

where sample size $n = 200$, $\sigma = 0.2$ and ϵ is $N(0,1)$ and independent of $\mathbf{X}$. For each scenario, we consider four different settings with dimension $p = 2000$ or 5000 and correlation $\rho = 0.0$ or 0.5. Scenario 0.1 is a linear model with three additive effects. The way X_1 is introduced is to make it marginally uncorrelated with the response Y (note that when $\rho = 0.0$, X_1 is not a relevant predictor). We added another variable X_{100} that has negligible correlation with X_1 and X_2 and a very small correlation with the response Y. Scenario 0.2 contains an interaction term X_1X_2 and a heteroscedatic noise term determined by X_{100}. Scenario 0.3 is an example of a rational model with interactions.

TABLE 1
The proportions that relevant predictors are placed among the top $[n/\log(n)]$ *by different screening methods under Scenarios* 0.1-0.3 *in Section 4.1.*

Method	Scenario 0.1			Scenario 0.2			Scenario 0.3		
	X_1	X_2	X_{100}	X_1	X_2	X_{100}	X_1	X_2	X_{100}
Setting 1: $p = 2000$, $\rho = 0.0$									
ISIS	-	1.00	1.00	0.02	0.01	0.46	0.00	0.00	0.09
DC-SIS	-	1.00	0.55	0.07	0.09	1.00	0.00	0.00	0.60
SIRI	-	1.00	0.30	0.32	0.25	0.97	1.00	0.99	1.00
Setting 2: $p = 2000$, $\rho = 0.5$									
ISIS	1.00	1.00	1.00	0.04	0.02	0.54	0.00	0.00	0.15
DC-SIS	0.02	1.00	0.71	0.55	0.53	1.00	0.03	0.00	0.59
SIRI	1.00	1.00	0.45	0.92	0.87	0.92	1.00	1.00	1.00
Setting 3: $p = 5000$, $\rho = 0.0$									
ISIS	-	1.00	1.00	0.02	0.00	0.43	0.00	0.00	0.06
DC-SIS	-	1.00	0.39	0.03	0.05	1.00	0.00	0.00	0.44
SIRI	-	1.00	0.14	0.15	0.16	0.99	0.99	1.00	1.00
Setting 4: $p = 5000$, $\rho = 0.5$									
ISIS	1.00	1.00	1.00	0.03	0.02	0.60	0.00	0.00	0.07
DC-SIS	0.05	1.00	0.71	0.41	0.44	1.00	0.00	0.02	0.61
SIRI	1.00	1.00	0.39	0.88	0.86	0.94	0.98	1.00	0.99

Proportions that relevant predictors are are placed among the top $[n/\log(n)]$ by different screening methods are shown in Table 1. Under Scenario 0.1 with linear models, we can see that ISIS and DC-SIS has better power than SIRI in detecting variables that are weakly correlated wth the response (X_{100} in this example). When the correlation between predictors $\rho = 0.5$ (Setting 2 and 4), iterative procedures, ISIS and SIRI, are more effective in detecting predictors that are marginally uncorrelated with the response (X_1 in this example) compared with DC-SIS. Under Scenario 0.2, ISIS based on linear models failed to detect the interaction term and often misses the predictor in the heteroscedastic noise term. When there are moderate correlations between two predictors X_1 and X_2 in the interaction term (Setting 2 and 4), DC-SIS picks up X_1 and X_2 about half of the time. However, when the two predictors are uncorrelated (Setting 1 and 3), DC-SIS failed to detect them most of the time. SIRI outperforms DC-SIS in detecting interactions for both settings with $\rho = 0.0$ and $\rho = 0.5$. Under Scenario 0.3, when there is a rational relationship between the response and the relevant predictors, SIRI significantly outperforms the other two methods in detecting the relevant predictors. Performances of different methods are only slightly affected as we increase the dimension from $p = 2000$ and $p = 5000$.

4.2. *Variable Selection Performance.* We further study the variable selection accuracy of SIRI and other existing methods in identifying relevant predictors and excluding irrelevant predictors. In the following examples, for both SIRI and COP, we implemented a fixed slicing scheme with 5 slices of equal size (*i.e.*, $H = 5$) and used a 10-fold CV procedure to determine the stepwise variable selection thresholds and the number of effective directions q in model (2.2) of Section 2.1. Specifically, the number of effective directions q was chosen from $\{0, 1, 2, 3, 4\}$, where $q = 0$ means that we skipped the variable selection step under simple model (2.2) in the iterative procedure described by Figure 2. The thresholds in addition and deletion steps were selected from the grid $\{(\nu_{i,a} = \chi^2(\alpha_i, q), \nu_{i,d} = \chi^2(\alpha_i - 0.05, q))\}$ for simple model (2.2) and from the grid $\{(\nu^*_{i,a} = \frac{n}{n-H(d+2)}\chi^2(\alpha_i, (H-1)(d+2)), \nu^*_{i,d} = \frac{n}{n-H(d+2)}\chi^2(\alpha_i - 0.05, (H-1)(d+2)))\}$ for augmented model (2.7), where $\chi^2(\alpha, \text{d.f.})$ is the 100αth quantile of $\chi^2(\text{d.f.})$ and $d = |\mathcal{C}|$ is the number of previously selected predictors. For a given p, the dimension of predictors, we chose $\{\alpha_i\} = \{1 - p^{-1}, 1 - 0.5p^{-1}, 1 - 0.1p^{-1}, 1 - 0.05p^{-1}, 1 - 0.01p^{-1}\}$.

The other variable selection methods to be compared with SIRI and COP include Lasso, ISIS-SCAD (SCAD with iterative sure independence screening), and hierNet, which is a Lasso-like procedure to detect multiplicative interactions between predictors under hierarchical constraints. The R packages glmnet, SIS, COP and hierNet are used to run Lasso, ISIS-SCAD, COP and hierNet, respectively. For Lasso and hierNet, we select the largest regularization parameter with estimated CV error less than or equal to the minimum estimated CV error plus one standard deviation of the estimate. The tuning parameters SCAD are also selected by CV.

For variable selections under index models, we generated the predictor variables $\mathbf{X} = (X_1, X_2, \ldots, X_p)^T$ from a multi-variate normal distribution with mean 0 and covariances $\text{Cov}(X_i, X_j) = \rho^{|i-j|}$ for $1 \le i, j \le p$, and simulated the response variable according to the following models:

$$
\begin{aligned}
\text{Scenario 1.1}: \quad & Y = \boldsymbol{\beta}^T\mathbf{X} + \sigma\epsilon, \quad n = 200, \sigma = 1.0, \rho = 0.5, \\
& \boldsymbol{\beta} = (3, 1.5, 2, 2, 2, 2, 2, 2, 0, \ldots, 0), \\
\text{Scenario 1.2}: \quad & Y = \frac{\sum_{j=1}^{3} X_j}{0.5 + (1.5 + \sum_{j=2}^{4} X_j)^2} + \sigma\epsilon, \\
& n = 200, \sigma = 0.2, \rho = 0.0, \\
\text{Scenario 1.3}: \quad & Y = \frac{\sigma\epsilon}{1.5 + \sum_{j=1}^{8} X_j}, \\
& n = 1000, \sigma = 0.2, \rho = 0.0,
\end{aligned}
$$

where n is the number of observations, p is the number of predictors and is

set as 1000 here, and the noise ϵ is independent of $\mathbf{X}$ and follows $N(0,1)$. Scenario 1.1 is a linear model which involves 8 true predictors and 992 irrelevant predictors. Scenario 1.2, a multi-index model with 4 true predictors, was studied in Li (1991) and Zhong et al. (2012), and there is a non-linear relationship between the response Y and two linear combinations of predictors $X_1 + X_2 + X_3$ and $X_2 + X_3 + X_4$. Scenario 1.3 is a single-index model with 8 true predictors and heteroscedastic noise.

For each simulation setting, we randomly generated 100 data sets each with n observations and applied variable selection methods to each data set. Two quantities, the average number of irrelevant predictors falsely selected as true predictors (which is referred to as FP) and the average number of true predictors falsely excluded as irrelevant predictors (which is referred to as FN), were used to measure the variable selection performance of each method. For example, under Scenario 1.1, the FPs and FNs range from 0 to 992 and from 0 to 8, respectively, with smaller values indicating better accuracies in variable selection. The FP- and FN-values of different methods together with their corresponding standard errors (in brackets) are reported in Table 2.

TABLE 2
False positive (FP) and false negative (FN) values of different variable selection methods under Scenario 1.1-1.3.

Method	Scenario 1.1		Scenario 1.2		Scenario 1.3	
	FP $(0, 992)$	FN $(0, 8)$	FP $(0, 996)$	FN $(0, 4)$	FP $(0, 992)$	FN $(0, 8)$
Lasso	0.59 (0.10)	0.00 (0.00)	0.08 (0.03)	1.07 (0.03)	0.00 (0.00)	8.00 (0.00)
ISIS-SCAD	0.35 (0.07)	0.00 (0.00)	0.60 (0.08)	1.02 (0.01)	5.08 (0.65)	7.97 (0.02)
hierNet	0.59 (0.10)	0.00 (0.00)	8.65 (0.36)	0.93 (0.03)	7.66 (0.48)	7.94 (0.02)
COP	0.69 (0.12)	0.06 (0.03)	1.84 (0.16)	0.98 (0.01)	1.26 (0.13)	3.32 (0.19)
SIRI-AE	0.01 (0.01)	0.09 (0.04)	0.13 (0.04)	0.07 (0.03)	0.43 (0.08)	4.82 (0.27)
SIRI-CE	0.26 (0.05)	0.08 (0.03)	0.55 (0.08)	0.09 (0.03)	2.02 (0.17)	0.51 (0.16)

Under Scenario 1.1, variable selection methods derived from linear models (Lasso, SCAD and hierNet) were able to detect all the relevant predictors (FN=0) with few false positives. On the other hand, COP, SIRI-AE and SIRI-CE missed some (about 10%) relevant predictors while excluded most irrelevant ones (lower FP vaues). The relatively high accuracy of methods developed for linear models is expected under this scenario, because the observations were simulated from a linear relationship. Under Scenario 1.2, Lasso achieved the lowest false positives, but it almost always missed one of the relevant predictor, X_4, because of its non-linear relationship with the response. The other methods developed under the linear model assump-

tion suffered from the same issue. However, SIRI-AE and SIRI-CE was able to detect most of the four relevant predictors (FN=0.09 and 0.07) with a comparable number of false positives. Under the heteroscedastic model in Scenario 1.3, the methods based on linear models failed to detect relevant predictors most of the time. Among other methods, SIRI-AE achieved the lowest number of false positives (FP=0.43) but missed about half of the relevant predictors (FN=4.82), while SIRI-CE selected most of the relevant predictors (FN=0.51) with a reasonably low false positives (FP=2.02). The performance of COP was in-between SIRI-AE and SIRI-CE with FN=3.32 and FP=1.26. A possible explanation for the better performance of SIRI-CE relative to SIRI-AE in this setting is because the generative model under Scenario 1.3 contains a singular point at $\sum_{j=1}^{8} X_j = -1.5$. Since the absolute error is less robust to outliers than the classification error, SIRI-AE is more sensitive to the inclusion of irrelevant predictors and more conservative in selecting predictors.

Next, we consider forward models containing different types interactions. Predictor variables $X_1, X_2, \ldots, X_p$ were generated as independent and identically distributed $N(0,1)$ random variables, and the response was generated under the following models given the predictors:

$$
\begin{aligned}
&\text{Scenario 2.1}: \quad Y = X_1X_2 + \sigma\epsilon, \quad n = 200,\\
&\text{Scenario 2.2}: \quad Y = X_1 + X_1X_2 + X_1X_3 + \sigma\epsilon, \quad n = 200,\\
&\text{Scenario 2.3}: \quad Y = X_1X_2 + X_1X_3 + \sigma\epsilon, \quad n = 200,\\
&\text{Scenario 2.4}: \quad Y = X_1X_2X_3 + \sigma\epsilon, \quad n = 200, 500 \text{ and } 1000,\\
&\text{Scenario 2.5}: \quad Y = X_1^2X_2 + \sigma\epsilon, \quad n = 200,\\
&\text{Scenario 2.6}: \quad Y = \frac{X_1}{X_2 + X_3} + \sigma\epsilon, \quad n = 200,
\end{aligned}
$$

where n is the number of observations, p is the number of predictors and is set as 1000 here, $\sigma = 0.2$ and ϵ is independent of $\mathbf{X}$ and follows $N(0,1)$. Scenario 2.1 and Scenario 2.3 contain predictors with pairwise multiplicative interactions and without main effects. The model under Scenario 2.2 has hierarchical interaction terms (X_1 has main effect). The three-way interaction model in Scenario 2.4 was simulated under three settings with different sample sizes: $n = 200$, $n = 500$ and $n = 1000$. Scenario 2.5 contains a quadratic interaction term and Scenario 2.6 has a rational relationship.

Because methods such as Lasso, SCAD and COP are not specifically designed for detecting interactions and are clearly at a disadvantage, we did not directly compare them with SIRI and hierNet. For the purpose of comparison, we created a benchmark method based on ISIS-SCAD by applying

ISIS-SCAD to an expanded set of predictors that includes all the terms up to k-way multiplicative interactions. The corresponding method, which we referred to as ISIS-SCAD-k, is an oracle benchmark under Scenario 2.1-2.3 when responses were generated according to 2-way multiplicative interactions. Since DC-SIS as a screening tool has the ability to detect individual predictors under the presence of interaction effects, we also augmented ISIS-SCAD with DC-SIS and denoted the method as DC-SIS-SCAD-k. In DC-SIS-SCAD-k, we first used DC-SIS to reduce the number of predictors from p to $[n/\log(n)]$. Then, we expanded the selected predictors to include up to k-way multiplicative interactions among them and applied ISIS-SCAD. Because DC-SIS-SCAD-k does not need to consider all the interaction terms among p predictors, it has a huge speed advantage over ISIS-SCAD-k but it may fail to detect all the predictors if the DC-SIS step does not retain all the relevant predictors. The FP- and FN-values (and their standard errors) of different methods including ISIS-SCAD-2 and DC-SIS-SCAD-2 under various scenarios are shown in Table 3, Table 4 and Table 5, respectively. Note that FP- and FN-values are calculated based on the number of predictors selected by a method, not based on the number of parameters used in building the model. For example, if X_3, X_4 and X_3X_4 all have non-zero coefficients from hierNet under Scenario 2.1, we count the number of false positives as 2, not 3.

TABLE 3
False positive (FP) and false negative (FN) values of different variable selection methods under Scenario 2.1-2.3.

Method	Scenario 2.1		Scenario 2.2		Scenario 2.3	
	FP $(0, 998)$	FN $(0, 2)$	FP $(0, 997)$	FN $(0, 3)$	FP $(0, 997)$	FN $(0, 3)$
ISIS-SCAD-2	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.06 (0.04)	0.00 (0.00)	0.03 (0.03)
DC-SIS-SCAD-2	0.00 (0.00)	0.00 (0.00)	0.25 (0.09)	0.11 (0.03)	1.56 (0.19)	1.81 (0.11)
hierNet	2.38 (0.33)	0.00 (0.00)	6.93 (0.56)	0.14 (0.05)	6.98 (0.57)	0.12 (0.05)
SIRI-AE	0.01 (0.01)	0.00 (0.00)	0.02 (0.01)	0.04 (0.02)	0.10 (0.04)	0.11 (0.05)
SIRI-CE	0.76 (0.13)	0.00 (0.00)	0.29 (0.06)	0.10 (0.04)	0.86 (0.12)	0.11 (0.05)

Under Scenarios 2.1-2.3 of Table 3, the oracle benchmark, ISIS-SCAD-2, correctly discovered most of the relevant predictors in the two-way interactions and did not pick up any irrelevant predictor. It is encouraging to see that the performance of the proposed method SIRI-AE was comparable with ISIS-SCAD-2 (in terms of both false positives and false negatives), although SIRI-AE did not assume the knowledge on the generative model. Moreover, since both ISIS-SCAD-2 and hierNet considered all the pairwise

interactions between p predictor variables, they have computational complexity $O(np^2)$ with $p = 1000$ and need much more computational resources compared with SIRI. On average ISIS-SCAD-2 and hierNet are more than 100 times slower than SIRI (see Table 6 for running time comparison of different methods). While we can dramatically increase the computational speed by using DC-SIS to screen variables before applying more refined variable selection methods, relevant predictors may be incorrectly filtered by the DC-SIS procedure as shown by DC-SIS-SCAD's higher false negatives under Scenario 2.3 of Table 3.

TABLE 4
False positive (FP) and false negative (FN) values of different variable selection methods under Scenario 2.4 *with different sample sizes.*

Method	Scenario 2.4 ($n = 200$)		Scenario 2.4 ($n = 500$)		Scenario 2.4 ($n = 1000$)	
	FP $(0, 997)$	FN $(0, 3)$	FP $(0, 997)$	FN $(0, 3)$	FP $(0, 997)$	FN $(0, 3)$
DC-SIS-SCAD-3	0.45 (0.12)	0.85 (0.12)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
hierNet	7.22 (0.64)	2.41 (0.08)	7.73 (1.17)	2.38 (0.08)	4.25 (1.17)	2.62 (0.06)
SIRI-AE	0.98 (0.12)	2.27 (0.06)	0.36 (0.09)	0.70 (0.07)	0.21 (0.06)	0.00 (0.00)
SIRI-CE	1.98 (0.16)	2.27 (0.07)	1.96 (0.17)	0.46 (0.05)	2.03 (0.19)	0.00 (0.00)

Under Scenario 2.4 with three-way interactions, the computational cost prevents us from directly applying ISIS-SCAD-3 to consider all the three-way interaction terms. So we only compared the performance of ISIS-SCAD-3 after variable screening using DC-SIS, that is, DC-SIS-SCAD-3 in Table 4. DC-SIS-SCAD-3 performed best under different sample sizes as it assumed the form of the underlying generative model. Among other methods, the performance of SIRI-AE improved relatively to hierNet as sample size increased. When sample size $n = 1000$, SIRI-AE was able to select all the relevant predictors with very low false positives.

TABLE 5
False positive (FP) and false negative (FN) values of different variable selection methods Scenario 2.5 *and* 2.6.

Method	Scenario 2.5		Scenario 2.6	
	FP $(0, 998)$	FN $(0, 2)$	FP $(0, 997)$	FN $(0, 3)$
ISIS-SCAD-2	0.04 (0.02)	1.09 (0.04)	0.00 (0.00)	3.00 (0.00)
DC-SIS-SCAD-2	2.38 (0.18)	0.51 (0.05)	0.81 (0.16)	2.96 (0.02)
hierNet	2.42 (0.44)	0.88 (0.05)	5.71 (0.59)	2.91 (0.03)
SIRI-AE	0.08 (0.03)	0.00 (0.00)	0.51 (0.11)	0.00 (0.00)
SIRI-CE	0.88 (0.11)	0.01 (0.01)	0.56 (0.11)	0.00 (0.00)

TABLE 6
Average running time (in seconds) of different variable selection methods under Scenarios 2.1-2.3, 2.5 *and* 2.6.

Method	Scenario 2.1	Scenario 2.2	Scenario 2.3	Scenario 2.5	Scenario 2.6
ISIS-SCAD-2	13890.86	9406.27	11581.55	10232.31	4220.24
DC-SIS-SCAD-2	29.24	25.77	31.90	37.03	25.68
hierNet	15213.01	26171.28	34733.13	37312.59	27255.16
SIRI	27.96	44.85	20.01	44.36	35.26

Simulations in Scenarios 2.1-2.4 were generated under the same model assumption as ISIS-SCAD-k and DC-SIS-SCAD-k, which gives them advantage in the comparisons. Under Scenarios 2.5 and 2.6 of Table 5, when the generative model goes beyond multiplicative interactions, we can see that SIRI-AE and SIRI-CE significantly outperformed other methods in detecting relevant predictors with low false positives.

5. Real Data Examples. We applied SIRI to two real data examples. The first example studies the problem of leukemia subtype classification with ultra-high dimensional features. In the second example, we treat gene expression level in embryonic stem cells as a continuous response variables, and are interested in selecting regulatory factors that interact with DNA and other factors to determine expression patterns of genes.

5.1. *Leukemia Classification.* For the first example, we applied SIRI-CE to select features for the classification of a leukemia data set from high density Affymetrix oligonucelotide arrays Golub et al. (1999) that has been previously analyzed by Tibshirani et al. (2002) using a nearest shrunken centroid method and Fan and Lv (2008) using a SIS-SCAD based linear discrimination method (SIS-SCAD-LD). The data set consists of 7129 genes and 72 samples from two classes: ALL (acute lymphocytic leukemia) with 47 samples and AML (acute mylogenous leukemia) with 25 samples. The data set was divided into a training set of 38 samples (27 in class ALL nad 11 in class AML) and a test set of 34 samples (20 in class ALL and 14 in class AML).

The classification results of SIRI-CE, SIS-SCAD-LD and nearest shrunken centroids method are shown in Table 7. The results of SIS-SCAD-LD and the nearest shrunken centroids method were extracted from Fan and Lv (2008) and Tibshirani et al. (2002), respectively. SIRI-CE and SIS-SCAD-LD both made no training error and one test error, whereas the nearest shrunken centroids method made one training error and two test errors. Comparing

TABLE 7
Leukemia classification results

Method	Training error	Test error	Number of genes
SIRI-CE	0/38	1/34	8
SIS-SCAD-LD	0/38	1/34	16
Nearest Shrunken Centroid	1/38	2/34	21

with SIS-SCAD-LD, SIRI used the smallest number of genes (8 genes) to achieve the same classification accuracy.

5.2. *Identifying Regulating Factors in Embryonic Stem Cells.* The mouse embryonic stem cells (ESCs) data set has previously been analyzed by Zhong et al. (2012) to identify important transcription factors (TFs) for regulating the expression of genes. The response variable, expression levels of 12408 genes, was quantified using RNA-seq technology in mouse ESCs (Cloonan et al., 2008). To understand the ESC development, it is important to identify key regulating TFs, whose binding profiles on promoter regions are associated with corresponding gene expression levels. To extract features that are associated with potential gene regulating TFs, Chen et al. (2008) performed ChIP-seq experiments on 12 TFs that are known to play different roles in ES-cell biology as components of the important signaling pathways, self-renewal regulators, and key reprogramming factors. For each pair of gene and one of these 12 TFs, a score named transcription factor association strength (TFAS) that was proposed by Ouyang, Zhou and Wong (2009) was calculated. In addition, Zhong et al. (2012) supplemented the data set with motif matching scores of 300 putative mouse TFs compiled from the TRANSFAC database. The TF motif matching scores were calculated based on the occurrences of TF binding motifs on gene promoter regions (Zhong et al., 2005). The data consists of a 12408×312 matrix with (i, j)th entry representing the score of the jth TF on the ith gene's promoter region.

In Zhong et al. (2012), the method COP selected a total of 42 predictors, which include 8 out of 12 TFASs and 34 out of 300 TF motif scores. Here, we used SIRI-AE to re-analyze the mouse ESCs data set and selected 34 predictors, which include all the 12 TFASs and 22 TF motif matching scores. The relative ranks of 12 TFASs from SIRI-AE and COP are shown in Table 8. Among the top-10 TFs ranked by SIRI-AE, 8 of them are known ES-cell TFs. SIRI-AE is also able to identify Nanog and Sox that are generally believed to be the master ESC regulators but were missed in the results of COP. A further study of the top-ranked TFs whose roles in ES cells have not been studied, such as AP2 (ranked 11 by SIRI), PAX (ranked 19 by SIRI) and

TABLE 8
The ranks of 12 *known ES-cell TFs (among 312 predictors) using SIRI-AE and COP*

TF names	Ranks	
	SIRI-AE	COP
E2f1	1	1
Zfx	3	3
Mycn	4	10
Klf4	5	19
Myc	6	-
Esrrb	8	-
Oct4	9	11
Tcfcp2l1	10	36
Nanog	14	-
Stat3	17	20
Sox2	18	-
Smad1	32	13

SP1 (ranked 22 by SIRI), could help us better understand transcriptional regulatory networks in embryonic stem cells.

6. Concluding Remarks. We study the problem of variable selection in high dimensions from an inverse modeling perspective. The contributions of the proposed procedure that we named SIRI is twofold. First, it is effective and computationally efficient in detecting interactions. Combined with independence screening, SIRI can be used to tackle the problem of ultra-high dimensionality when the number of predictors in the model is much larger than the number of observations. Second, SIRI does not impose any specific assumption on the relationship between the response and the predictors, and is a powerful tool for variable selections beyond linear models and detecting predictors with unknown form of interaction effects. As a trade-off, SIRI imposes a few assumptions on the distribution of the predictors. As demonstrated in our simulation studies, SIRI has competitive performance when the generative model is different from the inverse model assumption. However, we found that SIRI is not very robust against extreme outliers on values of the predictors. Data preprocessing, such as quantile normalization, is advised when extreme outliers are presented from exploratory analysis. We have implemented the SIRI procedure using programming language R, and the source code can be downloaded from http://www.people.fas.harvard.edu/~junliu/SIRI/ or requested from the authors directly.

Like other stepwise procedures such as linear stepwise regression, SIRI may encounter issues that are typical to stepwise variable selection methods

as discussed in Miller (1984). Imagine a simple classification example with two relevant predictors having the same mean and variance, but different correlations in two classes. Then, the predictors are undetectable to any stepwise procedure that selects one variables at a time. In less extreme scenarios, when relevant predictors have weak marginal effects but strong joint effects, iterative sampling procedures such as Gibbs sampling could be more powerful than stepwise procedures like SIRI. This motivates us to further study the problem of variable selection from a Bayesian perspective.

Finally, inverse models are not substitutes but complements to forward models. When a specific form is derived from solid scientific arguments, a forward perspective that treats the distribution of predictors as a nuisance can be more powerful in building predictive models. Depending on one's research questions and objectives, it may be helpful to alternate between the two perspectives in analyzing and interpreting data.

APPENDIX A: DETAILED PROOFS

A.1. Proof on Properties of Likelihood Ratio Test Statistic in Section 2.1. Given the set of relevant predictors indexed by $\mathcal{A}$ with size $|\mathcal{A}|$ in model (2.2), we let $\mathbf{x}_{\mathcal{A}}$ denote a $n \times |\mathcal{A}|$ matrix of observed variables in index set $\mathcal{A}$, and $\mathbf{x}_{\mathcal{A}}^{hj}$ $(1 \le j \le n_h)$ denote a $|\mathcal{A}|$-dimensional column vector of variables in index set $\mathcal{A}$ for jth observation in slice S_h $(1 \le h \le H)$. We denote $B_{\mathcal{A}} = \mathrm{Cov}\left(\mathbb{E}(\mathbf{X}_{\mathcal{A}}|S(Y))\right)$, $W_{\mathcal{A}} = \mathbb{E}\left(\mathrm{Cov}(\mathbf{X}_{\mathcal{A}}|S(Y))\right)$ and $\Omega_{\mathcal{A}} = B_{\mathcal{A}} + W_{\mathcal{A}}$. The corresponding sample estimates are given by

$$\widehat{B}_{\mathcal{A}} = \frac{1}{n}\sum_{h=1}^{H} n_h \left(\bar{\mathbf{x}}_{\mathcal{A}}^{h\cdot} - \bar{\mathbf{x}}_{\mathcal{A}}^{\cdot\cdot}\right)\left(\bar{\mathbf{x}}_{\mathcal{A}}^{h\cdot} - \bar{\mathbf{x}}_{\mathcal{A}}^{\cdot\cdot}\right)^T, \tag{A.1}$$

$$\widehat{W}_{\mathcal{A}} = \frac{1}{n}\sum_{h=1}^{H}\sum_{j=1}^{n_h} \left(\mathbf{x}_{\mathcal{A}}^{hj} - \bar{\mathbf{x}}_{\mathcal{A}}^{h\cdot}\right)\left(\mathbf{x}_{\mathcal{A}}^{hj} - \bar{\mathbf{x}}_{\mathcal{A}}^{h\cdot}\right)^T,$$

and $\widehat{\Omega}_{\mathcal{A}} = \widehat{B}_{\mathcal{A}} + \widehat{W}_{\mathcal{A}}$, where $\bar{\mathbf{x}}_{\mathcal{A}}^{\cdot\cdot} = \frac{1}{n}\sum_{h=1}^{H}\sum_{j=1}^{n_h} \mathbf{x}_{\mathcal{A}}^{hj}$ and $\bar{\mathbf{x}}_{\mathcal{A}}^{h\cdot} = \frac{1}{n_h}\sum_{j=1}^{n_h} \mathbf{x}_{\mathcal{A}}^{hj}$.

Szretter and Yohai (2009) proved the following proposition:

PROPOSITION 3. *Let C be the orthogonal matrix $[\mathbf{c}_1, \ldots, \mathbf{c}_p]$, where $\mathbf{c}_j$ is an eigenvector of $\widehat{B}_{\mathcal{A}}^{-1/2}\widehat{W}_{\mathcal{A}}\widehat{B}_{\mathcal{A}}^{-1/2} = C\Theta C^T$ corresponding to the eigenvalue θ_j, where $\theta_1 \ge \theta_2 \ge \ldots \ge \theta_p$. Θ is the diagonal matrix with $\theta_1, \theta_2, \ldots, \theta_p$ in the diagonal. C_r is the matrix with the first r columns of C.*
(a) The maximum likelihood estimate of $\Sigma = \Sigma_{\mathcal{A}} = \mathrm{Cov}\left(\mathbf{X}_{\mathcal{A}}|S(Y)\right)$ in model (2.2) is

$$\widehat{\Sigma}_{\mathcal{A}} = \widehat{W}_{\mathcal{A}} + \widehat{B}_{\mathcal{A}}^{1/2} C_{p-q} C_{p-q}^T \widehat{B}_{\mathcal{A}}^{1/2}. \tag{A.2}$$

(b) Let $\mathbf{u}_i$, $1 \le i \le p$, *be orthogonal eigenvectors of norm one of* $\widehat{\Sigma}_{\mathcal{A}}^{-1/2}\widehat{B}_{\mathcal{A}}\widehat{\Sigma}_{\mathcal{A}}^{-1/2}$ *corresponding to the eigenvalues* $\omega_1 \ge \omega_2 \ge \ldots \ge \omega_p$. *The maximum likelihood estimate of* $\mu_h = \mu_{\mathcal{A}}^{(h)} = \mathbb{E}\left(\mathbf{X}_{\mathcal{A}}|Y \in S_h\right)$ *in model (2.2) is*

$$\widehat{\mu}_{\mathcal{A}}^{(h)} = \widehat{\Sigma}_{\mathcal{A}}^{1/2} U_K U_q^T \widehat{\Sigma}_{\mathcal{A}}^{-1/2}\left(\bar{\mathbf{x}}_{\mathcal{A}}^{h\cdot} - \bar{\mathbf{x}}_{\mathcal{A}}^{\cdot\cdot}\right) + \bar{\mathbf{x}}_{\mathcal{A}}^{\cdot\cdot}, 1 \le h \le H,$$

where $U_q = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_q]$. *Then,* $\widehat{\mu}_{\mathcal{A}}^{(h)}$ *is the orthogonal projection, using the norm associated to* $\widehat{\Sigma}_{\mathcal{A}}$, *of* $\left(\bar{\mathbf{x}}_{\mathcal{A}}^{h\cdot} - \bar{\mathbf{x}}_{\mathcal{A}}^{\cdot\cdot}\right)$ *on the q-dimensional affine subspace* $\bar{\mathbf{x}}_{\mathcal{A}}^{\cdot\cdot} + \mathbb{V}^q$, *where* $\mathbb{V}^q$ *is spanned by* $\left(\widehat{\Sigma}_{\mathcal{A}}^{1/2}\mathbf{u}_1, \widehat{\Sigma}_{\mathcal{A}}^{1/2}\mathbf{u}_2, \ldots, \widehat{\Sigma}_{\mathcal{A}}^{1/2}\mathbf{u}_q\right)$.
(c) $\widehat{\Sigma}_{\mathcal{A}}$ *can also be written as*

$$\widehat{\Sigma}_{\mathcal{A}} = \frac{1}{n}\sum_{h=1}^{H}\sum_{j=1}^{n_h}\left(\mathbf{x}_{\mathcal{A}}^{hj} - \widehat{\mu}_{\mathcal{A}}^{(h)}\right)\left(\mathbf{x}_{\mathcal{A}}^{hj} - \widehat{\mu}_{\mathcal{A}}^{(h)}\right)^T.$$

$\widehat{\Sigma}_{\mathcal{A}}^{1/2}\mathbf{u}_j$ *is an eigenvector of* $\widehat{B}_{\mathcal{A}}\widehat{W}_{\mathcal{A}}^{-1}$ *corresponding to eigenvalue* $1/\theta_{p-i+1}$, $1 \le i \le p$.

Properties of the likelihood-ratio test statistic proposed in Section 2.1 are summarized in the following proposition:

PROPOSITION 4. *Given the current set of selected predictors indexed by* $\mathcal{C}$ *with dimension* $|\mathcal{C}| = d$ *and another predictor indexed by* $j \notin \mathcal{C}$, *the scaled log-likelihood-ratio test statistic for testing*

$$H_0 : \mathcal{A} = \mathcal{C} \text{ v.s. } H_1 : \mathcal{A} = \mathcal{C} \cup \{j\},$$

can be written as

$$\widehat{D}_{j|\mathcal{C}} = \frac{2}{n}\log\left(L_{j|\mathcal{C}}\right) = \sum_{k=1}^{q}\log\left(1 + \frac{\widehat{\lambda}_k^{d+1} - \widehat{\lambda}_k^d}{1 - \widehat{\lambda}_k^{d+1}}\right), \tag{A.3}$$

where q *is the dimension of subspace* $\mathbb{V}^q$ *defined in model (2.2),* λ_k^d *and* λ_k^{d+1} *are the* k*th largest eigenvalues of* $\Omega_{\mathcal{C}}^{-1}B_{\mathcal{C}}$ *and* $\Omega_{[\mathcal{C}\cup\{j\}]}^{-1}B_{[\mathcal{C}\cup\{j\}]}$, *respectively, and* $\widehat{\lambda}_k^d$ *and* $\widehat{\lambda}_k^{d+1}$ *are their estimates. For any fixed slicing scheme and the true relevant predictors indexed by* $\mathcal{A}$, *we let* λ_k *be the* k*th largest eigenvalue of* $\Omega_{\mathcal{A}}^{-1}B_{\mathcal{A}}$. *We further assume that* $\lambda_1 > \lambda_2 > \ldots > \lambda_q > 0$.
(a) Under the assumption that $\mathcal{A} \subset \mathcal{C}$,

$$n\widehat{D}_{j|\mathcal{C}} \simeq \sum_{i=1}^{q}\frac{n\left(\widehat{\lambda}_i^{d+1} - \widehat{\lambda}_i^d\right)}{1 - \widehat{\lambda}_i^{d+1}}, \tag{A.4}$$

which is asymptotically equivalent to the correlation pursuit (COP) test statistic defined in Zhong et al. *(2012) and has an asymptotic distribution of* $\chi^2(q)$.
(b) Under the same conditions as in (a),

$$\left(n\widehat{D}_{j|\mathcal{C}}\right)_{j\in\mathcal{C}^c} \xrightarrow{d} \left(\sum_{k=1}^{K} z_{kj}^2\right)_{j\in\mathcal{C}^c}, \text{ and } \max_{j\in\mathcal{C}^c}\left(n\widehat{D}_{j|\mathcal{C}}\right) \xrightarrow{d} \max_{j\in\mathcal{C}^c}\left(\sum_{i=1}^{q} z_{ij}^2\right),$$

where $\mathbf{z}_k = (z_{kj})_{j\in\mathcal{C}^c} \sim MVN\left(\mathbf{0}, [Corr(X_i, X_j|\mathbf{X}_\mathcal{C})]_{i,j\in\mathcal{C}^c}\right)$ *and* $\mathbf{z}_k$*'s are independent.*
(c) As $n \to \infty$,

$$\begin{aligned}\widehat{D}_{j|\mathcal{C}} &\xrightarrow{a.s.} D_{j|\mathcal{C}} \\ &= \log\left(1 + \frac{Var(M_j) - Cov(M_j, \mathbf{X}_\mathcal{C})\left[Cov(\mathbf{X}_\mathcal{C})\right]^{-1} Cov(M_j, \mathbf{X}_\mathcal{C})^T}{V_j}\right)\end{aligned}$$

where $M_j = \mathbb{E}(X_j|\mathbf{X}_\mathcal{C}, S(Y))$, $V_j = Var(X_j|\mathbf{X}_\mathcal{C}, S(Y))$ *and* $S(Y) = h$ *when* $Y \in S_h$ $(1 \le h \le H)$. V_j *does not depend on* Y *under the assumption in model (2.2). Furthermore,*

$$D_{j|\mathcal{C}} = 0 \text{ iff } \mathbb{E}(X_j|\mathbf{X}_\mathcal{C}, Y \in S_h) = \mathbb{E}(X_j|\mathbf{X}_\mathcal{C}), 1 \le h \le H.$$

Proof of Proposition 4. One can show that

$$\widehat{D}_{j|\mathcal{C}} = \frac{2}{n}\log\left(L_{j|\mathcal{C}}\right) = -\left(\log\left[\det\left(\widehat{\Omega}^{-1}_{[\mathcal{C}\cup\{j\}]}\widehat{\Sigma}_{[\mathcal{C}\cup\{j\}]}\right)\right] - \log\left[\det\left(\widehat{\Omega}^{-1}_{\mathcal{C}}\widehat{\Sigma}_{\mathcal{C}}\right)\right]\right).$$

To prove (A.3), we just need to show that

$$\log\left[\det\left(\widehat{\Omega}^{-1}_{\mathcal{C}}\widehat{\Sigma}_{\mathcal{C}}\right)\right] = \sum_{i=1}^{q}\log\left(1 - \widehat{\lambda}_i^d\right),$$

where $d = |\mathcal{C}|$ and $\widehat{\lambda}_i^d$ is the ith largest eigenvalue of $\widehat{\Omega}^{-1}_{\mathcal{C}}\widehat{B}_\mathcal{C}$ corresponding to eigenvector $\boldsymbol{\eta}_i$, $1 \le i \le d$. Since $\widehat{B}_\mathcal{C}\boldsymbol{\eta}_i = \widehat{\lambda}_i^d\widehat{\Omega}_\mathcal{C}\boldsymbol{\eta}_i$ and $\widehat{\Omega}_\mathcal{C} = \widehat{W}_\mathcal{C} + \widehat{B}_\mathcal{C}$,

$$\widehat{\Omega}_\mathcal{C}\boldsymbol{\eta}_i = \frac{1}{\widehat{\lambda}_i^d}\widehat{B}_\mathcal{C}\boldsymbol{\eta}_i = \frac{1}{1-\widehat{\lambda}_i^d}\widehat{W}_\mathcal{C}\boldsymbol{\eta}_i.$$

Then,

$$\widehat{W}_\mathcal{C}\widehat{B}^{-1}_\mathcal{C}\widehat{\Omega}_\mathcal{C}\boldsymbol{\eta}_i = \frac{1-\widehat{\lambda}_i^d}{\widehat{\lambda}_i^d}\widehat{\Omega}_\mathcal{C}\boldsymbol{\eta}_i,$$

and

$$\left(\widehat{B}_{\mathcal{C}}^{-1/2}\widehat{W}_{\mathcal{C}}\widehat{B}_{\mathcal{C}}^{-1/2}\right)\widehat{B}_{\mathcal{C}}^{-1/2}\widehat{\Omega}_{\mathcal{C}}\boldsymbol{\eta}_i = \frac{1-\widehat{\lambda}_i^d}{\widehat{\lambda}_i^d}\widehat{B}_{\mathcal{C}}^{-1/2}\widehat{\Omega}_{\mathcal{C}}\boldsymbol{\eta}_i.$$

Thus, the eigenvalues of $\widehat{B}_{\mathcal{C}}^{-1/2}\widehat{W}_{\mathcal{C}}\widehat{B}_{\mathcal{C}}^{-1/2}$ are given by $\theta_{d-i+1} = \frac{1-\widehat{\lambda}_i^d}{\widehat{\lambda}_i^d}$, $1 \le i \le d$. Let $\mathbf{c}_i$ be an eigenvector of $\widehat{B}_{\mathcal{C}}^{-1/2}\widehat{W}_{\mathcal{C}}\widehat{B}_{\mathcal{C}}^{-1/2} = C\Theta C^T$ corresponding to the eigenvalue θ_i. We will prove that the eigenvalues of $\widehat{\Sigma}_{\mathcal{C}}^{-1}\widehat{B}_{\mathcal{C}}$ are

$$\omega_i = \begin{cases} \frac{1}{\theta_{d-i+1}} = \frac{\widehat{\lambda}_i^d}{1-\widehat{\lambda}_i^d} & \text{if } 1 \le i \le q, \\ \frac{1}{1+\theta_{d-i+1}} = \widehat{\lambda}_i^d & \text{if } q+1 \le i \le d, \end{cases}$$

with corresponding eigenvectors given by $\mathbf{b}_i = \widehat{B}_{\mathcal{C}}^{-1/2}\mathbf{c}_{d-i+1}$. According to (A.2) in Proposition 3,

$$\begin{aligned} \widehat{B}_{\mathcal{C}}^{-1/2}\widehat{\Sigma}_{\mathcal{C}}\widehat{B}_{\mathcal{C}}^{-1/2}\mathbf{c}_{d-i+1} &= \widehat{B}_{\mathcal{C}}^{-1/2}\left(\widehat{W}_{\mathcal{C}} + \widehat{B}_{\mathcal{C}}^{1/2}C_{|\mathcal{A}|-q}C_{d-q}^T B_{\mathcal{C}}^{1/2}\right)\widehat{B}_{\mathcal{C}}^{-1/2}\mathbf{c}_{d-i+1} \\ &= \left(\widehat{B}_{\mathcal{C}}^{-1/2}\widehat{W}_{\mathcal{C}}\widehat{B}_{\mathcal{C}}^{-1/2} + C_{d-k}C_{d-q}^T\right)\mathbf{c}_{d-i+1}, \end{aligned}$$

where $\widehat{B}_{\mathcal{C}}^{-1/2}\widehat{W}_{\mathcal{C}}\widehat{B}_{\mathcal{C}}^{-1/2}\mathbf{c}_{d-i+1} = \theta_{d-i+1}$, $C_{d-q}C_{d-q}^T\mathbf{c}_{d-i+1}$ is 0 for $1 \le i \le q$ and 1 for $q+1 \le i \le d$. Thus,

$$\widehat{B}_{\mathcal{C}}^{-1/2}\widehat{\Sigma}_{\mathcal{C}}\widehat{B}_{\mathcal{C}}^{-1/2}\mathbf{c}_{d-i+1} = \frac{1}{\omega_i}\mathbf{c}_{d-i+1},$$

and

$$\widehat{\Sigma}_{\mathcal{C}}^{-1}\widehat{B}_{\mathcal{C}}\mathbf{b}_i = \omega_i\mathbf{b}_i, 1 \le i \le d.$$

Therefore,

$$\begin{aligned} \log\left[\det\left(\widehat{\Omega}_{\mathcal{C}}^{-1}\widehat{\Sigma}_{\mathcal{C}}\right)\right] &= \log\left[\det\left(\widehat{\Omega}_{\mathcal{C}}^{-1}\widehat{B}_{\mathcal{C}}\right)\right] - \log\left[\det\left(\widehat{\Sigma}_{\mathcal{C}}^{-1}\widehat{B}_{\mathcal{C}}\right)\right] \\ &= \sum_{i=1}^{q}\log\left(1-\widehat{\lambda}_i^d\right). \end{aligned}$$

To prove (a) and (b), note that

$$\begin{aligned} n\widehat{D}_{j|\mathcal{C}} &= -n\left(\sum_{i=1}^{q}\log\left(1-\widehat{\lambda}_i^{d+1}\right) - \sum_{i=1}^{q}\log\left(1-\widehat{\lambda}_i^d\right)\right) \\ &= n\sum_{i=1}^{q}\log\left(1+\frac{\widehat{\lambda}_i^{d+1}-\widehat{\lambda}_i^d}{1-\widehat{\lambda}_i^{d+1}}\right). \end{aligned}$$

Given that $\mathcal{A} \subset \mathcal{C}$, Zhong et al. (2012) have showed that

$$\frac{n\left(\widehat{\lambda}_i^{d+1} - \widehat{\lambda}_i^{d}\right)}{1 - \widehat{\lambda}_i^{d+1}}, i = 1, 2, \ldots, q,$$

are asymptotically independent and identically distributed as $\chi^2(1)$. Thus,

$$\frac{\widehat{\lambda}_i^{d+1} - \widehat{\lambda}_i^{d}}{1 - \widehat{\lambda}_i^{d+1}} \xrightarrow{P} 0, i = 1, 2, \ldots, q,$$

and

$$n\widehat{D}_{j|\mathcal{C}} = n\sum_{i=1}^{q} \log\left(1 + \frac{\widehat{\lambda}_i^{d+1} - \widehat{\lambda}_i^{d}}{1 - \widehat{\lambda}_i^{d+1}}\right) \simeq \sum_{i=1}^{q} \frac{n\left(\widehat{\lambda}_i^{d+1} - \widehat{\lambda}_i^{d}\right)}{1 - \widehat{\lambda}_i^{d+1}}.$$

Conditioning on variables in $\mathcal{C}$, variables in $\mathcal{C}^c$ follows a multivariate normal distribution with the same mean and variance across slices. Then, the proofs of (a) and (b) directly follow from the Theorem 1 and Theorem 2 in Zhong et al. (2012).

For (c), since $\widehat{\lambda}_i^{d} \xrightarrow{P} \lambda_i^{d}$, for $i = 1, 2, \ldots, d$,

$$\begin{aligned}
\widehat{D}_{j|\mathcal{C}} \xrightarrow{P} & -\sum_{i=1}^{q} \log\left(1 - \lambda_i^{d+1}\right) + \sum_{i=1}^{q} \log\left(1 - \lambda_i^{d}\right) \\
= & -\log\left[\det\left(\Omega_{[\mathcal{C}\cup\{j\}]}^{-1} W_{[\mathcal{C}\cup\{j\}]}\right)\right] + \log\left[\det\left(\Omega_{\mathcal{C}}^{-1} W_{\mathcal{C}}\right)\right] \\
= & \log\left[\frac{\det\left(\Omega_{[\mathcal{C}\cup\{j\}]}\right)}{\det\left(\Omega_{\mathcal{C}}\right)}\right] - \log\left[\frac{\det\left(W_{[\mathcal{C}\cup\{j\}]}\right)}{\det\left(W_{\mathcal{C}}\right)}\right] \\
= & \log\left[\mathrm{Var}\left(X_j\right) - \mathrm{Cov}\left(X_j, \mathbf{X}_{\mathcal{C}}\right)\left[\mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}}\right)\right]^{-1}\mathrm{Cov}\left(X_j, \mathbf{X}_{\mathcal{C}}\right)^T\right] \\
& -\log\left(V_j\right) \\
= & \log\left(1 + \frac{\mathrm{Var}\left(M_j\right) - \mathrm{Cov}\left(M_j, \mathbf{X}_{\mathcal{C}}\right)\left[\mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}}\right)\right]^{-1}\mathrm{Cov}\left(M_j, \mathbf{X}_{\mathcal{C}}\right)^T}{V_j}\right).
\end{aligned}$$

The last equality follows from $\mathrm{Var}\left(X_j\right) = \mathrm{Var}\left(M_j\right) + \mathbb{E}(V_j) = \mathrm{Var}\left(M_j\right) + V_j$ and $\mathrm{Cov}\left(X_j, \mathbf{X}_{\mathcal{C}}\right) = \mathrm{Cov}\left(M_j, \mathbf{X}_{\mathcal{C}}\right)$, where $M_j = \mathbb{E}\left(X_j|\mathbf{X}_{\mathcal{C}}, S(Y)\right)$, $V_j = \mathrm{Var}\left(X_j|\mathbf{X}_{\mathcal{C}}, S(Y)\right)$ and V_j is a constant that does not depend on $\mathbf{X}_{\mathcal{C}}$ or $S(Y)$. Note that

$$\mathrm{Var}\left(M_j\right) \geq \mathrm{Cov}\left(M_j, \mathbf{X}_{\mathcal{C}}\right)\left[\mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}}\right)\right]^{-1}\mathrm{Cov}\left(M_j, \mathbf{X}_{\mathcal{C}}\right)^T$$

where the equality holds if and only if $M_j = \mathbb{E}\left(X_j|\mathbf{X}_{\mathcal{C}}, S(Y)\right)$ is a linear combination of $\mathbf{X}_{\mathcal{C}}$ that does not depend on $S(Y)$, that is, $\mathbb{E}\left(X_j|\mathbf{X}_{\mathcal{C}}, Y \in S_h\right) = \mathbb{E}\left(X_j|\mathbf{X}_{\mathcal{C}}\right)$ for $1 \leq h \leq H$ under the normality assumption. □

A.2. Proof of Theorem 1 in Section 2.1. To prove Theorem 1, we will need the following two lemmas.

LEMMA 1. *Under the same conditions as in Theorem 1, there exist* $\kappa > 0$ *and* $\xi_1 > 0$ *such that for any set of predictors indexed by* $\mathcal{C}$ *and* $\mathcal{C}^c \cap \mathcal{A} \neq \emptyset$,

$$\max_{j \in \mathcal{C}^c \cap \mathcal{A}} \left[\frac{Var(M_j) - Cov(M_j, \mathbf{X}_\mathcal{C}) \left[Cov(\mathbf{X}_\mathcal{C})\right]^{-1} Cov(M_j, \mathbf{X}_\mathcal{C})^T}{V_j} \right] \geq \xi_1 n^{-\kappa},$$

where $M_j = \mathbb{E}(X_j | \mathbf{X}_\mathcal{C}, S(Y))$, $V_j = Var(X_j | \mathbf{X}_\mathcal{C}, S(Y))$, *and* V_j *is a constant that does not depend on* $\mathbf{X}_\mathcal{C}$ *or* $S(Y)$.

LEMMA 2. *Under the same conditions as in Theorem 1, for any set of predictors indexed by* $\mathcal{C}$, *we let* $\widehat{\lambda}_i^\mathcal{C}$ *be the* i*th largest eigenvalue of* $\widehat{\Omega}_\mathcal{C}^{-1} \widehat{B}_\mathcal{C}$ *and let* $\lambda_i^\mathcal{C}$ *be the* i*th largest eigenvalue of* $\Omega_\mathcal{C}^{-1} B_\mathcal{C}$. *Then, for* $0 < \epsilon < 1$ *and* $i = 1, 2, \ldots, q$, *there exist positive constants* C_1 *and* C_2 *such that*

$$\begin{aligned} & Pr\left(\max_{\mathcal{C} \subset \{1,2,\ldots,p\}} \left| \log\left(1 - \widehat{\lambda}_i^\mathcal{C}\right) - \log\left(1 - \lambda_i^\mathcal{C}\right) \right| > \epsilon \right) \\ \leq \; & 2p(p+1) C_1 \exp\left(-C_2 n \frac{\tau_{min}^4 \epsilon^2}{64 \tau_{max}^2 p^2} \right) \end{aligned} \tag{A.5}$$

We will start with the proofs of two lemmas.

PROOF OF LEMMA 1. Let $\mathcal{B} = \mathcal{C} \cap \mathcal{A}$, $\mathcal{E} = \mathcal{C} \cap \mathcal{A}^c$, and $\mathcal{D} = \mathcal{C}^c \cap \mathcal{A} \neq \emptyset$. Under model (2.2),

$$\mathbf{X}_\mathcal{D} | \mathbf{X}_\mathcal{B}, Y \in S_h \sim N\left(\alpha_{\mathcal{D}|\mathcal{B}}^{(h)} + \boldsymbol{\beta}_{\mathcal{D}|\mathcal{B}}^T \mathbf{X}_\mathcal{B}, \Sigma_1 = \Sigma_{\mathcal{D}|\mathcal{B}} \right).$$

Let $\alpha_\mathcal{D}^{(h)} = \left(\alpha_{j \in \mathcal{D}}^{(h)}\right)^T$, where $\alpha_j^{(h)}$ is defined in (2.5). Then, $\alpha_\mathcal{D}^{(h)} = \Psi \alpha_{\mathcal{D}|\mathcal{B}}^{(h)}$, where Ψ is a $|\mathcal{D}|$ by $|\mathcal{D}|$ matrix that satisfies $\Psi^T \Psi = \Sigma_1^{-1} \Delta_\mathcal{D}^2 \Sigma_1^{-1}$, $\Delta_\mathcal{D} = \text{diag}\left(\sigma_{j|\mathcal{A}-\{j\}}^2\right)_{j \in \mathcal{D}}$ and $\sigma_{j|\mathcal{A}-\{j\}}^2$ is defined in (2.5). Under Condition 1, $\sigma_{j|\mathcal{A}-\{j\}}^2 \leq \tau_{\max}$ and $\lambda_{\max}\left(\Sigma_1^{-1}\right) \leq \frac{1}{\tau_{\min}}$, and $\lambda_{\max}\left(\Psi^T \Psi\right) \leq \left(\frac{\tau_{\max}}{\tau_{\min}}\right)^2$. Thus,

$$\begin{aligned} \text{trace}\left(\text{Var}\left(\alpha_\mathcal{D}(Y)\right)\right) &= \text{trace}\left(\Psi \text{Var}\left(\alpha_{\mathcal{D}|\mathcal{B}}(Y)\right) \Psi^T\right) \\ &\leq \left(\frac{\tau_{\max}}{\tau_{\min}}\right)^2 \text{trace}\left(\text{Var}\left(\alpha_{\mathcal{D}|\mathcal{B}}(Y)\right)\right), \end{aligned}$$

where $\alpha_\mathcal{D}(Y) = \alpha_\mathcal{D}^{(h)}$ and $\alpha_{\mathcal{D}|\mathcal{B}}(Y) = \alpha_{\mathcal{D}|\mathcal{B}}^{(h)}$ when $Y \in S_h$. Furthermore,

$$\text{trace}\left(\text{Var}\left(\alpha_{\mathcal{D}|\mathcal{B}}(Y)\right)\right) \geq |\mathcal{D}| \left(\frac{\tau_{\min}}{\tau_{\max}}\right)^2 \xi n^{-\kappa}.$$

Assume

$$\mathrm{Cov}\left(\mathbf{X}_{\mathcal{E}\cup\mathcal{D}}|\mathbf{X}_{\mathcal{B}}\right)=\begin{pmatrix}\Sigma_0 & \Sigma_{01}\\ \Sigma_{10} & \Sigma_1\end{pmatrix},$$

Under model (2.2), conditioning on $\mathbf{X}_{\mathcal{A}}=\mathbf{X}_{\mathcal{B}\cup\mathcal{D}}$, the distribution of $\mathbf{X}_{\mathcal{E}}$ does not depend on the slice. Therefore, the conditional distribution of $\mathbf{X}_{\mathcal{D}}$ given $\mathbf{X}_{\mathcal{C}}=\mathbf{X}_{\mathcal{B}\cup\mathcal{E}}$ can be written as

$$\mathbf{X}_{\mathcal{D}}|\mathbf{X}_{\mathcal{C}},Y\in S_h\sim N\left(\alpha^{(h)}_{\mathcal{D}|\mathcal{C}}+\boldsymbol{\beta}^T_{\mathcal{D}|\mathcal{C}}\mathbf{X}_{\mathcal{C}},\Sigma_{\mathcal{D}|\mathcal{C}}\right),$$

where $\alpha^{(h)}_{\mathcal{D}|\mathcal{C}}=\alpha_0+\mathbf{M}\alpha^{(h)}_{\mathcal{D}|\mathcal{B}}$, $\mathbf{M}=\Sigma_1^{\frac{1}{2}}\left(\mathbf{I}_{|\mathcal{D}|}-\mathbf{N}^T\left(\mathbf{I}_{|\mathcal{E}|}+\mathbf{N}\mathbf{N}^T\right)^{-1}\mathbf{N}\right)\Sigma_1^{-\frac{1}{2}}$, $\mathbf{N}=\Sigma_0^{-\frac{1}{2}}\Sigma_{01}\Sigma_1^{-\frac{1}{2}}$ and α_0 is a constant that does not depend on the slice. Since $\lambda_{\max}\left(\mathbf{N}\mathbf{N}^T\right)\leq\frac{\lambda_{\max}(\Sigma_0)}{\lambda_{\min}(\Sigma_0)}\leq\frac{\tau_{\max}}{\tau_{\min}}$, $\lambda_{\min}\left(\Sigma_1\right)\geq\tau_{\min}$ and $\lambda_{\min}\left(\Sigma_1^{-1}\right)\geq\frac{1}{\tau_{\max}}$, we have

$$\lambda_{\min}\left(\mathbf{I}_{|\mathcal{D}|}-\mathbf{N}^T\left(\mathbf{I}_{|\mathcal{E}|}+\mathbf{N}\mathbf{N}^T\right)^{-1}\mathbf{N}\right)\geq\frac{1}{1+\frac{\tau_{\max}}{\tau_{\min}}}=\frac{\tau_{\min}}{\tau_{\max}+\tau_{\min}},$$

$$\begin{aligned}
&\mathrm{trace}\left(\mathrm{Var}\left(\alpha_{\mathcal{D}|\mathcal{C}}(Y)\right)\right)=\mathrm{trace}\left(\mathbf{M}\mathrm{Var}\left(\alpha_{\mathcal{D}|\mathcal{B}}(Y)\right)\mathbf{M}^T\right)\\
\geq\ &\lambda_{\min}\left(\Sigma_1^{-1}\right)\lambda_{\min}\left(\Sigma_1\right)\lambda^2_{\min}\left(\mathbf{I}_{|\mathcal{D}|}-\mathbf{N}^T\left(\mathbf{I}_{|\mathcal{E}|}+\mathbf{N}\mathbf{N}^T\right)^{-1}\mathbf{N}\right)\\
&\times\mathrm{trace}\left(\mathrm{Var}\left(\alpha_{\mathcal{D}|\mathcal{B}}(Y)\right)\right)\\
\geq\ &\left(\frac{\tau_{\min}}{\tau_{\max}+\tau_{\min}}\right)^2\left(\frac{\tau_{\min}}{\tau_{\max}}\right)\left(\mathrm{Var}\left(\alpha_{\mathcal{D}|\mathcal{B}}(Y)\right)\right)\\
\geq\ &|\mathcal{D}|\left(\frac{\tau_{\min}}{\tau_{\max}+\tau_{\min}}\right)^2\left(\frac{\tau_{\min}}{\tau_{\max}}\right)^3\xi n^{-\kappa}
\end{aligned}$$

Thus, there exists $j\in\mathcal{D}=\mathcal{C}^c\cap\mathcal{A}$ such that

$$\mathrm{Var}\left(\alpha_{j|\mathcal{C}}(Y)\right)\geq\left(\frac{\tau_{\min}}{\tau_{\max}+\tau_{\min}}\right)^2\left(\frac{\tau_{\min}}{\tau_{\max}}\right)^3\xi n^{-\kappa}.$$

For such j, we have $M_j=\alpha_{j|\mathcal{C}}(Y)+\boldsymbol{\beta}^T_{j|\mathcal{C}}\mathbf{X}_{\mathcal{C}}$, $V_j=\Sigma_{j|\mathcal{C}}\leq\tau_{\max}$, and

$$\begin{aligned}
&\mathrm{Var}\left(M_j\right)-\mathrm{Cov}\left(M_j,\mathbf{X}_{\mathcal{C}}\right)\left[\mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}}\right)\right]^{-1}\mathrm{Cov}\left(M_j,\mathbf{X}_{\mathcal{C}}\right)^T\\
=\ &\mathrm{Var}\left(\alpha_{j|\mathcal{C}}(Y)\right)-\mathrm{Cov}\left(\alpha_{j|\mathcal{C}}(Y),\mathbb{E}\left(\mathbf{X}_{\mathcal{C}}|S(Y)\right)\right)\left[\mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}}\right)\right]^{-1}\\
&\mathrm{Cov}\left(\alpha_{j|\mathcal{C}}(Y),\mathbb{E}\left(\mathbf{X}_{\mathcal{C}}|S(Y)\right)\right)^T.
\end{aligned}$$

Let $\mathbf{T}_1 = \mathrm{Cov}\left(\mathbb{E}\left(\mathbf{X}_{\mathcal{C}}|S(Y)\right)\right)$ and $\mathbf{T}_2 = \mathbb{E}\left(\mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}}|S(Y)\right)\right) = \mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}}|S(Y)\right)$. Since $\lambda_{\min}\left(\mathbf{T}_2\right) \geq \tau_{\min}$ and $\lambda_{\max}\left(\mathbf{T}_1\right) \leq \lambda_{\max}\left(\mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}}\right)\right) \leq \tau_{\max}$, we have $\lambda_{\min}\left(\mathbf{T}_1^{-\frac{1}{2}}\mathbf{T}_2\mathbf{T}_1^{-\frac{1}{2}}\right) \geq \frac{\tau_{\min}}{\tau_{\max}}$ and

$$
\begin{aligned}
& \mathrm{Cov}\left(\alpha_{j|\mathcal{C}}(Y), \mathbb{E}\left(\mathbf{X}_{\mathcal{C}}|S(Y)\right)\right)\left[\mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}}\right)\right]^{-1}\mathrm{Cov}\left(\alpha_{j|\mathcal{C}}(Y), \mathbb{E}\left(\mathbf{X}_{\mathcal{C}}|S(Y)\right)\right)^T \\
= \; & \mathrm{Cov}\left(\alpha_{j|\mathcal{C}}(Y), \mathbb{E}\left(\mathbf{X}_{\mathcal{C}}|S(Y)\right)\right)\left(\mathbf{T}_1+\mathbf{T}_2\right)^{-1}\mathrm{Cov}\left(\alpha_{j|\mathcal{C}}(Y), \mathbb{E}\left(\mathbf{X}_{\mathcal{C}}|S(Y)\right)\right)^T \\
\leq \; & \frac{\mathrm{Cov}\left(\alpha_{j|\mathcal{C}}(Y), \mathbb{E}\left(\mathbf{X}_{\mathcal{C}}|S(Y)\right)\right)\mathbf{T}_1^{-1}\mathrm{Cov}\left(\alpha_{j|\mathcal{C}}(Y), \mathbb{E}\left(\mathbf{X}_{\mathcal{C}}|S(Y)\right)\right)^T}{1+\lambda_{\min}\left(\mathbf{T}_1^{-\frac{1}{2}}\mathbf{T}_2\mathbf{T}_1^{-\frac{1}{2}}\right)} \\
\leq \; & \frac{\tau_{\max}}{\tau_{\min}+\tau_{\max}}\mathrm{Var}\left(\alpha_{j|\mathcal{C}}(Y)\right).
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
& \frac{\mathrm{Var}\left(M_j\right) - \mathrm{Cov}\left(M_j, \mathbf{X}_{\mathcal{C}}\right)\left[\mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}}\right)\right]^{-1}\mathrm{Cov}\left(M_j, \mathbf{X}_{\mathcal{C}}\right)^T}{V_j} \\
\geq \; & \frac{1}{\tau_{\max}}\frac{\tau_{\min}}{\tau_{\min}+\tau_{\max}}\mathrm{Var}\left(\alpha_{j|\mathcal{C}}(Y)\right) \\
\geq \; & \frac{1}{\tau_{\max}}\left(\frac{\tau_{\min}}{\tau_{\max}+\tau_{\min}}\right)^3\left(\frac{\tau_{\min}}{\tau_{\max}}\right)^3\xi n^{-\kappa},
\end{aligned}
$$

where $\xi_1 = \frac{1}{\tau_{\max}}\left(\frac{\tau_{\min}}{\tau_{\max}+\tau_{\min}}\right)^3\left(\frac{\tau_{\min}}{\tau_{\max}}\right)^3\xi$. □

PROOF OF LEMMA 2. Since $\mathbf{X}_{\mathcal{C}}$ follows either a multivariate normal distribution or a finite mixture of multivariate normal distributions, under Condition 1, one can show that for $i = 1, 2, \ldots, q$ and any $\epsilon > 0$ there exist constant C_1 and C_2 such that

$$
\Pr\left(\max_{\mathcal{C}\subset\{1,2,\ldots,p\}}\left|\widehat{\lambda}_i^{\mathcal{C}} - \lambda_i^{\mathcal{C}}\right| > \epsilon\right) \leq 2p(p+1)C_1\exp\left(-C_2 n\frac{\tau_{\min}^2\epsilon^2}{16p^2}\right)
$$

following similar arguments in the proof of Lemma 2 in Zhong et al. (2012). Since $\Omega_{\mathcal{C}} = W_{\mathcal{C}} + B_{\mathcal{C}}$, where $\Omega_{\mathcal{C}} = \mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}}\right)$ and $W_{\mathcal{C}} = \mathbb{E}\left(\mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}}|S(Y)\right)\right) = \mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}}|S(Y)\right)$ under model (2.2), $\lambda_{\max}\left(\Omega_{\mathcal{C}}\right) \leq \tau_{\max}$ and $\lambda_{\min}\left(W_{\mathcal{C}}\right) \geq \tau_{\min}$ Thus,

$$
\lambda_1^{\mathcal{C}} = \max_{||\boldsymbol{\eta}||=1}\frac{\boldsymbol{\eta}^T B_{\mathcal{C}}\boldsymbol{\eta}}{\boldsymbol{\eta}^T\Omega_{\mathcal{C}}\boldsymbol{\eta}} = 1 - \min_{||\boldsymbol{\eta}||=1}\frac{\boldsymbol{\eta}^T W_{\mathcal{C}}\boldsymbol{\eta}}{\boldsymbol{\eta}^T\Omega_{\mathcal{C}}\boldsymbol{\eta}} \leq 1 - \frac{\min_{||\boldsymbol{\eta}||=1}\boldsymbol{\eta}^T W_{\mathcal{C}}\boldsymbol{\eta}}{\max_{||\boldsymbol{\eta}||=1}\boldsymbol{\eta}^T\Omega_{\mathcal{C}}\boldsymbol{\eta}} \leq 1 - \frac{\tau_{\min}}{\tau_{\max}},
$$

and

$$1-\lambda_i^{\mathcal{C}} \geq 1-\lambda_1 \geq \frac{\tau_{\min}}{\tau_{\max}}, \text{ for } i=1,2,\ldots,q.$$

Therefore, for $i=1,2,\ldots,q$ and $0<\epsilon<1$, there exist constant C_1 and C_2 such that

$$\begin{aligned}
& \Pr\left(\max_{\mathcal{C}\subset\{1,2,\ldots,p\}} \left|\log\left(1-\widehat{\lambda}_i^{\mathcal{C}}\right)-\log\left(1-\lambda_i^{\mathcal{C}}\right)\right|>\epsilon\right) \\
\leq\ & \Pr\left(\max_{\mathcal{C}\subset\{1,2,\ldots,p\}} \frac{\left|\widehat{\lambda}_i^{\mathcal{C}}-\lambda_i^{\mathcal{C}}\right|}{1-\lambda_i^{\mathcal{C}}}>\frac{\epsilon}{2}\right) \\
\leq\ & \Pr\left(\max_{\mathcal{C}\subset\{1,2,\ldots,p\}} \left|\widehat{\lambda}_i^{\mathcal{C}}-\lambda_i^{\mathcal{C}}\right|>\frac{\tau_{\min}}{2\tau_{\max}}\epsilon\right) \\
\leq\ & 2p(p+1)C_1\exp\left(-C_2 n\frac{\tau_{\min}^4\epsilon^2}{64\tau_{\max}^2p^2}\right).
\end{aligned}$$

as $n\to\infty$. □

Proof of Theorem 1. Let $R_{\mathcal{C}}=\sum_{i=1}^q\log\left(1-\widehat{\lambda}_i^{\mathcal{C}}\right)-\sum_{i=1}^q\log\left(1-\lambda_i^{\mathcal{C}}\right)$. Then, according to Lemma 2, for $0<\epsilon<1$, there exist constant C_1 and C_2 such that

$$\Pr\left(\max_{\mathcal{C}\subset\{1,2,\ldots,p\}}|R_{\mathcal{C}}|>\epsilon\right)\leq 2p(p+1)qC_1\exp\left(-C_2n\frac{\tau_{\min}^4\epsilon^2}{64\tau_{\max}^2p^2q^2}\right).$$

Under Condition 2, $p=o\left(n^{\rho}\right)$ with $2\rho+2\kappa<1$, and for any positive constant C,

$$\begin{aligned}
& \Pr\left(\max_{\mathcal{C}\subset\{1,2,\ldots,p\}}|R_{\mathcal{C}}|>Cn^{-\kappa}\right) \\
\leq\ & 2p(p+1)qC_1\exp\left(-C_2n^{1-2\kappa-2\rho}\frac{\tau_{\min}^4C^2}{64\tau_{\max}^2q^2}\right)\to 0
\end{aligned}$$

as $n\to\infty$. For $j\notin\mathcal{C}$ and $d=|\mathcal{C}|$,

$$\begin{aligned}
\widehat{D}_{j|\mathcal{C}} &= -\sum_{i=1}^q\log\left(1-\widehat{\lambda}_i^{d+1}\right)+\sum_{i=1}^q\log\left(1-\widehat{\lambda}_i^{d}\right) \\
&= -\sum_{i=1}^q\log\left(1-\lambda_i^{d+1}\right)+\sum_{i=1}^q\log\left(1-\lambda_i^{d}\right)-R_{[\mathcal{C}\cup\{j\}]}+R_{\mathcal{C}} \\
&= \log\left(1+\frac{\operatorname{Var}\left(M_j\right)-\operatorname{Cov}\left(M_j,\mathbf{X}_{\mathcal{C}}\right)\left[\operatorname{Cov}\left(\mathbf{X}_{\mathcal{C}}\right)\right]^{-1}\operatorname{Cov}\left(M_j,\mathbf{X}_{\mathcal{C}}\right)^T}{V_j}\right) \\
&\quad -R_{[\mathcal{C}\cup\{j\}]}+R_{\mathcal{C}},
\end{aligned}$$

where $M_j = \mathbb{E}(X_j|\mathbf{X}_\mathcal{C}, S(Y))$, $V_j = \mathrm{Var}(X_j|\mathbf{X}_\mathcal{C}, S(Y))$, and V_j is a constant that does not depend on $\mathbf{X}_\mathcal{C}$ or $S(Y)$ under model (2.2).

When $\mathcal{C}^c \cap \mathcal{A} \neq \emptyset$, according to Lemma 1, there exist $\kappa > 0$ and $\xi_1 > 0$ such that

$$\max_{j \in \mathcal{C}^c \cap \mathcal{A}} \left[\frac{\mathrm{Var}(M_j) - \mathrm{Cov}(M_j, \mathbf{X}_\mathcal{C}) [\mathrm{Cov}(\mathbf{X}_\mathcal{C})]^{-1} \mathrm{Cov}(M_j, \mathbf{X}_\mathcal{C})^T}{V_j} \right] \geq \xi_1 n^{-\kappa},$$

Then, for sufficiently large n, there exists $j \in \mathcal{C}^c \cap \mathcal{A}$ such that

$$\log\left(1 + \frac{\mathrm{Var}(M_j) - \mathrm{Cov}(M_j, \mathbf{X}_\mathcal{C}) [\mathrm{Cov}(\mathbf{X}_\mathcal{C})]^{-1} \mathrm{Cov}(M_j, \mathbf{X}_\mathcal{C})^T}{V_j}\right) \geq \frac{\xi_1}{2} n^{-\kappa},$$

and

$$\widehat{D}_{j|\mathcal{C}} \geq \frac{\xi_1}{2} n^{-\kappa} - \left(\left| R_{[\mathcal{C} \cup \{j\}]} \right| + |R_\mathcal{C}| \right).$$

Let $c = \frac{\xi_1}{4}$. Since

$$\Pr\left(\max_{\mathcal{C} \subset \{1,2,\ldots,p\}} |R_\mathcal{C}| > \frac{c}{2} n^{-\kappa} \right) \to 0,$$

we have

$$\Pr\left(\min_{\mathcal{C}: \mathcal{C}^c \cap \mathcal{A} \neq \emptyset} \max_{j \in \mathcal{C}^c \cap \mathcal{A}} \widehat{D}_{j|\mathcal{C}} \geq c n^{-\kappa} \right) \to 1,$$

as $n \to \infty$.

When variable $\mathcal{C}^c \cap \mathcal{A} = \emptyset$, for $j \in \mathcal{C}^c \subset \mathcal{A}^c$, $M_j = \mathbb{E}(X_j|\mathbf{X}_\mathcal{C}, S(Y)) = \mathbb{E}(X_j|\mathbf{X}_\mathcal{C})$ is a linear combination of $\mathbf{X}_\mathcal{C}$ under model (2.2), and

$$\frac{\mathrm{Var}(M_j) - \mathrm{Cov}(M_j, \mathbf{X}_\mathcal{C}) [\mathrm{Cov}(\mathbf{X}_\mathcal{C})]^{-1} \mathrm{Cov}(M_j, \mathbf{X}_\mathcal{C})^T}{V_j} = 0.$$

Thus,

$$\widehat{D}_{j|\mathcal{C}} \leq \left(\left| R_{[\mathcal{C} \cup \{j\}]} \right| + |R_\mathcal{C}| \right),$$

and

$$\Pr\left(\max_{\mathcal{C}: \mathcal{C}^c \cap \mathcal{A} = \emptyset} \max_{j \in \mathcal{C}^c} \widehat{D}_{j|\mathcal{C}} \geq C n^{-\kappa} \right) \leq \Pr\left(\max_{\mathcal{C} \subset \{1,2,\ldots,p\}} |R_\mathcal{C}| \geq \frac{C}{2} n^{-\kappa} \right) \to 0$$

for any positive constant C as $n \to \infty$. □

A.3. Proof of Proposition 2 in Section 2.2. We start with a simple case to get some intuitions behind the proof. Suppose there are two different sets of minimally relevant predictors $\{X_1\}$ and $\{X_2\}$. Under model (2.7) and Condition 1, the support for the joint distribution of two predictors X_1 and X_2 is the entire 2-dimensional space $\mathbb{R}^2$. If $X_1 \perp\!\!\!\perp S(Y)|X_2$ and $X_2 \perp\!\!\!\perp S(Y)|X_1$, then we have $\Pr(S(Y)|X_1 = x_1, X_2 = x_2) = \Pr(S(Y)|X_2 = x_2) = \Pr(S(Y)|X_1 = x_1)$ for any $x_1, x_2 \in \mathbb{R}$. Thus, both $\Pr(S(Y)|X_2)$ and $\Pr(S(Y)|X_1)$ are constants, and $X_1 \perp\!\!\!\perp S(Y)$ and $X_2 \perp\!\!\!\perp S(Y)$, which is contradictory to the assumption that $\{X_1\}$ and $\{X_2\}$ are minimally relevant.

Proof of Proposition 2. Suppose there are two different sets of minimally relevant predictors indexed by $\mathcal{B}$ and $\mathcal{C}$, respectively. Define $\mathcal{A}_1 = \mathcal{B} \cap \mathcal{C}$, $\mathcal{A}_2 = \mathcal{B} \cap \mathcal{C}^c$ and $\mathcal{A}_3 = \mathcal{B}^c \cap \mathcal{C}$. Then, we must have $\mathcal{A}_2 \neq \emptyset$ and $\mathcal{A}_3 \neq \emptyset$. The conditional distribution of $\mathbf{X}_{\mathcal{A}_2}$ given $\mathbf{X}_{\mathcal{A}_1}$ and slice S_h can be written as

$$\mathbf{X}_{\mathcal{A}_2}|\mathbf{X}_{\mathcal{A}_1}, Y \in S_h \sim N\left(\alpha^{(h)}_{\mathcal{A}_2|\mathcal{A}_1} + \mathbf{B}^{(h)}_{\mathcal{A}_2|\mathcal{A}_1}\mathbf{X}_{\mathcal{A}_1}, \Sigma_2^{(h)} = \Sigma^{(h)}_{\mathcal{A}_2|\mathcal{A}_1}\right).$$

Let $\alpha^{(h)} = \alpha^{(h)}_{\mathcal{A}_2|\mathcal{A}_1}$ and $\mathbf{B}^{(h)} = \mathbf{B}^{(h)}_{\mathcal{A}_2|\mathcal{A}_1} = \left(\mathbf{B}_j^{(h)}\right)_{j\in\mathcal{A}_1}$. Since $\mathcal{B}$ is minimally relevant, at least one of $\alpha^{(h)}$, $\mathbf{B}_j^{(h)}$ $(j \in \mathcal{A}_1)$ and $\Sigma_2^{(h)}$ has to be different across slices. The conditional distribution of $\mathbf{X}_{\mathcal{A}_3}$ given $\mathbf{X}_{\mathcal{A}_1\cup\mathcal{A}_2} = \mathbf{X}_{\mathcal{B}}$ is the same across slices,

$$\mathbf{X}_{\mathcal{A}_3}|\mathbf{X}_{\mathcal{B}}, Y \in S_h \sim N\left(a + \mathbf{b}\mathbf{X}_{\mathcal{A}_1} + \mathbf{c}\mathbf{X}_{\mathcal{A}_2}, \Sigma_3 = \Sigma_{\mathcal{B}}\right).$$

The conditional distribution of $\mathbf{X}_{\mathcal{A}_3}$ given $\mathbf{X}_{\mathcal{A}_1}$ and slice S_h is

$$\mathbf{X}_{\mathcal{A}_3}|\mathbf{X}_{\mathcal{A}_1}, Y \in S_h \sim N\left(a + \mathbf{c}\alpha^{(h)} + \left(\mathbf{b} + \mathbf{c}\mathbf{B}^{(h)}\right)\mathbf{X}_{\mathcal{A}_1}, \Sigma_3 + \mathbf{c}\Sigma_2^{(h)}\mathbf{c}^T\right),$$

Since $\mathcal{A}_3 \in \mathcal{C}$ and $\mathcal{C}$ is minimally relevant, at least one of $\mathbf{c}\alpha^{(h)}$, $\mathbf{c}\mathbf{B}_j^{(h)}$ $(j \in \mathcal{A}_1)$ and $\mathbf{M}^{(h)} = \mathbf{c}\Sigma_2^{(h)}\mathbf{c}^T$ has to be different across slices. Given $\mathbf{X}_{\mathcal{A}_1\cup\mathcal{A}_3} = \mathbf{X}_{\mathcal{C}}$ and slice S_h, the conditional distribution of $\mathbf{c}\mathbf{X}_{\mathcal{A}_2}$ is

$$\mathbf{c}\mathbf{X}_{\mathcal{A}_2}|\mathbf{X}_{\mathcal{C}}, Y \in S_h \sim N\left(\gamma^{(h)} + \mathbf{D}^{(h)}\mathbf{X}_{\mathcal{A}_1} + \mathbf{C}^{(h)}\mathbf{X}_{\mathcal{A}_3}, \Sigma^{(h)}\right),$$

where $\mathbf{C}^{(h)} = \mathbf{M}^{(h)}\left(\Sigma_3 + \mathbf{M}^{(h)}\right)^{-1}$, $\gamma^{(h)} = -\mathbf{C}^{(h)}a + \left(\mathbf{I}_{|\mathcal{A}_3|} - \mathbf{C}^{(h)}\right)\mathbf{c}\alpha^{(h)}$, $\mathbf{D}^{(h)} = -\mathbf{C}^{(h)}\mathbf{b} + \left(\mathbf{I}_{|\mathcal{A}_3|} - \mathbf{C}^{(h)}\right)\mathbf{c}\mathbf{B}^{(h)}$, $\Sigma^{(h)} = \mathbf{M}^{(h)} - \mathbf{M}^{(h)}\left(\Sigma_3 + \mathbf{M}^{(h)}\right)^{-1}\mathbf{M}^{(h)}$.

First, if $\mathbf{M}^{(h)}$ $(h = 1, 2, \ldots, H)$ are different across slices. Then $\mathbf{N}^{(h)} = \Sigma_3^{-\frac{1}{2}}\mathbf{M}^{(h)}\Sigma_3^{-\frac{1}{2}} = \Gamma^{(h)}\Lambda^{(h)}\left[\Gamma^h\right]^{-1}$ are different across slices (note that under

Condition 1, $\Sigma_3^{\frac{1}{2}}$ is invertible). We have

$$\begin{aligned}\Sigma_3^{-\frac{1}{2}}\Sigma^{(h)}\Sigma_3^{-\frac{1}{2}} &= \mathbf{N}^{(h)} - \mathbf{N}^{(h)}\left(\mathbf{I}_{|\mathcal{A}_3|} + \mathbf{N}^{(h)}\right)^{-1}\mathbf{N}^{(h)} \\ &= \Gamma^{(h)}\Lambda^{(h)}\left(\mathbf{I}_{|\mathcal{A}_3|} + \Lambda^{(h)}\right)^{-1}\left[\Gamma^h\right]^{-1}.\end{aligned} \tag{A.6}$$

Thus, $\Sigma_3^{-\frac{1}{2}}\Sigma^{(h)}\Sigma_3^{-\frac{1}{2}}$ and $\Sigma^{(h)}$ are different across slices.

Second, if $\mathbf{M}^{(h)} = \mathbf{M}$ $(h = 1, 2, \ldots, H)$ are the same across slices. Then at least one of $\mathbf{c}\alpha^{(h)}$ and $\mathbf{cB}_j^{(h)}$ $(j \in \mathcal{A}_1)$ has to be different across slices. Without loss of generality, assume $\mathbf{c}\alpha^{(h)}$ are different across slices, that is, $\operatorname{trace}(\operatorname{Var}(\mathbf{c}\alpha(Y))) > 0$, where $\alpha(Y) = \alpha^{(h)}$ when $Y \in S_h$. Under Condition 1, $\lambda_{\max}(\mathbf{N}) \le \frac{\tau_{\max}}{\tau_{\min}}$, $\lambda_{\min}(\Sigma_3) \ge \tau_{\min}$ and $\lambda_{\min}\left(\Sigma_3^{-1}\right) \ge \frac{1}{\tau_{\max}}$. Then,

$$\mathbf{C}^{(h)} = \mathbf{C} = \mathbf{I}_{\mathcal{A}_3} - \mathbf{M}(\Sigma_3 + \mathbf{M})^{-1} = \Sigma_3^{\frac{1}{2}}\left(\mathbf{I}_{\mathcal{A}_3} - \mathbf{N}(\mathbf{I}_{\mathcal{A}_3} + \mathbf{N})^{-1}\right)\Sigma_3^{-\frac{1}{2}},$$

and

$$\begin{aligned}&\operatorname{trace}(\operatorname{Var}(\gamma(Y))) = \operatorname{trace}\left(\mathbf{C}\operatorname{Var}(c\alpha(Y))\mathbf{C}^T\right) \\ \ge\ & \lambda_{\min}\left(\Sigma_3^{-1}\right)\lambda_{\min}(\Sigma_3)\,\lambda_{\min}^2\left(\mathbf{I}_{\mathcal{A}_3} - \mathbf{N}(\mathbf{I}_{\mathcal{A}_3} + \mathbf{N})^{-1}\right)\operatorname{trace}(\operatorname{Var}(\mathbf{c}\alpha(Y))) \\ \ge\ & \left(\frac{\tau_{\max}}{\tau_{\max} + \tau_{\min}}\right)^2\left(\frac{\tau_{\max}}{\tau_{\min}}\right)\operatorname{trace}(\operatorname{Var}(\mathbf{c}\alpha(Y))) > 0,\end{aligned}$$

where $\gamma(Y) = \gamma^{(h)}$ when $Y \in S_h$. Thus, $\gamma^{(h)}$ are different across slices.

Therefore, given $\mathbf{X}_{\mathcal{C}}$ and slice S_h, the conditional distribution of $\mathbf{cX}_{\mathcal{A}_2}$ depends on slice S_h, which is contradictory with the previous assumption that $\mathcal{A}_2 \in \mathcal{C}^c$ and the conditional distribution of $\mathbf{X}_{\mathcal{A}_2}$ given $\mathbf{X}_{\mathcal{C}}$ is the same across slices. So we must have $\mathcal{B} = \mathcal{C}$. □

A.4. Proof on Properties of Augmented Likelihood Ratio Test Statistic in Section 2.2. Properties of the likelihood-ratio test statistic proposed in Section 2.2 are summarized in the following proposition:

PROPOSITION 5. *Given the current set of selected predictors indexed by $\mathcal{C}$ with dimension $|\mathcal{C}| = d$ and another predictor indexed by $j \notin \mathcal{C}$, the scaled log-likelihood-ratio test statistic for testing*

$$H_0 : \mathcal{A} = \mathcal{C} \text{ v.s. } H_1 : \mathcal{A} = \mathcal{C} \cup \{j\},$$

under the augmented model (2.7) can be written as

$$\widehat{D}^*_{j|\mathcal{C}} = \log \widehat{\sigma}^2_{j|\mathcal{C}} - \sum_{h=1}^{H} \frac{n_h}{n} \log\left[\widehat{\sigma}^{(h)}_{j|\mathcal{C}}\right]^2, \tag{A.7}$$

where $\left[\widehat{\sigma}^{(h)}_{j|\mathcal{C}}\right]^2$ *is the estimated variance by regressing* X_j *on* $\mathbf{X}_\mathcal{C}$ *in slice* S_h*, and* $\widehat{\sigma}^2_{j|\mathcal{C}}$ *is the estimated variance by regressing* X_j *on* $\mathbf{X}_\mathcal{C}$ *using all the observations.*
(a) For any fixed slicing scheme and $\mathcal{A} \subset \mathcal{C}$*,*

$$\widehat{D}^*_{j|\mathcal{C}} \sim \log\left(1+\frac{Q_0}{\sum_{h=1}^H Q_h}\right) - \sum_{h=1}^H \frac{n_h}{n}\log\left(\frac{Q_h/n_h}{\sum_{h=1}^H Q_h/n}\right) \xrightarrow{d} \chi^2((H-1)(d+2)),$$

where $Q_0 \sim \chi^2((H-1)(d+1))$ *and* $Q_h \sim \chi^2(n_h-(d+1))$ *(*$1 \le h \le H$*) are mutually independent.*
(b) Under the same condition as in (a),

$$\left(n\widehat{D}^*_{j|\mathcal{C}}\right)_{j\in\mathcal{C}^c} \xrightarrow{d} \left(\sum_{i=1}^{(H-1)(d+1)} z^2_{ij} + \sum_{i=1}^{H-1}\widetilde{z}^2_{ij}\right)_{j\in\mathcal{C}^c},$$

where $\mathbf{z}_i$*'s and* $\widetilde{\mathbf{z}}_i$*'s are mutually independent with*

$$\mathbf{z}_i = (z_{ij})_{j\in\mathcal{C}^c} \sim MVN\left(\mathbf{0}, \left[Corr\left(X_j, X_k|\mathbf{X}_\mathcal{C}\right)\right]_{j,k\in\mathcal{C}^c}\right),$$

and

$$\widetilde{\mathbf{z}}_i = (\widetilde{z}_{ij})_{j\in\mathcal{C}^c} \sim MVN\left(\mathbf{0}, \left[Corr^2\left(X_j, X_k|\mathbf{X}_\mathcal{C}\right)\right]_{j,k\in\mathcal{C}^c}\right).$$

(c) For any fixed slicing scheme, as $n\to\infty$*,*

$$\begin{aligned}&\widehat{D}^*_{j|\mathcal{C}} \xrightarrow{a.s.} D^*_{j|\mathcal{C}}\\ =\ &\log\left(1+\frac{Var(M_j)-Cov\left(M_j,\mathbf{X}_\mathcal{C}\right)\left[Cov\left(\mathbf{X}_\mathcal{C}\right)\right]^{-1}Cov\left(M_j,\mathbf{X}_\mathcal{C}\right)^T}{\mathbb{E}\left(V_j\right)}\right)\\ &+\log\mathbb{E}\left(V_j\right)-\mathbb{E}\log\left(V_j\right)\end{aligned}$$

where $M_j = \mathbb{E}\left(X_j|\mathbf{X}_\mathcal{C}, S(Y)\right)$*,* $V_j = Var\left(X_j|\mathbf{X}_\mathcal{C}, S(Y)\right)$ *and* $S(Y)=h$ *when* $Y\in S_h$ *(*$1\le h\le H$*). Furthermore,*

$$\begin{aligned}D^*_{j|\mathcal{C}} = 0 \quad \textit{iff} \quad &\mathbb{E}\left(X_j|\mathbf{X}_\mathcal{C}, Y\in S_h\right) = \mathbb{E}\left(X_j|\mathbf{X}_\mathcal{C}\right) \textit{ and}\\ &Var\left(X_j|\mathbf{X}_\mathcal{C}, Y\in S_h\right) = Var\left(X_j|\mathbf{X}_\mathcal{C}\right),\end{aligned}$$

for $1\le h\le H$*.*

Proof of Proposition 5. (a) Let $\widetilde{\mathbf{x}}_\mathcal{C} = [\mathbf{1}_n, \mathbf{x}_\mathcal{C}]$, where $\mathbf{1}_n$ is a n-dimensional column vector of 1 and $\mathbf{I}_n$ is a n by n identity matrix. We denote $P_0 = \widetilde{\mathbf{x}}_\mathcal{C}\left(\widetilde{\mathbf{x}}^T_\mathcal{C}\widetilde{\mathbf{x}}_\mathcal{C}\right)^{-1}\widetilde{\mathbf{x}}^T_\mathcal{C}$,

$$P_h = diag\left(0,\ldots,\widetilde{\mathbf{x}}^{(h)}_\mathcal{C}\left(\widetilde{\mathbf{x}}^{(h)T}_\mathcal{C}\widetilde{\mathbf{x}}^{(h)}_\mathcal{C}\right)^{-1}\widetilde{\mathbf{x}}^{(h)T}_\mathcal{C},\ldots,0\right),$$

$R_h = diag(0, \ldots, \mathbf{I}_{n_h}, \ldots, 0) - P_h$ and $R_0 = \mathbf{I}_n - P_0$. Then,

$$\left[\widehat{\sigma}^{(h)}_{j|\mathcal{C}}\right]^2 = \frac{1}{n_h}\mathbf{x}_j^T R_h \mathbf{x}_j = \frac{1}{n_h}\sigma^2_{j|\mathcal{C}} Q_h = \frac{1}{n}\frac{\sigma^2_{j|\mathcal{C}} Q_h}{s_h}, \text{ for } h = 1, 2, \ldots, H$$

and

$$\begin{aligned}\widehat{\sigma}^2_{j|\mathcal{C}} &= \frac{1}{n}\mathbf{x}_j^T R_0 \mathbf{x}_j = \frac{1}{n}\mathbf{x}_j^T \left(\sum_{h=1}^H R_h\right)\mathbf{x}_j + \frac{1}{n}\mathbf{x}_j^T\left(\sum_{h=1}^H P_h - P_0\right)\mathbf{x}_j \\ &= \frac{1}{n}\sum_{h=1}^H \sigma^2_{j|\mathcal{C}} Q_h + \frac{1}{n}\sigma^2_{j|\mathcal{C}} Q_0,\end{aligned}$$

where $\sigma^2_{j|\mathcal{C}}$ is the conditional variance of X_j given $\mathbf{X}_\mathcal{C}$ for $j \in \mathcal{C}^c$ and $\mathcal{A} \subset \mathcal{C}$.

The augmented likelihood-ratio test statistic in (A.7) can be written as

$$\begin{aligned}\widehat{D}^*_{j|\mathcal{C}} &= -\left(\sum_{h=1}^H s_h \log\left[\widehat{\sigma}^{(h)}_{j|\mathcal{C}}\right]^2 - \log\widehat{\sigma}^2_{j|\mathcal{C}}\right) \\ &= \left(\log\left(\frac{1}{n}\sum_{h=1}^H \sigma^2_{j|\mathcal{C}} Q_h + \frac{1}{n}\sigma^2_{j|\mathcal{C}} Q_0\right) - \sum_{h=1}^H s_h \log\left(\frac{1}{n}\frac{\sigma^2_{j|\mathcal{C}} Q_h}{s_h}\right)\right) \\ &= \left(\log\left(1 + \frac{Q_0}{\sum_{h=1}^H Q_h}\right) - \sum_{h=1}^H s_h \log\left(\frac{Q_h/s_h}{\sum_{h=1}^H Q_h}\right)\right).\end{aligned}$$

Note that both $\left(\sum_{h=1}^H P_h - P_0\right)$ and R_h's are orthogonal to $\tilde{\mathbf{x}}_\mathcal{C}$. Given that $j \in \mathcal{C}^c$, according to *Cochran's theorem*, $Q_0 = \mathbf{x}_j^T\left(\sum_{h=1}^H P_h - P_0\right)\mathbf{x}_j/\sigma^2_{j|\mathcal{C}}$ and $Q_h = \mathbf{x}_j^T R_h \mathbf{x}_j/\sigma^2_{j|\mathcal{C}}$ are independent, and

$Q_0 \sim \chi^2((H-1)(d+1))$, and $Q_h \sim \chi^2(n_h-(d+1))$, for $h = 1, 2, \ldots, H$, and $d = |\mathcal{C}|$.

Let $n'_h = n_h - (d+1)$ and $n' = n - H(d+1)$. For any fixed slicing scheme, as $n \to \infty$, $n'_h \to \infty$, $n' \to \infty$ and $n'_h/n_h \to 1$, $n'/n \to 1$ given that $d/n \to 0$. Since

$$\frac{Q_0}{n'} \xrightarrow{P} 0, \ \frac{\sum_{h=1}^H Q_h}{n'} \xrightarrow{P} 1, \text{ and } \frac{Q_0}{\sum_{h=1}^H Q_h} = \frac{Q_0/n'}{\sum_{h=1}^H Q_h/n'} \xrightarrow{P} 0,$$

we have

$$U_j \equiv n\log\left(1 + \frac{Q_0}{\sum_{h=1}^H Q_h}\right) \xrightarrow{P} \frac{n}{n'}\frac{Q_0}{\sum_{h=1}^H Q_h/n'} \xrightarrow{P} Q_0 \sim \chi^2((H-1)(d+1)).$$

Since

$$\frac{Q_h/s_h}{\sum_{h=1}^H Q_h} = \frac{n'_h/n_h}{n'/n}\frac{Q_h/n'_h}{\sum_{h=1}^H Q_h/n'} \xrightarrow{P} 1, \text{ for } h = 1, 2, \ldots, H,$$

and

$$\frac{\sum_{h=1}^H Q_h}{n} = \frac{n'}{n}\frac{\sum_{h=1}^H Q_h}{n'} \xrightarrow{P} 1,$$

we have

$$\begin{aligned}
\widetilde{U}_j &\equiv -n\sum_{h=1}^H s_h \log\left(\frac{Q_h/s_h}{\sum_{h=1}^H Q_h}\right) = -n\sum_{h=1}^H s_h \log\left(1 + \frac{Q_h/s_h}{\sum_{h=1}^H Q_h} - 1\right) \\
&\simeq -n\sum_{h=1}^H s_h\left(\frac{Q_h/s_h}{\sum_{h=1}^H Q_h} - 1\right) + \frac{1}{2}\sum_{h=1}^H n_h\left(\frac{Q_h/s_h}{\sum_{h=1}^H Q_h} - 1\right)^2 \\
&= \frac{1}{2}\sum_{h=1}^H n_h\left(\frac{Q_h/n_h}{\sum_{h=1}^H Q_h/n} - 1\right)^2 = \frac{1}{2}\frac{\sum_{h=1}^H n_h\left(Q_h/n_h - \sum_{h=1}^H Q_h/n\right)^2}{\left(\sum_{h=1}^H Q_h/n\right)^2} \\
&\simeq \frac{1}{2}\sum_{h=1}^H n_h\left(Q_h/n_h - \sum_{h=1}^H Q_h/n\right)^2 \\
&= \sum_{h=1}^H\left(\frac{Q_h}{\sqrt{2n_h}}\right)^2 - \left(\sum_{h=1}^H \sqrt{\frac{n_h}{n}}\frac{Q_h}{\sqrt{2n_h}}\right)^2.
\end{aligned}$$

Let $q_h = \frac{Q_h}{\sqrt{2n_h}}$, for $h = 1, 2, \ldots, H$. Then

$$q_h = \sqrt{\frac{n'_h}{n_h}}\frac{\sum_{i=1}^{n'_h}\left[z_{ij}^{(h)}\right]^2}{\sqrt{2n'_h}},$$

where $z_{ij}^{(h)} \sim N(0,1)$ independently for $i = 1, 2, \ldots, n'_h$ and $h = 1, 2, \ldots, H$. Thus, according to central limit theorem, $q_h \xrightarrow{d} N(0,1)$, for $h = 1, 2, \ldots, H$, and let $\mathbf{q} = (q_1, \ldots, q_H)^T$. Then,

$$\widetilde{U}_j \simeq \sum_{h=1}^H q_h^2 - \left(\sum_{h=1}^H \sqrt{\frac{n_h}{n}} q_h\right)^2 = \mathbf{q}^T\left(\mathbf{I}_H - JJ^T\right)\mathbf{q}$$

where $J = \left(\sqrt{n_1/n}, \ldots, \sqrt{n_H/n}\right)^T$ and $J^TJ = 1$. According to Cochran's theorem, $\widetilde{U}_j$ is asymptotically $\chi^2(H-1)$. Since Q_0 is independent of Q_h $(h = 1, 2, \ldots, H)$, U_j is asymptotically independent of $\widetilde{U}_j$ and

$$n\widehat{D}^*_{j|\mathcal{C}} = U_j + \widetilde{U}_j \xrightarrow{d} \chi^2((H-1)(d+2)).$$

To prove (b), for any $j \in \mathcal{C}^c$, we denote

$$U_j \simeq Q_j^{(0)} = \frac{\mathbf{x}_j^T \left(\sum_{h=1}^H P_h - P_0\right) \mathbf{x}_j}{\sigma_{j|\mathcal{C}}^2} = \sum_{i=1}^{(H-1)(d+1)} z_{ij}^2,$$

where $z_{ij} \sim N(0,1)$ independently for the same j and $i = 1, 2, \ldots, (H-1)(d+1)$, and $\mathrm{Cov}(z_{ij}, z_{ij'}) = \mathrm{Corr}\left(X_j, X_{j'} | \mathbf{X}_\mathcal{C}\right)$ for $j' \neq j$. For any j and $h \in \{1, 2, \ldots, H\}$, we denote

$$Q_j^{(h)} = \frac{\mathbf{x}_j^T R_h \mathbf{x}_j}{\sigma_{j|\mathcal{C}}^2} = \sum_{i=1}^{n_h'} \left[z_{ij}^{(h)}\right]^2, \text{ and } q_j^{(h)} = \frac{Q_j^{(h)}}{\sqrt{2n_h}},$$

where $z_{ij}^{(h)} \sim N(0,1)$ independently for the same j and $i = 1, 2, \ldots, n_h'$, and $\mathrm{Cov}(z_{ij}^{(h)}, z_{ij'}^{(h)}) = \mathrm{Corr}\left(X_j, X_{j'} | \mathbf{X}_\mathcal{C}\right)$ for $j' \neq j$. Since $\mathrm{Cov}\left(\left[z_{ij}^{(h)}\right]^2, \left[z_{ij'}^{(h)}\right]^2\right) = 2\mathrm{Corr}^2\left(X_j, X_{j'} | \mathbf{X}_\mathcal{C}\right)$, for $h = 1, 2, \ldots . H$ and $j \neq j'$, we have

$$\begin{aligned} \mathrm{Cov}\left(q_j^{(h)}, q_{j'}^{(h)}\right) &= \frac{1}{2n_h} \mathrm{Cov}\left(Q_j^{(h)}, Q_{j'}^{(h)}\right) \\ &= \frac{n_h'}{n_h} \mathrm{Corr}^2\left(X_j, X_{j'} | \mathbf{X}_\mathcal{C}\right) \to \mathrm{Corr}^2\left(X_j, X_{j'} | \mathbf{X}_\mathcal{C}\right). \end{aligned}$$

Hence

$$\widetilde{U}_j \simeq \sum_{h=1}^H \left[q_j^{(h)}\right]^2 - \left(\sum_{h=1}^H \sqrt{\frac{n_h}{n}} \left[q_j^{(h)}\right]\right)^2 \simeq \sum_{i=1}^{(H-1)} \tilde{z}_{ij}^2,$$

where $\tilde{z}_{ij} \sim N(0,1)$ independent for the same j and $i = 1, 2, \ldots, (H-1)$, and for $j' \neq j$, $\mathrm{Cov}(\tilde{z}_{ij}, \tilde{z}_{ij'}) = \mathrm{Corr}^2\left(X_j, X_{j'} | \mathbf{X}_\mathcal{C}\right)$. Therefore,

$$\left(n\widehat{D}_{j|\mathcal{C}}^*\right)_{j \in \mathcal{C}^c} = \left(U_j + \widetilde{U}_j\right)_{j \in \mathcal{C}^c} \xrightarrow{d} \left(\sum_{i=1}^{(H-1)(d+1)} z_{ij}^2 + \sum_{i=1}^{(H-1)} \tilde{z}_{ij}^2\right)_{j \in \mathcal{C}^c}.$$

(c) For any fixed slicing scheme, as $n \to \infty$, $n_h \to \infty$ for $h = 1, 2, \ldots, H$. Under the normality assumption, as $n_h \to \infty$,

$$\begin{aligned} &\left[\widehat{\sigma}_{j|\mathcal{C}}^{(h)}\right]^2 \xrightarrow{P} \left[\sigma_{j|\mathcal{C}}^{(h)}\right]^2 \\ = \;& \mathrm{Var}\left(X_j | Y \in S_h\right) - \mathrm{Cov}\left(X_j, \mathbf{X}_\mathcal{C} | Y \in S_h\right) \left[\mathrm{Cov}\left(\mathbf{X}_\mathcal{C} | Y \in S_h\right)\right]^{-1} \times \\ & \mathrm{Cov}\left(X_j, \mathbf{X}_\mathcal{C} | Y \in S_h\right)^T \\ = \;& \mathrm{Var}\left(X_j | \mathbf{X}_\mathcal{C}, Y \in S_h\right), \end{aligned}$$

and as $n \to \infty$,

$$\begin{aligned} & \widehat{\sigma}^2_{j|\mathcal{C}} \xrightarrow{P} \sigma^2_{j|\mathcal{C}} \\ = \; & \mathrm{Var}\left(X_j\right) - \mathrm{Cov}\left(X_j, \mathbf{X}_\mathcal{C}\right) \left[\mathrm{Cov}\left(\mathbf{X}_\mathcal{C}\right)\right]^{-1} \mathrm{Cov}\left(X_j, \mathbf{X}_\mathcal{C}\right)^T \\ = \; & \mathrm{Var}\left(X_j\right) - \mathrm{Cov}\left(\mathbb{E}\left(X_j|\mathbf{X}_\mathcal{C}, S(Y)\right), \mathbf{X}_\mathcal{C}\right) \left[\mathrm{Cov}\left(\mathbf{X}_\mathcal{C}\right)\right]^{-1} \times \\ & \mathrm{Cov}\left(\mathbb{E}\left(X_j|\mathbf{X}_\mathcal{C}, S(Y)\right), \mathbf{X}_\mathcal{C}\right)^T. \end{aligned}$$

Let $M_j = \mathbb{E}\left(X_j|\mathbf{X}_\mathcal{C}, S(Y)\right)$ and $V_j = \mathrm{Var}\left(X_j|\mathbf{X}_\mathcal{C}, S(Y)\right)$. Then, $\mathrm{Var}\left(X_j\right) = \mathrm{Var}\left(M_j\right) + \mathbb{E}\left(V_j\right)$, and

$$\begin{aligned} \widehat{D}^*_{j|\mathcal{C}} \; = \; & \log \widehat{\sigma}^2_{j|\mathcal{C}} - \sum_{h=1}^{H} s_h \log \left[\widehat{\sigma}^{(h)}_{j|\mathcal{C}}\right]^2 \\ \xrightarrow{P} \; & \log \sigma^2_{j|\mathcal{C}} - \sum_{h=1}^{H} s_h \log \left[\sigma^{(h)}_{j|\mathcal{C}}\right]^2 \\ = \; & \log\left(\mathbb{E}\left(V_j\right) + \mathrm{Var}\left(M_j\right) - \mathrm{Cov}\left(M_j, \mathbf{X}_\mathcal{C}\right) \left[\mathrm{Cov}\left(\mathbf{X}_\mathcal{C}\right)\right]^{-1} \mathrm{Cov}\left(M_j, \mathbf{X}_\mathcal{C}\right)^T\right) \\ & - \sum_{h=1}^{H} s_h \log\left(\mathrm{Var}\left(X_j|\mathbf{X}_\mathcal{C}, Y \in S_h\right)\right). \end{aligned}$$

Since $\mathrm{Var}\left(X_j|\mathbf{X}_\mathcal{C}, Y \in S_h\right)$ is a constant that does not depend on $\mathbf{X}_\mathcal{C}$ under the normality assumption,

$$\mathbb{E}\log(V_j) = \sum_{h=1}^{H} s_h \log\left(\mathrm{Var}\left(X_j|\mathbf{X}_\mathcal{C}, Y \in S_h\right)\right),$$

and thus

$$\begin{aligned} \widehat{D}^*_{j|\mathcal{C}} \; \xrightarrow{P} \; & \log\left(\mathbb{E}\left(V_j\right) + \mathrm{Var}\left(M_j\right) - \mathrm{Cov}\left(M_j, \mathbf{X}_\mathcal{C}\right) \left[\mathrm{Cov}\left(\mathbf{X}_\mathcal{C}\right)\right]^{-1} \mathrm{Cov}\left(M_j, \mathbf{X}_\mathcal{C}\right)^T\right) \\ & -\mathbb{E}\log(V_j) \\ = \; & \log\left(1 + \frac{\mathrm{Var}\left(M_j\right) - \mathrm{Cov}\left(M_j, \mathbf{X}_\mathcal{C}\right) \left[\mathrm{Cov}\left(\mathbf{X}_\mathcal{C}\right)\right]^{-1} \mathrm{Cov}\left(M_j, \mathbf{X}_\mathcal{C}\right)^T}{\mathbb{E}\left(V_j\right)}\right) \\ & + \log\left(\mathbb{E}V_j\right) - \mathbb{E}\log\left(V_j\right). \end{aligned}$$

Note that

$$\mathrm{Var}\left(M_j\right) \geq \mathrm{Cov}\left(M_j, \mathbf{X}_\mathcal{C}\right) \left[\mathrm{Cov}\left(\mathbf{X}_\mathcal{C}\right)\right]^{-1} \mathrm{Cov}\left(M_j, \mathbf{X}_\mathcal{C}\right)^T$$

where the equality holds if and only if $M_j = \mathbb{E}\left(X_j|\mathbf{X}_\mathcal{C}, S(Y)\right)$ is a linear combination of $\mathbf{X}_\mathcal{C}$ that does not depend on $S(Y)$, that is, $M_j =$

$\mathbb{E}(X_j|\mathbf{X}_\mathcal{C}, S(Y)) = \mathbb{E}(X_j|\mathbf{X}_\mathcal{C})$ under the normality assumption. Furthermore, according to Jensen's inequality

$$\log(\mathbb{E}V_j) \geq \mathbb{E}\log(V_j),$$

where the equality holds if and only if $V_j = \mathrm{Var}(X_j|\mathbf{X}_\mathcal{C}, S(Y)) = EV_j$, or equivalently, $\mathrm{Var}(X_j|\mathbf{X}_\mathcal{C}, Y \in S_h)$ is a constant for $h = 1, 2, \ldots, H$. Combined with $M_j = \mathbb{E}(X_j|\mathbf{X}_\mathcal{C})$, $\mathrm{Var}(X_j|\mathbf{X}_\mathcal{C}) = \mathbb{E}(V_j|\mathbf{X}_\mathcal{C}) + \mathrm{Var}(M_j|\mathbf{X}_\mathcal{C}) = V_j = \mathrm{Var}(X_j|\mathbf{X}_\mathcal{C}, S(Y))$. □

A.5. Proof of Theorem 2 in Section 2.2. To prove Theorem 2, we will need the following lemma.

LEMMA 3. *Under the same condition as in Theorem 2, for* $0 < \epsilon < 1$, *there exist positive constants* C_1 *and* C_2 *such that*

$$Pr\left(\max_{\mathcal{C}\subset\{1,2,\ldots,p\}} \max_{j\in\mathcal{C}^c} \left|\log\widehat{\sigma}^2_{j|\mathcal{C}} - \log\sigma^2_{j|\mathcal{C}}\right| > \epsilon\right) \leq \frac{p(p+1)}{2} C_1 \exp\left(-C_2 n \frac{\epsilon^2}{p^2L^2}\right),$$

and

$$\begin{aligned} & Pr\left(\max_{\mathcal{C}\subset\{1,2,\ldots,p\}} \max_{j\in\mathcal{C}^c} \left|\sum_{h=1}^H s_h \log\left[\widehat{\sigma}^{(h)}_{j|\mathcal{C}}\right]^2 - \sum_{h=1}^H s_h \log\left[\sigma^{(h)}_{j|\mathcal{C}}\right]^2\right| > \epsilon\right) \\ \leq\ & \frac{Hp(p+1)}{2} C_1 \exp\left(-C_2 n \frac{\epsilon^2}{H^2p^2L^2}\right), \end{aligned}$$

where $L = \frac{4}{\tau_{min}}\left(3\left(\frac{\tau_{max}}{\tau_{min}}\right)^{3/2} + 1\right)$.

PROOF OF LEMMA 3. We denote $V_\mathcal{C} = \mathrm{Cov}(\mathbf{X}_\mathcal{C}) = (v_{j_1,j_2})_{j_1,j_2\in\mathcal{C}}$, $r_{j,\mathcal{C}} = \mathrm{Cov}(X_j, \mathbf{X}_\mathcal{C})^T$ and $v_{j,j} = \mathrm{Var}(X_j)$, and $\widehat{V}_\mathcal{C} = (\widehat{v}_{j_1,j_2})_{j_1,j_2\in\mathcal{C}}$, $\widehat{r}_{j,\mathcal{C}}$ and $\widehat{v}_{j,j}$ are the corresponding sample estimates. Then, $\sigma^2_{j|\mathcal{C}} = v_{j,j} - r^T_{j,\mathcal{C}}V^{-1}_\mathcal{C} r_{j,\mathcal{C}}$, $\widehat{\sigma}^2_{j|\mathcal{C}} = \widehat{v}_{j,j} - \widehat{r}^T_{j,\mathcal{C}}\widehat{V}^{-1}_\mathcal{C}\widehat{r}_{j,\mathcal{C}}$, and

$$\begin{aligned} & \left|\widehat{\sigma}^2_{j|\mathcal{C}} - \sigma^2_{j|\mathcal{C}}\right| \\ \leq\ & |\widehat{v}_{j,j} - v_{j,j}| + \left|\widehat{r}^T_{j,\mathcal{C}}\widehat{V}^{-1}_\mathcal{C}\widehat{r}_{j,\mathcal{C}} - \widehat{r}^T_{j,\mathcal{C}}V^{-1}_\mathcal{C}\widehat{r}_{j,\mathcal{C}}\right| + \left|\widehat{r}^T_{j,\mathcal{C}}V^{-1}_\mathcal{C}\widehat{r}_{j,\mathcal{C}} - r^T_{j,\mathcal{C}}V^{-1}_\mathcal{C} r_{j,\mathcal{C}}\right| \end{aligned}$$

Let $M = \max_{j_1,j_2\in\{1,2,\ldots,p\}} |\widehat{v}_{j_1,j_2} - v_{j_1,j_2}|$. We have

$$\max_{j\in\{1,2,\ldots,p\}} |\widehat{v}_{j,j} - v_{j,j}| \leq M,$$

$$\left|\widehat{r}_{j,\mathcal{C}}^T V_{\mathcal{C}}^{-1}\widehat{r}_{j,\mathcal{C}} - r_{j,\mathcal{C}}^T V_{\mathcal{C}}^{-1} r_{j,\mathcal{C}}\right| = \left|\left(\widehat{r}_{j,\mathcal{C}} - r_{j,\mathcal{C}}\right)^T V_{\mathcal{C}}^{-1}\left(\widehat{r}_{j,\mathcal{C}} + r_{j,\mathcal{C}}\right)\right|,$$

and

$$\begin{aligned}\left|\widehat{r}_{j,\mathcal{C}}^T \widehat{V}_{\mathcal{C}}^{-1}\widehat{r}_{j,\mathcal{C}} - \widehat{r}_{j,\mathcal{C}}^T V_{\mathcal{C}}^{-1}\widehat{r}_{j,\mathcal{C}}\right| &= \left|\left(\widehat{V}_{\mathcal{C}}^{-1}\widehat{r}_{j,\mathcal{C}}\right)^T\left(\widehat{V}_{\mathcal{C}} - V_{\mathcal{C}}\right)\left(V_{\mathcal{C}}^{-1}\widehat{r}_{j,\mathcal{C}}\right)\right| \\ &= ||\boldsymbol{\eta}_1||||\boldsymbol{\eta}_2||\left|\boldsymbol{\eta}_1^{*T}\left(\widehat{V}_{\mathcal{C}} - V_{\mathcal{C}}\right)\boldsymbol{\eta}_2^*\right|,\end{aligned}$$

where $\boldsymbol{\eta}_1 = \widehat{V}_{\mathcal{C}}^{-1}\widehat{r}_{j,\mathcal{C}}$, $\boldsymbol{\eta}_2 = V_{\mathcal{C}}^{-1}\widehat{r}_{j,\mathcal{C}}$, and $\boldsymbol{\eta}_1^* = \frac{\boldsymbol{\eta}_1}{||\boldsymbol{\eta}_1||}$, $\boldsymbol{\eta}_2^* = \frac{\boldsymbol{\eta}_2}{||\boldsymbol{\eta}_2||}$.

Since

$$\tau_{\min} \leq \min_{\mathcal{C}\subset\{1,2,\ldots,p\}} \lambda_{\min}\{V_{\mathcal{C}}\} \leq \max_{\mathcal{C}\subset\{1,2,\ldots,p\}} \lambda_{\max}\{V_{\mathcal{C}}\} \leq \tau_{\max},$$

using similar arguments to the proof of Lemma 1 in Wang (2009) with $p = o\left(n^{\rho}\right)$ and $2\rho + 2\kappa < 1$, we have

$$2^{-1}\tau_{\min} \leq \min_{\mathcal{C}\subset\{1,2,\ldots,p\}} \lambda_{\min}\{\widehat{V}_{\mathcal{C}}\} \leq \max_{\mathcal{C}\subset\{1,2,\ldots,p\}} \lambda_{\max}\{\widehat{V}_{\mathcal{C}}\} \leq 2\tau_{\max}.$$

Then, $||\widehat{V}_{\mathcal{C}}^{-1/2}\widehat{r}_{j,\mathcal{C}}||^2 = \widehat{r}_{j,\mathcal{C}}^T\widehat{V}_{\mathcal{C}}^{-1}\widehat{r}_{j,\mathcal{C}} \leq \widehat{v}_{j,j} \leq 2\tau_{\max}$, $||\widehat{V}_{\mathcal{C}}^{1/2}|| \leq \sqrt{2\tau_{\max}}$, and $||\widehat{V}_{\mathcal{C}}^{-1/2}|| \leq \sqrt{\frac{2}{\tau_{\min}}}$. Thus,

$$||\boldsymbol{\eta}_1|| = ||\widehat{V}_{\mathcal{C}}^{-1}\widehat{r}_{j,\mathcal{C}}|| \leq ||\widehat{V}_{\mathcal{C}}^{-1/2}||||\widehat{V}_{\mathcal{C}}^{-1/2}\widehat{r}_{j,\mathcal{C}}|| \leq 2\sqrt{\frac{\tau_{\max}}{\tau_{\min}}},$$

and

$$||\boldsymbol{\eta}_2|| = ||V_{\mathcal{C}}^{-1}\widehat{r}_{j,\mathcal{C}}|| \leq ||V_{\mathcal{C}}^{-1}||||\widehat{V}_{\mathcal{C}}^{1/2}||||\widehat{V}_{\mathcal{C}}^{-1/2}\widehat{r}_{j,\mathcal{C}}|| \leq \frac{2\tau_{\max}}{\tau_{\min}}.$$

Furthermore,

$$\begin{aligned}&\left|\boldsymbol{\eta}_1^{*T}\left(\widehat{V}_{\mathcal{C}} - V_{\mathcal{C}}\right)\boldsymbol{\eta}_2^*\right| \leq \max_{j_1,j_2\in\{1,2,\ldots,p\}} |\widehat{v}_{j_1,j_2} - v_{j_1,j_2}| \sum_{j_1,j_2} |\boldsymbol{\eta}_{j_1}^*||\boldsymbol{\eta}_{j_2}^*| \\ = \;& M\left(\sum_{j_1}|\boldsymbol{\eta}_{j_1}^*|\right)\left(\sum_{j_2}|\boldsymbol{\eta}_{j_2}^*|\right) \leq Mp.\end{aligned}$$

Hence,

$$\left|\widehat{r}_{j,\mathcal{C}}^T\widehat{V}_{\mathcal{C}}^{-1}\widehat{r}_{j,\mathcal{C}} - \widehat{r}_{j,\mathcal{C}}^T V_{\mathcal{C}}^{-1}\widehat{r}_{j,\mathcal{C}}\right| \leq 4\left(\frac{\tau_{\max}}{\tau_{\min}}\right)^{3/2} Mp.$$

Since $||\widehat{r}_{j,\mathcal{C}} - r_{j,\mathcal{C}}|| \leq \max_{j_1,j_2\in\{1,2,\ldots,p\}} |\widehat{v}_{j_1,j_2} - v_{j_1,j_2}|\sqrt{p} = M\sqrt{p}$, $||V_{\mathcal{C}}^{-1}\widehat{r}_{j,\mathcal{C}}|| \leq \frac{2\tau_{\max}}{\tau_{\min}}$, and $||V_{\mathcal{C}}^{-1}r_{j,\mathcal{C}}|| \leq \sqrt{\frac{\tau_{\max}}{\tau_{\min}}}$, we have

$$\begin{aligned}\left|\left(\widehat{r}_{j,\mathcal{C}} - r_{j,\mathcal{C}}\right)^T V_{\mathcal{C}}^{-1}\left(\widehat{r}_{j,\mathcal{C}} + r_{j,\mathcal{C}}\right)\right| &\leq ||\widehat{r}_{j,\mathcal{C}} - r_{j,\mathcal{C}}||\left(||V_{\mathcal{C}}^{-1}\widehat{r}_{j,\mathcal{C}}|| + ||V_{\mathcal{C}}^{-1}r_{j,\mathcal{C}}||\right) \\ &\leq \left(\frac{2\tau_{\max}}{\tau_{\min}} + \sqrt{\frac{\tau_{\max}}{\tau_{\min}}}\right)M\sqrt{p},\end{aligned}$$

and

$$\begin{aligned}\left|\widehat{r}_{j,\mathcal{C}}^T V_{\mathcal{C}}^{-1}\widehat{r}_{j,\mathcal{C}} - r_{j,\mathcal{C}}^T V_{\mathcal{C}}^{-1} r_{j,\mathcal{C}}\right| &\leq \left(\frac{2\tau_{\max}}{\tau_{\min}} + \sqrt{\frac{\tau_{\max}}{\tau_{\min}}}\right) M\sqrt{p} \\ &\leq \left(2\left(\frac{\tau_{\max}}{\tau_{\min}}\right)^{3/2} + 1\right) M\sqrt{p}.\end{aligned}$$

Therefore,

$$\max_{\mathcal{C}\subset\{1,2,\ldots,p\}} \max_{j\in\mathcal{C}^c} \left|\widehat{\sigma}_{j|\mathcal{C}}^2 - \sigma_{j|\mathcal{C}}^2\right| \leq 2\left(3\left(\frac{\tau_{\max}}{\tau_{\min}}\right)^{3/2} + 1\right) Mp.$$

Since $\sigma_{j|\mathcal{C}}^2 = v_{j,j} - r_{j,\mathcal{C}}^T V_{\mathcal{C}}^{-1} r_{j,\mathcal{C}} \geq v_{j,j} \geq \tau_{\min}$,

$$\max_{\mathcal{C}\subset\{1,2,\ldots,p\}} \max_{j\in\mathcal{C}^c} \frac{\left|\widehat{\sigma}_{j|\mathcal{C}}^2 - \sigma_{j|\mathcal{C}}^2\right|}{\sigma_{j|\mathcal{C}}^2} \leq \frac{2}{\tau_{\min}}\left(3\left(\frac{\tau_{\max}}{\tau_{\min}}\right)^{3/2} + 1\right) Mp.$$

Let $L = \frac{4}{\tau_{\min}}\left(3\left(\frac{\tau_{\max}}{\tau_{\min}}\right)^{3/2} + 1\right)$. For $0 < \epsilon < 1$, we have

$$\begin{aligned}&\Pr\left(\max_{\mathcal{C}\subset\{1,2,\ldots,p\}} \max_{j\in\mathcal{C}^c} \left|\log\widehat{\sigma}_{j|\mathcal{C}}^2 - \log\sigma_{j|\mathcal{C}}^2\right| > \epsilon\right) \\ \leq\ &\Pr\left(\max_{\mathcal{C}\subset\{1,2,\ldots,p\}} \max_{j\in\mathcal{C}^c} \frac{\left|\widehat{\sigma}_{j|\mathcal{C}}^2 - \sigma_{j|\mathcal{C}}^2\right|}{\sigma_{j|\mathcal{C}}^2} > \frac{\epsilon}{2}\right) \\ \leq\ &\Pr\left(M = \max_{j_1,j_2\in\{1,2,\ldots,p\}} |\widehat{v}_{j_1,j_2} - v_{j_1,j_2}| > \frac{\epsilon}{pL}\right) \\ \leq\ &\sum_{j_1,j_2\in\{1,2,\ldots,p\}} \Pr\left(|\widehat{v}_{j_1,j_2} - v_{j_1,j_2}| > \frac{\epsilon}{pL}\right) \\ \leq\ &\frac{p(p+1)}{2} C_1 \exp\left(-C_2 n \frac{\epsilon^2}{p^2 L^2}\right),\end{aligned}$$

where the last inequality follows from Bernstein inequality since predictors $\mathbf{X}$ follows either a multivariate normal distribution or a finite mixture of multivariate normal distributions. Similarly,

$$\begin{aligned}&\Pr\left(\max_{\mathcal{C}\subset\{1,2,\ldots,p\}} \max_{j\in\mathcal{C}^c} \left|\log\left[\widehat{\sigma}_{j|\mathcal{C}}^{(h)}\right]^2 - \log\left[\sigma_{j|\mathcal{C}}^{(h)}\right]^2\right| > \epsilon\right) \\ \leq\ &\frac{p(p+1)}{2} C_1 \exp\left(-C_2 n \frac{s_h \epsilon^2}{p^2 L^2}\right).\end{aligned}$$

Thus,

$$\begin{aligned}
&\Pr\left(\max_{\mathcal{C}\subset\{1,2,\ldots,p\}}\max_{j\in\mathcal{C}^c}\left|\sum_{h=1}^{H}s_h\log\left[\widehat{\sigma}_{j|\mathcal{C}}^{(h)}\right]^2-\sum_{h=1}^{H}s_h\log\left[\sigma_{j|\mathcal{C}}^{(h)}\right]^2\right|>\epsilon\right)\\
\leq&\sum_{h=1}^{H}\Pr\left(\max_{\mathcal{C}\subset\{1,2,\ldots,p\}}\max_{j\in\mathcal{C}^c}\left|\log\left[\widehat{\sigma}_{j|\mathcal{C}}^{(h)}\right]^2-\log\left[\sigma_{j|\mathcal{C}}^{(h)}\right]^2\right|>\frac{\epsilon}{s_hH}\right)\\
\leq&\sum_{h=1}^{H}\frac{p(p+1)}{2}C_1\exp\left(-C_2n\frac{\epsilon^2}{s_hH^2p^2L^2}\right)\\
\leq&\frac{Hp(p+1)}{2}C_1\exp\left(-C_2n\frac{\epsilon^2}{H^2p^2L^2}\right).
\end{aligned}$$

□

Proof of Theorem 2. We denote $R_{j|\mathcal{C}}=\log\widehat{\sigma}_{j|\mathcal{C}}^2-\log\sigma_{j|\mathcal{C}}^2$ and

$$\widetilde{R}_{j|\mathcal{C}}=\sum_{h=1}^{H}s_h\log\left[\widehat{\sigma}_{j|\mathcal{C}}^{(h)}\right]^2-\sum_{h=1}^{H}s_h\log\left[\sigma_{j|\mathcal{C}}^{(h)}\right]^2.$$

According to Lemma 3, for $0<\epsilon<1$, there exist constant C_1 and C_2 such that

$$\Pr\left(\max_{\mathcal{C}\subset\{1,2,\ldots,p\}}\max_{j\in\mathcal{C}^c}\left|R_{j|\mathcal{C}}\right|>\epsilon\right)\leq\frac{p(p+1)}{2}C_1\exp\left(-C_2n\frac{\epsilon^2}{p^2L^2}\right),$$

and

$$\Pr\left(\max_{\mathcal{C}\subset\{1,2,\ldots,p\}}\max_{j\in\mathcal{C}^c}\left|\widetilde{R}_{j|\mathcal{C}}\right|>\epsilon\right)\leq\frac{Hp(p+1)}{2}C_1\exp\left(-C_2n\frac{\epsilon^2}{H^2p^2L^2}\right),$$

where $L=\frac{4}{\tau_{\min}}\left(3\left(\frac{\tau_{\max}}{\tau_{\min}}\right)^{3/2}+1\right)$. Under Condition 2, $p=o\left(n^{\rho}\right)$ and $2\rho+2\kappa<1$,

$$\begin{aligned}
&\Pr\left(\max_{\mathcal{C}\subset\{1,2,\ldots,p\}}\max_{j\in\mathcal{C}^c}\left|R_{j|\mathcal{C}}\right|>Cn^{-\kappa}\right)\\
\leq&\frac{p(p+1)}{2}C_1\exp\left(-C_2n^{1-2\kappa-2\rho}\frac{C^2}{L^2}\right)\to0,
\end{aligned}$$

$$\begin{aligned}
&\Pr\left(\max_{\mathcal{C}\subset\{1,2,\ldots,p\}}\max_{j\in\mathcal{C}^c}\left|\widetilde{R}_{j|\mathcal{C}}\right|>Cn^{-\kappa}\right)\\
\leq&\frac{Hp(p+1)}{2}C_1\exp\left(-C_2n^{1-2\kappa-2\rho}\frac{C^2}{H^2L^2}\right)\to0,
\end{aligned}$$

for any positive constant C as $n \to \infty$. According to Proposition 5,

$$\begin{aligned}
\widehat{D}^*_{j|\mathcal{C}} &= \log \widehat{\sigma}^2_{j|\mathcal{C}} - \sum_{h=1}^{H} s_h \log \left[\widehat{\sigma}^{(h)}_{j|\mathcal{C}}\right]^2 \\
&= \log \sigma^2_{j|\mathcal{C}} - \sum_{h=1}^{H} s_h \log \left[\sigma^{(h)}_{j|\mathcal{C}}\right]^2 + R_{j|\mathcal{C}} - \widetilde{R}_{j|\mathcal{C}} \\
&= \log\left(1 + \frac{\operatorname{Var}(M_j) - \operatorname{Cov}(M_j, \mathbf{X}_{\mathcal{C}}) \left[\operatorname{Cov}(\mathbf{X}_{\mathcal{C}})\right]^{-1} \operatorname{Cov}(M_j, \mathbf{X}_{\mathcal{C}})^T}{\mathbb{E}(V_j)}\right) + \\
&\quad \log(\mathbb{E}V_j) - \mathbb{E}\log(V_j) + R_{j|\mathcal{C}} - \widetilde{R}_{j|\mathcal{C}},
\end{aligned}$$

where $M_j = \mathbb{E}(X_j|\mathbf{X}_{\mathcal{C}}, S(Y))$ and $V_j = \operatorname{Var}(X_j|\mathbf{X}_{\mathcal{C}}, S(Y))$.

When $\mathcal{C}^c \cap \mathcal{A} \neq \emptyset$ and all the relevant predictors indexed by $\mathcal{A}$ are stepwise detectable with constant $\kappa \geq 0$, then there exists $m \geq 0$ such that $\cup_{i=0}^{m-1} \mathcal{T}_i \subset \mathcal{C}$ and $\mathcal{C}^c \cap \mathcal{T}_m \neq \emptyset$. According to Definition 3, there exists $j \in \mathcal{C}^c \cap \mathcal{T}_m$ and $\xi_1, \xi_2 > 0$ such that either

$$\frac{\operatorname{Var}(M_j) - \operatorname{Cov}(M_j, \mathbf{X}_{\mathcal{C}}) \left[\operatorname{Cov}(\mathbf{X}_{\mathcal{C}})\right]^{-1} \operatorname{Cov}(M_j, \mathbf{X}_{\mathcal{C}})^T}{\mathbb{E}(V_j)} \geq \xi_1 n^{-\kappa},$$

that is, with sufficiently large n,

$$\log\left(1 + \frac{\operatorname{Var}(M_j) - \operatorname{Cov}(M_j, \mathbf{X}_{\mathcal{C}}) \left[\operatorname{Cov}(\mathbf{X}_{\mathcal{C}})\right]^{-1} \operatorname{Cov}(M_j, \mathbf{X}_{\mathcal{C}})^T}{\mathbb{E}(V_j)}\right) \geq \frac{\xi_1}{2} n^{-\kappa}.$$

or

$$\log(\mathbb{E}V_j) - \mathbb{E}\log(V_j) \geq \xi_2 n^{-\kappa}$$

Let $c = \min\left(\frac{\xi_1}{4}, \frac{\xi_2}{2}\right)$. Therefore,

$$\begin{aligned}
\widehat{D}^*_{j|\mathcal{C}} &\geq \log\left(1 + \frac{\operatorname{Var}(M_j) - \operatorname{Cov}(M_j, \mathbf{X}_{\mathcal{C}}) \left[\operatorname{Cov}(\mathbf{X}_{\mathcal{C}})\right]^{-1} \operatorname{Cov}(M_j, \mathbf{X}_{\mathcal{C}})^T}{\mathbb{E}(V_j)}\right) \\
&\quad + \log(\mathbb{E}V_j) - \mathbb{E}\log(V_j) - \left(\left|R_{j|\mathcal{C}}\right| + \left|\widetilde{R}_{j|\mathcal{C}}\right|\right) \\
&\geq 2cn^{-\kappa} - \left(\left|R_{j|\mathcal{C}}\right| + \left|\widetilde{R}_{j|\mathcal{C}}\right|\right)
\end{aligned}$$

Since

$$\Pr\left(\max_{\mathcal{C} \subset \{1,2,\ldots,p\}} \max_{j \in \mathcal{C}^c} \left|R_{j|\mathcal{C}}\right| > \frac{c}{2} n^{-\kappa}\right) \to 0,$$

and

$$\Pr\left(\max_{\mathcal{C} \subset \{1,2,\ldots,p\}} \max_{j \in \mathcal{C}^c} \left|\widetilde{R}_{j|\mathcal{C}}\right| > \frac{c}{2} n^{-\kappa}\right) \to 0,$$

we have

$$\Pr\left(\min_{\mathcal{C}:\mathcal{C}^c\cap\mathcal{A}\neq\emptyset}\max_{j\in\mathcal{C}^c\cap\mathcal{A}}\widehat{D}^*_{j|\mathcal{C}}\geq cn^{-\kappa}\right)\to 1,$$

as $n\to\infty$.

When $\mathcal{C}^c\cap\mathcal{A}=\emptyset$ under model (2.7), for any $j\in\mathcal{C}^c$, $M_j=\mathbb{E}\left(X_j|\mathbf{X}_\mathcal{C},S(Y)\right)=\mathbb{E}\left(X_j|\mathbf{X}_\mathcal{C}\right)$, which is a linear combination of predictors in $\mathbf{X}_\mathcal{C}$, and $V_j=\mathrm{Var}\left(X_j|\mathbf{X}_\mathcal{C},S(Y)\right)=\mathrm{Var}\left(X_j|\mathbf{X}_\mathcal{C}\right)$, which is a constant that does not depend on $\mathbf{X}_\mathcal{C}$ or $S(Y)$. Then,

$$\frac{\mathrm{Var}\left(M_j\right)-\mathrm{Cov}\left(M_j,\mathbf{X}_\mathcal{C}\right)\left[\mathrm{Cov}\left(\mathbf{X}_\mathcal{C}\right)\right]^{-1}\mathrm{Cov}\left(M_j,\mathbf{X}_\mathcal{C}\right)^T}{\mathbb{E}\left(V_j\right)}=0,$$

and

$$\log\left(\mathbb{E}V_j\right)-\mathbb{E}\log\left(V_j\right)=0.$$

Thus,

$$\widehat{D}^*_{j|\mathcal{C}}\leq\left|R_{j|\mathcal{C}}\right|+\left|\widetilde{R}_{j|\mathcal{C}}\right|,$$

and

$$\Pr\left(\max_{\mathcal{C}:\mathcal{C}^c\cap\mathcal{A}=\emptyset}\max_{j\in\mathcal{C}^c}\widehat{D}^*_{j|\mathcal{C}}<Cn^{-\kappa}\right)\to 1,$$

for any positive constant C as $n\to\infty$. □

A.6. Proof of Theorem 3 in Section 2.3.

Proof of Theorem 3. (a) We denote

$$\begin{aligned}D^*_j &= \log\sigma^2_j-\sum_{h=1}^{H}s_h\log\left[\sigma^{(h)}_j\right]^2\\ &= \frac{\mathrm{Var}\left(\mathbb{E}\left(X_j|S(Y)\right)\right)}{\mathbb{E}\left(\mathrm{Var}\left(X_j|S(Y)\right)\right)}+\log\mathbb{E}\left(\mathrm{Var}\left(X_j|S(Y)\right)\right)-\mathbb{E}\log\left[\mathrm{Var}\left(X_j|S(Y)\right)\right].\end{aligned}$$

First, we will prove that

$$P\left(\max_{j\in\{1,2,\ldots,p\}}\left|\widehat{D}^*_j-D^*_j\right|>Cn^{-\kappa}\right)\to 0,$$

for any constant $C>0$ as $n\to\infty$. There exist $C_1,C_2>0$ such that

$$\begin{aligned}&P\left(\max_{j\in\{1,2,\ldots,p\}}\left|\log\widehat{\sigma}^2_j-\log\sigma^2_j\right|>\epsilon\right)\leq P\left(\max_{j\in\{1,2,\ldots,p\}}\frac{\left|\widehat{\sigma}^2_j-\sigma^2_j\right|}{\sigma^2_j}>\frac{\epsilon}{2}\right)\\ \leq\ &P\left(\max_{j\in\{1,2,\ldots,p\}}\left|\widehat{\sigma}^2_j-\sigma^2_j\right|>\frac{\epsilon\tau_{\min}}{2}\right)\leq\sum_{j\in\{1,2,\ldots,p\}}P\left(\left|\widehat{\sigma}^2_j-\sigma^2_j\right|>\frac{\epsilon\tau_{\min}}{2}\right)\\ \leq\ &pC_1\exp\left(-C_2n\frac{\epsilon^2\tau_{\min}}{4}\right),\end{aligned}$$

where the last inequality follows from Bernstein inequality since predictors $\mathbf{X}$ follows either a multivariate normal distribution or a finite mixture of multivariate normal distributions. Since $\log(p) = o(n^\gamma)$ with $\gamma > 0$ and $\gamma + 2\kappa < 1$,

$$\begin{aligned} & P\left(\max_{j\in\{1,2,\ldots,p\}} \left|\log \widehat{\sigma}_j^2 - \log \sigma_j^2\right| > Cn^{-\kappa}\right) \le pC_1 \exp\left(-C_2 n^{1-2\kappa}\frac{C^2\tau_{\min}}{4}\right) \\ \le \;& C_1 \exp\left(-C_2 C^2 n^{1-2\kappa}\frac{\tau_{\min}}{4} + n^\gamma\right) \to 0, \end{aligned}$$

for any constant $C > 0$ as $n \to \infty$. Similarly,

$$\begin{aligned} & P\left(\max_{j\in\{1,2,\ldots,p\}} \left|\sum_{h=1}^H s_h \left(\log\left[\widehat{\sigma}_j^{(h)}\right]^2 - \log\left[\sigma_j^{(h)}\right]^2\right)\right| > Cn^{-\kappa}\right) \\ \le \;& \sum_{h=1}^H P\left(\max_{j\in\{1,2,\ldots,p\}} \left|\log\left[\widehat{\sigma}_j^{(h)}\right]^2 - \log\left[\sigma_j^{(h)}\right]^2\right| > \frac{Cn^{-\kappa}}{s_h H}\right) \\ \le \;& HpC_1 \exp\left(-C_2 n^{1-2\kappa}\frac{C^2\tau_{\min}}{4H^2}\right) \\ \le \;& HC_1 \exp\left(-C_2 C^2 n^{1-2\kappa}\frac{\tau_{\min}}{4H^2} + n^\gamma\right) \to 0, \end{aligned}$$

for any constant $C > 0$ as $n \to \infty$. Thus,

$$P\left(\max_{j\in\{1,2,\ldots,p\}} \left||\widehat{D}_j^* - D_j^*\right| > Cn^{-\kappa}\right) \to 0,$$

for any constant $C > 0$ as $n \to \infty$.

Second, note that if $\frac{\mathrm{Var}(\mathbb{E}(X_j|S(Y)))}{\mathbb{E}(\mathrm{Var}(X_j|S(Y)))} \ge \xi_1 n^{-\kappa}$, then for sufficiently large n,

$$n\log\left[1 + \frac{\mathrm{Var}\left(\mathbb{E}(X_j|S(Y))\right)}{\mathbb{E}\left(\mathrm{Var}(X_j|S(Y))\right)}\right] \ge \frac{\xi_1}{2} n^{1-\kappa}.$$

When all the relevant predictors indexed by $\mathcal{A}$ are independtly detectable, there exists $c = \min\{\frac{\xi_1}{4}, \frac{\xi_2}{2}\} > 0$ such that

$$D_j^* \ge 2cn^{-\kappa}, \text{ for } j \in \mathcal{A}.$$

The events $\left\{\min_{j\in\mathcal{A}} D_j^* < cn^{-\kappa}\right\} \subset \left\{\max_{j\in\{1,2,\ldots,p\}} \left|\widehat{D}_j^* - D_j^*\right| > cn^{-\kappa}\right\}$, and

$$P\left(\min_{j\in\mathcal{A}} D_j^* < cn^{-\kappa}\right) \le P\left(\max_{j\in\{1,2,\ldots,p\}} \left|\widehat{D}_j^* - D_j^*\right| > cn^{-\kappa}\right) \to 0,$$

as $n \to \infty$.

(b) We denote $\widehat{M}_c = \{j : \widehat{D}_j^* \geq cn^{-\kappa}, \text{ for } 1 \leq j \leq p\}$ and $M_{\frac{c}{2}} = \{j : D_j^* \geq \frac{c}{2}n^{-\kappa}, \text{ for } 1 \leq j \leq p\}$. We will prove that there exists $C > 0$ such that $\left|M_{\frac{c}{2}}\right| \leq Cn^{\kappa+\eta}$. Since

$$\left|M_{\frac{c}{2}}\right| \frac{c}{2} n^{-\kappa} \leq \sum_{j=1}^{p} \widehat{D}_j^*,$$

we just need to prove that

$$\sum_{j=1}^{p} \widehat{D}_j^* \leq Cn^{\eta},$$

for some positive constant C. First,

$$\begin{aligned} \sum_{j=1}^{p} \log\left[1 + \frac{\operatorname{Var}\left(\mathbb{E}\left(X_j|S(Y)\right)\right)}{\mathbb{E}\left(\operatorname{Var}\left(X_j|S(Y)\right)\right)}\right] &\leq \sum_{j=1}^{p} \frac{\operatorname{Var}\left(\mathbb{E}\left(X_j|S(Y)\right)\right)}{\mathbb{E}\left(\operatorname{Var}\left(X_j|S(Y)\right)\right)} \\ &\leq \frac{1}{\tau_{\min}} \sum_{j=1}^{p} \operatorname{Var}\left(\mathbb{E}\left(X_j|S(Y)\right)\right), \end{aligned}$$

and

$$\sum_{j=1}^{p} \operatorname{Var}\left(\mathbb{E}\left(X_j|S(Y)\right)\right) = \sum_{j\in\mathcal{A}^c} \operatorname{Var}\left(\mathbb{E}\left(X_j|S(Y)\right)\right) + \sum_{j\in\mathcal{A}} \operatorname{Var}\left(\mathbb{E}\left(X_j|S(Y)\right)\right).$$

Under model (2.7), $\mathbb{E}\left(\mathbf{X}_{\mathcal{A}^c}|S(Y)\right) = \alpha + \boldsymbol{\beta}^T \mathbb{E}\left(\mathbf{X}_{\mathcal{A}}|S(Y)\right)$. Since

$$\begin{aligned} &\sum_{j\in\mathcal{A}^c} \operatorname{Var}\left(\mathbb{E}\left(X_j|S(Y)\right)\right) = \operatorname{trace}\left(\operatorname{Cov}\left[\mathbb{E}\left(\mathbf{X}_{\mathcal{A}^c}|S(Y)\right)\right]\right) \\ = \;& \operatorname{trace}\left(\boldsymbol{\beta}^T \operatorname{Cov}\left[\mathbb{E}\left(\mathbf{X}_{\mathcal{A}}|S(Y)\right)\right]\boldsymbol{\beta}\right) \\ \leq \;& \lambda_{\max}\left(\boldsymbol{\beta}^T\boldsymbol{\beta}\right) \operatorname{trace}\left(\operatorname{Cov}\left[\mathbb{E}\left(\mathbf{X}_{\mathcal{A}}|S(Y)\right)\right]\right) \\ = \;& \lambda_{\max}\left(\boldsymbol{\beta}^T\boldsymbol{\beta}\right) \sum_{j\in\mathcal{A}} \operatorname{Var}\left(\mathbb{E}\left(X_j|S(Y)\right)\right). \end{aligned}$$

and

$$\begin{aligned} &\lambda_{\max}\left(\boldsymbol{\beta}^T\boldsymbol{\beta}\right) \\ \leq \;& \frac{\lambda_{\max}\left(\operatorname{Cov}\left(\mathbf{X}_{\mathcal{A}^c}, \mathbf{X}_{\mathcal{A}}|S(Y)\right)\left[\operatorname{Cov}\left(\mathbf{X}_{\mathcal{A}}|S(Y)\right)\right]^{-1} \operatorname{Cov}\left(\mathbf{X}_{\mathcal{A}^c}, \mathbf{X}_{\mathcal{A}}|S(Y)\right)^T\right)}{\lambda_{\min}\left(\operatorname{Cov}\left(\mathbf{X}_{\mathcal{A}}|S(Y)\right)\right)} \\ \leq \;& \frac{\lambda_{\max}\left(\operatorname{Cov}\left(\mathbf{X}_{\mathcal{A}^c}|S(Y)\right)\right)}{\lambda_{\min}\left(\operatorname{Cov}\left(\mathbf{X}_{\mathcal{A}}|S(Y)\right)\right)} \leq \frac{\tau_{\max}}{\tau_{\min}}, \end{aligned}$$

we have

$$\begin{aligned}\sum_{j=1}^{p}\log\left[1+\frac{\mathrm{Var}\left(\mathbb{E}\left(X_j|S(Y)\right)\right)}{\mathbb{E}\left(\mathrm{Var}\left(X_j|S(Y)\right)\right)}\right] &\leq \frac{1}{\tau_{\min}}\sum_{j=1}^{p}\mathrm{Var}\left(\mathbb{E}\left(X_j|S(Y)\right)\right)\\ &\leq \frac{1}{\tau_{\min}}\left(1+\frac{\tau_{\max}}{\tau_{\min}}\right)\sum_{j\in\mathcal{A}}\mathrm{Var}\left(\mathbb{E}\left(X_j|S(Y)\right)\right).\end{aligned}$$

Second, for $j\in\mathcal{A}^c$, $\mathrm{Var}(X_j|\mathbf{X}_{\mathcal{A}},S(Y))$ is a constant, and

$$\begin{aligned}&\log\left[\mathbb{E}\left(\mathrm{Var}(X_j|S(Y))\right)\right]-\mathbb{E}\left[\log\left(\mathrm{Var}(X_j|S(Y))\right)\right]\\ \leq\ &\mathbb{E}\left(\mathrm{Var}(X_j|S(Y))\right)\mathbb{E}\left(\frac{1}{\mathrm{Var}(X_j|S(Y))}\right)-1\leq\frac{\mathbb{E}\left(\mathrm{Var}(X_j|S(Y))\right)}{\mathrm{Var}(X_j|\mathbf{X}_{\mathcal{A}},S(Y))}-1\\ =\ &\frac{\mathbb{E}\left[\mathrm{Var}\left(\mathbb{E}(X_j|\mathbf{X}_{\mathcal{A}},S(Y))|S(Y)\right)\right]}{\mathrm{Var}(X_j|\mathbf{X}_{\mathcal{A}},S(Y))}\leq\frac{1}{\tau_{\min}}\mathbb{E}\left[\mathrm{Var}\left(\mathbb{E}(X_j|\mathbf{X}_{\mathcal{A}},S(Y))|S(Y)\right)\right].\end{aligned}$$

So

$$\begin{aligned}&\sum_{j\in\mathcal{A}^c}\left(\log\left[\mathbb{E}\left(\mathrm{Var}(X_j|S(Y))\right)\right]-\mathbb{E}\left[\log\left(\mathrm{Var}(X_j|S(Y))\right)\right]\right)\\ \leq\ &\frac{1}{\tau_{\min}}\sum_{j\in\mathcal{A}^c}\mathbb{E}\left[\mathrm{Var}\left(\mathbb{E}(X_j|\mathbf{X}_{\mathcal{A}},S(Y))|S(Y)\right)\right]\\ =\ &\frac{1}{\tau_{\min}}\mathrm{trace}\left(\mathbb{E}\left[\mathrm{Cov}\left(\mathbb{E}(\mathbf{X}_{\mathcal{A}^c}|\mathbf{X}_{\mathcal{A}},S(Y))|S(Y)\right)\right]\right)\\ =\ &\frac{1}{\tau_{\min}}\mathrm{trace}\left(\boldsymbol{\beta}^T\mathbb{E}\left[\mathrm{Cov}\left(\mathbf{X}_{\mathcal{A}}|S(Y)\right)\right]\boldsymbol{\beta}^T\right)\\ \leq\ &\frac{1}{\tau_{\min}}\lambda_{\max}\left(\boldsymbol{\beta}^T\boldsymbol{\beta}\right)\sum_{j\in\mathcal{A}}\mathbb{E}\left[\mathrm{Var}\left(X_j|S(Y)\right)\right]\leq\frac{\tau_{\max}}{\tau_{\min}^2}\sum_{j\in\mathcal{A}}\mathbb{E}\left[\mathrm{Var}\left(X_j|S(Y)\right)\right].\end{aligned}$$

Therefore,

$$\begin{aligned}&\sum_{j=1}^{p}\left(\log\left[\mathbb{E}\left(\mathrm{Var}(X_j|S(Y))\right)\right]-\mathbb{E}\left[\log\left(\mathrm{Var}(X_j|S(Y))\right)\right]\right)\\ \leq\ &\frac{1}{\tau_{\min}}\left(1+\frac{\tau_{\max}}{\tau_{\min}}\right)\sum_{j\in\mathcal{A}}\mathbb{E}\left[\mathrm{Var}\left(X_j|S(Y)\right)\right].\end{aligned}$$

Moreover, $\sum_{j\in\mathcal{A}}\mathrm{Var}\left(X_j|S(Y)\right)\leq|\mathcal{A}|\tau_{\max}\leq\tau_{\max}\xi_0 n^{\eta}$, and

$$\begin{aligned}\sum_{j=1}^{p}D_j^* &= \sum_{j=1}^{p}\left(\log\left[1+\frac{\mathrm{Var}\left(\mathbb{E}(X_j|S(Y))\right)}{\mathbb{E}\left(\mathrm{Var}(X_j|S(Y))\right)}\right]\right.\\ &\quad\left.+\log\left[\mathbb{E}\left(\mathrm{Var}(X_j|S(Y))\right)\right]-\mathbb{E}\left[\log\left(\mathrm{Var}(X_j|S(Y))\right)\right]\right)\\ &\leq \frac{1}{\tau_{\min}}\left(1+\frac{\tau_{\max}}{\tau_{\min}}\right)\sum_{j\in\mathcal{A}}\mathrm{Var}\left(X_j|S(Y)\right)\leq\frac{\tau_{\max}}{\tau_{\min}}\left(1+\frac{\tau_{\max}}{\tau_{\min}}\right)\xi_0 n^{\eta}.\end{aligned}$$

Thus, there exists $C > 0$ such that

$$\left|M_{\frac{c}{2}}\right| \leq \frac{2}{c} n^{\kappa} \sum_{j=1}^{p} D_j^* \leq \frac{\tau_{\max}}{\tau_{\min}} \left(1 + \frac{\tau_{\max}}{\tau_{\min}}\right) \frac{2\xi_0}{c} n^{\kappa+\eta} \leq C n^{\kappa+\eta}.$$

Then,

$$P\left(\left|\widehat{M}_c\right| > C n^{\kappa+\eta}\right) \leq P\left(\left|\widehat{M}_c\right| > \left|M_{\frac{c}{2}}\right|\right),$$

and the event $\left\{\left|\widehat{M}_c\right| > \left|M_{\frac{c}{2}}\right|\right\} \subset \left\{\exists j \text{ such that } D_j^* < \frac{c}{2} n^{-\kappa} \text{ and } \widehat{D}_j^* \geq c n^{-\kappa}\right\}$ $\subset \left\{\max_{j\in\{1,2,\ldots,p\}} \left|\widehat{D}_j^* - D_j^*\right| > \frac{c}{2} n^{-\kappa}\right\}$. Thus, according to the results in (a),

$$P\left(\left|\widehat{M}_c\right| > C n^{\kappa+\eta}\right) \leq P\left(\max_{j\in\{1,2,\ldots,p\}} \left|\widehat{D}_j^* - D_j^*\right| > \frac{c}{2} n^{-\kappa}\right) \to 0,$$

as $n \to \infty$. □

A.7. Proof on Choices of Slicing Schemes in Section 3. Suppose $(S_1, S_2, \ldots, S_H)$ is the true slicing scheme under model (2.2) or (2.7), and $S(Y)$ denotes the slice membership of the slicing scheme, *i.e.*, $S(Y) = h$ if $Y \in S_h$. We say that a slicing scheme $\widetilde{S}(Y)$ is a refinement of $S(Y)$, which is denoted by $\widetilde{S}(Y) \preceq S(Y)$, if there exists a function g such that $S(Y) = g(\widetilde{S}(Y))$.

For any slicing scheme $\widetilde{S}(Y)$, as $n \to \infty$, the limit of the likelihood-ratio test statistic under model (2.2) is given by

$$\begin{aligned} & D_{j|\mathcal{C},\widetilde{S}(Y)} = \lim_{n\to\infty} \widehat{D}_{j|\mathcal{C},\widetilde{S}(Y)} \\ = \; & \log\left[\mathrm{Cov}\left(X_j\right) - \mathrm{Cov}\left(X_j, \mathbf{X}_{\mathcal{C}}\right)\left[\mathrm{Var}\left(\mathbf{X}_{\mathcal{C}}\right)\right]^{-1} \mathrm{Var}\left(X_j, \mathbf{X}_{\mathcal{C}}\right)^T\right] - \\ & \log\left[\mathbb{E}\left(\mathrm{Var}\left(X_j \mid \widetilde{S}(Y)\right)\right) - \mathbb{E}\left(\mathrm{Cov}\left(X_j, \mathbf{X}_{\mathcal{C}} \mid \widetilde{S}(Y)\right)\right) \times \right. \\ & \left. \left[\mathbb{E}\left(\mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}} \mid \widetilde{S}(Y)\right)\right)\right]^{-1} \mathbb{E}\left(\mathrm{Cov}\left(X_j, \mathbf{X}_{\mathcal{C}} \mid \widetilde{S}(Y)\right)\right)^T\right], \end{aligned}$$

and the limit of the augmented likelihood-ratio test statistic under model (2.7) is given by

$$\begin{aligned} & D^*_{j|\mathcal{C},\widetilde{S}(Y)} = \lim_{n\to\infty} \widehat{D}^*_{j|\mathcal{C},\widetilde{S}(Y)} \\ = \; & \log\left[\mathrm{Var}\left(X_j\right) - \mathrm{Cov}\left(X_j, \mathbf{X}_{\mathcal{C}}\right)\left[\mathrm{Var}\left(\mathbf{X}_{\mathcal{C}}\right)\right]^{-1} \mathrm{Cov}\left(X_j, \mathbf{X}_{\mathcal{C}}\right)^T\right] - \\ & \mathbb{E}\left(\log\left[\mathrm{Var}\left(X_j \mid \widetilde{S}(Y)\right) - \right.\right. \\ & \left.\left. \mathrm{Cov}\left(X_j, \mathbf{X}_{\mathcal{C}} \mid \widetilde{S}(Y)\right)\left[\mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}} \mid \widetilde{S}(Y)\right)\right]^{-1} \mathrm{Cov}\left(X_j, \mathbf{X}_{\mathcal{C}} \mid \widetilde{S}(Y)\right)^T\right]\right) \end{aligned}$$

For the true slicing scheme $S(Y)$ or a slicing scheme $\widetilde{S}(Y)$ that is a refinement of $S(Y)$, *i.e.*, $\widetilde{S}(Y) \preceq S$, under model (2.2), we have

$$\begin{aligned} & D_{j|\mathcal{C},\widetilde{S}(Y)} = D_{j|\mathcal{C},S(Y)} \\ = \; & \log\left[\mathrm{Var}\left(X_j\right) - \mathrm{Cov}\left(X_j, \mathbf{X}_{\mathcal{C}}\right)\left[\mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}}\right)\right]^{-1}\mathrm{Cov}\left(X_j, \mathbf{X}_{\mathcal{C}}\right)^T\right] \\ & - \log\left(\mathrm{Var}(X_j \mid \mathbf{X}_{\mathcal{C}}, S(Y))\right), \end{aligned}$$

where $\mathrm{Var}(X_j \mid \mathbf{X}_{\mathcal{C}}, S(Y))$ is a constant that does not depend on $\mathbf{X}_{\mathcal{C}}$ or $S(Y)$. Similary, under the augmented model (2.7),

$$\begin{aligned} & D^*_{j|\mathcal{C},\widetilde{S}(Y)} = D^*_{j|\mathcal{C},S(Y)} \\ = \; & \log\left[\mathrm{Var}\left(X_j\right) - \mathrm{Cov}\left(X_j, \mathbf{X}_{\mathcal{C}}\right)\left[\mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}}\right)\right]^{-1}\mathrm{Cov}\left(X_j, \mathbf{X}_{\mathcal{C}}\right)^T\right] \\ & -\mathbb{E}\left(\log\left(\mathrm{Var}(X_j \mid \mathbf{X}_{\mathcal{C}}, S(Y))\right)\right). \end{aligned}$$

For a slicing scheme $\widetilde{S}(Y)$ that is "coarser" than the true slicing scheme $S(Y)$, *i.e.*, $S(Y) \preceq \widetilde{S}(Y)$, we have the following theorem.

PROPOSITION 6. *Suppose $\widetilde{S}(Y)$ is a slicing scheme such that $S(Y) \preceq \widetilde{S}(Y)$, where $S(Y)$ is the true slicing scheme.*
(a) Under model (2.2),

$$D_{j|\mathcal{C},S(Y)} \geq D_{j|\mathcal{C},\widetilde{S}(Y)},$$

where the equality holds if $\mathcal{A} \subset \mathcal{C}$, where $\mathcal{A}$ is the index set of relevant predictors.
(b) Under model (2.7),

$$D^*_{j|\mathcal{C},S(Y)} \geq D^*_{j|\mathcal{C},\widetilde{S}(Y)},$$

where the equality holds if $\mathcal{A} \subset \mathcal{C}$, where $\mathcal{A}$ is the index set of relevant predictors.

PROOF OF PROPOSITION 6. (a) Since

$$\begin{aligned} \mathrm{Var}\left(X_j \mid \widetilde{S}(Y)\right) \; = \; & \mathbb{E}\left(\mathrm{Var}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, \widetilde{S}(Y)\right) \mid \widetilde{S}(Y)\right) \\ & + \mathrm{Var}\left(\mathbb{E}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, \widetilde{S}(Y)\right) \mid \widetilde{S}(Y)\right), \end{aligned}$$

we have,

$$\begin{aligned}
& \mathbb{E}\left(\mathrm{Var}\left(X_j \mid \widetilde{S}(Y)\right)\right) - \mathbb{E}\left(\mathrm{Cov}\left(X_j, \mathbf{X}_{\mathcal{C}} \mid \widetilde{S}(Y)\right)\right) \times \\
& \left[\mathbb{E}\left(\mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}} \mid \widetilde{S}(Y)\right)\right)\right]^{-1} \mathbb{E}\left(\mathrm{Cov}\left(X_j, \mathbf{X}_{\mathcal{C}} \mid \widetilde{S}(Y)\right)\right)^T \\
= \; & \mathbb{E}\left(\mathrm{Var}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, \widetilde{S}(Y)\right)\right) + \mathbb{E}\left[\mathrm{Var}\left(\mathbb{E}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, \widetilde{S}(Y)\right) \mid \widetilde{S}(Y)\right)\right] - \\
& \mathbb{E}\left[\mathrm{Cov}\left(\mathbb{E}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, \widetilde{S}(Y)\right), \mathbf{X}_{\mathcal{C}} \mid \widetilde{S}(Y)\right)\right] \left[\mathbb{E}\left(\mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}} \mid \widetilde{S}(Y)\right)\right)\right]^{-1} \\
& \mathbb{E}\left[\mathrm{Cov}\left(\mathbb{E}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, \widetilde{S}(Y)\right), \mathbf{X}_{\mathcal{C}} \mid \widetilde{S}(Y)\right)\right]^T \\
\geq \; & \mathbb{E}\left(\mathrm{Var}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, \widetilde{S}(Y)\right)\right),
\end{aligned}$$

where the equality holds if $\mathbb{E}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, \widetilde{S}(Y)\right)$ is a linear combination of $\mathbf{X}_{\mathcal{C}}$ that does not depend on $\widetilde{S}(Y)$. Since $S(Y) \preceq \widetilde{S}(Y)$, the σ-algebra $\sigma(\widetilde{S}(Y)) \subset \sigma(S(Y))$. Thus,

$$\begin{aligned}
\mathrm{Var}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, \widetilde{S}(Y)\right) &= \mathbb{E}\left(\mathrm{Var}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, S(Y)\right) \middle| \mathbf{X}_{\mathcal{C}}, \widetilde{S}(Y)\right) \\
&\quad + \mathrm{Var}\left(\mathbb{E}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, S(Y)\right) \middle| \mathbf{X}_{\mathcal{C}}, \widetilde{S}(Y)\right) \\
&\geq \mathbb{E}\left(\mathrm{Var}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, S(Y)\right) \middle| \mathbf{X}_{\mathcal{C}}, \widetilde{S}(Y)\right),
\end{aligned}$$

and

$$\mathbb{E}\left(\mathrm{Var}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, \widetilde{S}(Y)\right)\right) \geq \mathbb{E}\left(\mathrm{Var}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, S(Y)\right)\right),$$

where the equality holds if $\mathbb{E}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, S(Y)\right)$ is a linear combination of $\mathbf{X}_{\mathcal{C}}$ that does not depend on $S(Y)$. Because $S(Y)$ is the true slicing scheme, under model (2.2), $\mathrm{Var}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, S(Y)\right)$ is a constant that does not depend on $\mathbf{X}_{\mathcal{C}}$ or $S(Y)$, that is, $\mathbb{E}\left(\mathrm{Var}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, S(Y)\right)\right) = \mathrm{Var}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, S(Y)\right)$. Therefore,

$$\begin{aligned}
& \mathbb{E}\left(\mathrm{Var}\left(X_j \mid \widetilde{S}(Y)\right)\right) - \mathbb{E}\left(\mathrm{Cov}\left(X_j, \mathbf{X}_{\mathcal{C}} \mid \widetilde{S}(Y)\right)\right) \times \\
& \left[\mathbb{E}\left(\mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}} \mid \widetilde{S}(Y)\right)\right)\right]^{-1} \mathbb{E}\left(\mathrm{Cov}\left(X_j, \mathbf{X}_{\mathcal{C}} \mid \widetilde{S}(Y)\right)\right)^T \\
\geq \; & \mathrm{Var}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, S(Y)\right),
\end{aligned}$$

and

$$\begin{aligned} D_{j|\mathcal{C},S(Y)} &= \log\left[\mathrm{Var}\left(X_j\right) - \mathrm{Cov}\left(X_j, \mathbf{X}_\mathcal{C}\right)\left[\mathrm{Cov}\left(\mathbf{X}_\mathcal{C}\right)\right]^{-1}\mathrm{Cov}\left(X_j, \mathbf{X}_\mathcal{C}\right)^T\right] - \\ &\quad \log\left(\mathrm{Var}(X_j \mid \mathbf{X}_\mathcal{C}, S(Y))\right) \\ &\geq \log\left[\mathrm{Var}\left(X_j\right) - \mathrm{Cov}\left(X_j, \mathbf{X}_\mathcal{C}\right)\left[\mathrm{Cov}\left(\mathbf{X}_\mathcal{C}\right)\right]^{-1}\mathrm{Cov}\left(X_j, \mathbf{X}_\mathcal{C}\right)^T\right] - \\ &\quad \log\left[\mathbb{E}\left(\mathrm{Var}\left(X_j \mid \widetilde{S}(Y)\right)\right) - \mathbb{E}\left(\mathrm{Cov}\left(X_j, \mathbf{X}_\mathcal{C} \mid \widetilde{S}(Y)\right)\right) \times \right. \\ &\qquad \left.\left[\mathbb{E}\left(\mathrm{Cov}\left(\mathbf{X}_\mathcal{C} \mid \widetilde{S}(Y)\right)\right)\right]^{-1}\mathbb{E}\left(\mathrm{Cov}\left(X_j, \mathbf{X}_\mathcal{C} \mid \widetilde{S}(Y)\right)\right)^T\right] \\ &= D_{j|\mathcal{C},\widetilde{S}(Y)}. \end{aligned}$$

When $\mathcal{A} \subset \mathcal{C}$, under model (2.2), the predictor indexed by $j \in \mathcal{C}^c$ has the same conditional distribution across difference slices given $\mathbf{X}_\mathcal{C}$. The conditional expectations $\mathbb{E}\left(X_j \mid \mathbf{X}_\mathcal{C}, S(Y)\right)$ and $\mathbb{E}\left(X_j \mid \mathbf{X}_\mathcal{C}, \widetilde{S}(Y)\right)$ are linear combinations of $\mathbf{X}_\mathcal{C}$ that do not depend on $S(Y)$ or $\widetilde{S}(Y)$. Thus, the equalities hold in this case.

To prove (b), note that

$$\begin{aligned} &\mathrm{Var}\left(X_j \mid \widetilde{S}(Y)\right) - \mathrm{Cov}\left(X_j, \mathbf{X}_\mathcal{C} \mid \widetilde{S}(Y)\right)\left[\mathrm{Cov}\left(\mathbf{X}_\mathcal{C} \mid \widetilde{S}(Y)\right)\right]^{-1} \times \\ &\mathrm{Cov}\left(X_j, \mathbf{X}_\mathcal{C} \mid \widetilde{S}(Y)\right)^T \\ = \;& \mathbb{E}\left(\mathrm{Var}\left(X_j \mid \mathbf{X}_\mathcal{C}, \widetilde{S}(Y)\right) \mid \widetilde{S}(Y)\right) + \mathrm{Var}\left(\mathbb{E}\left(X_j \mid \mathbf{X}_\mathcal{C}, \widetilde{S}(Y)\right) \mid \widetilde{S}(Y)\right) - \\ &\mathrm{Cov}\left(\mathbb{E}\left(X_j \mid \mathbf{X}_\mathcal{C}, \widetilde{S}(Y)\right), \mathbf{X}_\mathcal{C} \mid \widetilde{S}(Y)\right)\left[\mathrm{Cov}\left(\mathbf{X}_\mathcal{C} \mid \widetilde{S}(Y)\right)\right]^{-1} \times \\ &\mathrm{Cov}\left(\mathbb{E}\left(X_j \mid \mathbf{X}_\mathcal{C}, \widetilde{S}(Y)\right), \mathbf{X}_\mathcal{C} \mid \widetilde{S}(Y)\right)^T \\ \geq \;& \mathbb{E}\left(\mathrm{Var}\left(X_j \mid \mathbf{X}_\mathcal{C}, \widetilde{S}(Y)\right) \mid \widetilde{S}(Y)\right). \end{aligned}$$

where the equality holds if $\mathbb{E}\left(X_j \mid \mathbf{X}_\mathcal{C}, \widetilde{S}(Y)\right)$ is a linear combination of $\mathbf{X}_\mathcal{C}$ that does not depend on $\widetilde{S}(Y)$. Since $S(Y) \preceq \widetilde{S}(Y)$, the σ-algebra $\sigma(\widetilde{S}(Y)) \subset \sigma(S(Y))$. Thus,

$$\begin{aligned} \mathrm{Var}\left(X_j \mid \mathbf{X}_\mathcal{C}, \widetilde{S}(Y)\right) &= \mathbb{E}\left(\mathrm{Var}\left(X_j \mid \mathbf{X}_\mathcal{C}, S(Y)\right)\middle|\mathbf{X}_\mathcal{C}, \widetilde{S}(Y)\right) \\ &\quad + \mathrm{Var}\left(\mathbb{E}\left(X_j \mid \mathbf{X}_\mathcal{C}, S(Y)\right)\middle|\mathbf{X}_\mathcal{C}, \widetilde{S}(Y)\right) \\ &\geq \mathbb{E}\left(\mathrm{Var}\left(X_j \mid \mathbf{X}_\mathcal{C}, S(Y)\right)\middle|\mathbf{X}_\mathcal{C}, \widetilde{S}(Y)\right), \end{aligned}$$

and

$$\begin{aligned}
&\operatorname{Var}\left(X_j \mid \widetilde{S}(Y)\right) - \operatorname{Cov}\left(X_j, \mathbf{X}_{\mathcal{C}} \mid \widetilde{S}(Y)\right)\left[\operatorname{Cov}\left(\mathbf{X}_{\mathcal{C}} \mid \widetilde{S}(Y)\right)\right]^{-1} \times \\
&\operatorname{Cov}\left(X_j, \mathbf{X}_{\mathcal{C}} \mid \widetilde{S}(Y)\right)^T \\
\geq\ & \mathbb{E}\left(\operatorname{Var}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, \widetilde{S}(Y)\right)\middle|\widetilde{S}(Y)\right) \geq \mathbb{E}\left(\operatorname{Var}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, S(Y)\right)\middle|\widetilde{S}(Y)\right),
\end{aligned}$$

where the equalities hold if $\mathbb{E}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, S(Y)\right)$ and $\mathbb{E}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, \widetilde{S}(Y)\right)$ are linear combinations of $\mathbf{X}_{\mathcal{C}}$ that do not depend on $S(Y)$ or $\widetilde{S}(Y)$. According to Jensen's inequality,

$$\log\left[\mathbb{E}\left(\operatorname{Var}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, S(Y)\right)\middle|\widetilde{S}(Y)\right)\right] \geq \mathbb{E}\left(\log\left[\operatorname{Var}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, S(Y)\right)\right]\middle|\widetilde{S}(Y)\right).$$

where the equality holds if $\operatorname{Var}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, S(Y)\right)$ is a constant that does not depend on $\mathbf{X}_{\mathcal{C}}$ or $S(Y)$. Therefore,

$$\begin{aligned}
&\mathbb{E}\left(\log\left[\operatorname{Var}\left(X_j \mid \widetilde{S}(Y)\right) - \operatorname{Cov}\left(X_j, \mathbf{X}_{\mathcal{C}} \mid \widetilde{S}(Y)\right) \times \right.\right. \\
&\qquad \left.\left.\left[\operatorname{Cov}\left(\mathbf{X}_{\mathcal{C}} \mid \widetilde{S}(Y)\right)\right]^{-1} \operatorname{Cov}\left(X_j, \mathbf{X}_{\mathcal{C}} \mid \widetilde{S}(Y)\right)^T\right]\right) \\
\geq\ & \mathbb{E}\left(\log\left[\mathbb{E}\left(\operatorname{Var}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, \widetilde{S}(Y)\right)\middle|\widetilde{S}(Y)\right)\right]\right) \\
\geq\ & \mathbb{E}\left(\mathbb{E}\left(\log\left[\operatorname{Var}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, S(Y)\right)\right] \mid \widetilde{S}(Y)\right)\right) \\
=\ & \mathbb{E}\left(\log\left[\operatorname{Var}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, S(Y)\right)\right]\right),
\end{aligned}$$

and

$$\begin{aligned}
D^*_{j|\mathcal{C},S(Y)} =\ & \log\left[\operatorname{Var}\left(X_j\right) - \operatorname{Cov}\left(X_j, \mathbf{X}_{\mathcal{C}}\right)\left[\operatorname{Cov}\left(\mathbf{X}_{\mathcal{C}}\right)\right]^{-1} \operatorname{Cov}\left(X_j, \mathbf{X}_{\mathcal{C}}\right)^T\right] - \\
& \mathbb{E}\left(\log\left[\operatorname{Var}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, S(Y)\right)\right]\right) \\
\geq\ & \log\left[\operatorname{Var}\left(X_j\right) - \operatorname{Cov}\left(X_j, \mathbf{X}_{\mathcal{C}}\right)\left[\operatorname{Cov}\left(\mathbf{X}_{\mathcal{C}}\right)\right]^{-1} \operatorname{Cov}\left(X_j, \mathbf{X}_{\mathcal{C}}\right)^T\right] - \\
& \mathbb{E}\left(\log\left[\operatorname{Var}\left(X_j \mid \widetilde{S}(Y)\right) - \operatorname{Cov}\left(X_j, \mathbf{X}_{\mathcal{C}} \mid \widetilde{S}(Y)\right) \times \right.\right. \\
& \qquad \left.\left.\left[\operatorname{Cov}\left(\mathbf{X}_{\mathcal{C}} \mid \widetilde{S}(Y)\right)\right]^{-1} \operatorname{Cov}\left(X_j, \mathbf{X}_{\mathcal{C}} \mid \widetilde{S}(Y)\right)^T\right]\right) \\
=\ & D^*_{j|\mathcal{C},\widetilde{S}(Y)}.
\end{aligned}$$

When $\mathcal{A} \subset \mathcal{C}$ under the augmented model (2.7), the predictor indexed by $j \in \mathcal{C}^c$ has the same conditional distribution across difference slices given $\mathbf{X}_{\mathcal{C}}$. So $\mathbb{E}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, S(Y)\right)$ and $\mathbb{E}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, \widetilde{S}(Y)\right)$ are linear combinations of $\mathbf{X}_{\mathcal{C}}$ that do not depend on $S(Y)$ or $\widetilde{S}(Y)$, and $\operatorname{Var}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, S(Y)\right)$ is a constant that does not depend on $\mathbf{X}_{\mathcal{C}}$ or $S(Y)$. Thus, the equalities hold in this case. □

ACKNOWLEDGEMENTS

We thank Wenxuan Zhong for sharing the mouse embryonic stem cells data set, Tingting Zhang for helpful discussion on the COP procedure, Runze Li and Wei Zhong for providing the R code for DC-SIS.

REFERENCES

Bien, J., Taylor, J. and Tibshirani, R. (2012). A Lasso for hierarchical interactions. *arXiv preprint arXiv:1205.5050*.

Chen, C.-H. and Li, K.-C. (1998). Can SIR be as popular as multiple linear regression? *Statistica Sinica* **8** 289–316.

Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V. B., Wong, E., Orlov, Y. L., Zhang, W., Jiang, J. et al. (2008). Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133** 1106–1117.

Cloonan, N., Forrest, A. R., Kolle, G., Gardiner, B. B., Faulkner, G. J., Brown, M. K., Taylor, D. F., Steptoe, A. L., Wani, S., Bethel, G. et al. (2008). Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nature Methods* **5** 613–619.

Cook, R. D. (2004). Testing predictor contributions in sufficient dimension reduction. *The Annals of Statistics* **32** 1062–1092.

Cook, R. D. (2007). Fisher lecture: Dimension reduction in regression. *Statistical Science* **22** 1–26.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics* **32** 407–499.

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

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

Friedman, J., Hastie, T., Höfling, H. and Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics* **1** 302–332.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A. et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286** 531–537.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association* **86** 316–327.

Li, L. (2007). Sparse sufficient dimension reduction. *Biometrika* **94** 603–613.

Li, L., Cook, R. D. and Nachtsheim, C. J. (2005). Model-free variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67** 285–299.

Li, R., Zhong, W. and Zhu, L. (2012). Feature screening via distance correlation learning. *arXiv preprint arXiv:1205.4701*.

Miller, A. J. (1984). Selection of subsets of regression variables. *Journal of the Royal Statistical Society: Series A (General)* 389–425.

Murphy, T. B., Dean, N. and Raftery, A. E. (2010). Variable selection and updating in model-based discriminant analysis for high dimensional data with food authenticity applications. *The Annals of Applied Statistics* **4** 396.

Ouyang, Z., Zhou, Q. and Wong, W. H. (2009). ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proceedings of the National Academy of Sciences* **106** 21521–21526.

Simon, N. and Tibshirani, R. (2012). A permutation ppproach to testing interactions in many dimensions. *arXiv preprint arXiv:1206.6519*.

Szretter, M. E. and Yohai, V. J. (2009). The sliced inverse regression algorithm as a maximum likelihood procedure. *Journal of Statistical Planning and Inference* **139** 3570–3578.

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

Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences* **99** 6567–6572.

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

Zhang, Y. and Liu, J. S. (2007). Bayesian inference of epistatic interactions in case-control studies. *Nature Genetics* **39** 1167–1173.

Zhong, W., Zeng, P., Ma, P., Liu, J. S. and Zhu, Y. (2005). RSIR: regularized sliced inverse regression for motif discovery. *Bioinformatics* **21** 4169–4175.

Zhong, W., Zhang, T., Zhu, Y. and Liu, J. S. (2012). Correlation pursuit: forward stepwise variable selection for index models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101** 1418–1429.

Department of Statistics
Harvard University
1 Oxford Street, Cambridge
MA 02138, USA
E-mail: bjiang@fas.harvard.edu
jliu@stat.harvard.edu