Abstract

A statistical model is proposed for variable selection with interaction detection under the sliced inverse regression 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 distribution of predictors given the response. A two-stage stepwise procedure based on likelihood ratio test is developed to select relevant predictors and a corresponding independence pre-screening procedure is derived. Various asymptotic properties of the proposed procedure are established and, in particular, its variable selection performance under a diverging number of predictors and sample size is investigated. The performance of the proposed procedure in comparison with some existing method 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 and dimension reduction methods, largely due to the tremendous advance in data collection techniques such as those in biology, internet, and marketing. It has now been widely recognized by both general scientists and quantitative modelers the importance of discovering among many possibilities factors that are truly influential on the responses. Under linear regression models, various regularization methods have been proposed for simultaneously estimating regression coefficients and selecting predictors. Many promising algorithms, such as the Lasso (Tibshirani, 1996; Zou, 2006; Friedman, 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 that first independently selects variables based on their correlations with the
response and then applies variable selection methods in the second step.

The dimension of the variables to be considered can be even larger with the inclusion
of interaction terms between predictors. Consider the following example with a regression
model for a univariate response variable $Y$ and $p$ independent normally distributed predictor
variables $X = (X_1, X_2, \ldots, X_p)^T$,

$$Y = X_1X_2 + \epsilon,$$  \hspace{1cm} (1)

where $\epsilon \sim N(0, 0.1)$. Since there are $\binom{p}{2}$ interactions of order 2, 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, say $p = 1000$, of predictor
variables. Recently, there has been considerable effort in fitting interaction models in the
statistical literatures. For example, Bien et al. (2012) developed hierNet, an extension
of the Lasso to consider interactions in a model if one or both variables are marginally
important (called “hierarchical interactions” by the authors). Li et al. (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.

When the relationship between the response and predictors is beyond linear, the per-
formance of the variable selection methods based on the linear model assumption can be
compromised. In his seminal paper on dimension reduction, Li (1991) proposed a semi-
parametric index model of the form

$$Y = f(\beta_1^T X, \beta_2^T X, \ldots, \beta_K^T X, \epsilon),$$

where $f$ is an unknown link function and $\epsilon$ is a stochastic error independent of $X$. A
sliced inverse regression (SIR) method was developed by Li (1991) to estimate the so-called
sufficient dimension reduction (SDR) directions $\beta_1, \ldots, \beta_K$. 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. Li et al.
(2005) used a backward subset selection method based on various $\chi^2$-tests derived in Cook
(2004). Li (2007) developed sparse SIR (SSIR) algorithm to obtain shrinkage estimates of
the SDR directions. In Zhong et al. (2012), the authors proposed a stepwise procedure
called correlation pursuit (COP) for index models under the SDR framework.

Most of the aforementioned methods are derived from a “forward” model perspective,
that is, a model for the conditional distribution of $Y$ given $X$. When the predictor variables
$X$ can be treated as random, $Y$ and $X$ have a joint distribution and there is a different
modeling perspective based on the conditional distribution of $X$ given $Y$, which is called
an “inverse” model or a “backward” model. Indeed, this inverse modeling perspective has
been taken by several researchers and led to several new developments in variable selection
methods. Cook (2007) proposed various inverse regression models for dimension reduction,
which have deep connections with the SIR method. Szretter and Yohai (2009) showed that
the SIR procedure corresponds to the maximum likelihood estimate where the observations
of $X$ in each slice follows multivariate normal distributions with means in an affine manifold.
Witten and Tibshirani (2009) proposed the Scout method based on covariance regularization
Figure 1: Left panel: contour plot for the joint distribution of $Y$ and $X_1$ in example (1). 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

for a joint model of $Y$ and $X$. Simon and Tibshirani (2012) proposed a permutation-based method for testing interactions by exploring the connection between the “forward” logistic model and the “backward” normal mixture model when the response $Y$ is binary. Another classical method derived from the “inverse” model perspective is Naïve Bayes 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. Raftery and Dean (2006) and Murphy et al. (2010) proposed a variable selection method using Bayesian information criterion (BIC) for model-based discriminant analysis (Fraley and Raftery, 2002). 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 $X$ are discrete. Zhang et al. (2009) developed a Bayesian partition model for expression quantitative trait loci (eQTLs) studies based on a joint model of high-dimensional continuous responses (gene expressions) and discrete predictor variables (genetic markers).

The “inverse” modeling perspective can also shed lights on interaction detections. Figure 1(a) shows the contour plot for the joint distribution of $Y$ and $X_1$ in example (1). If we slice the response into five slices and look at the means and variances of $X_1$ within different slices (as shown in Figure 1(b)), we can see that although $X_1$ is marginally uncorrelated with $Y$ and the conditional means of $X_1$ in different slices are the same, the conditional variances of $X_1$ are quite different across slices. If we take an “inverse” modeling perspective, then instead of screening all the $\binom{p}{2} = O(p^2)$ interactions of order 2 in a regression model, we can
discover the importance of $X_1$ and $X_2$ by examining the conditional variances of $p$ individual predictors given the slices of the response $Y$, which only requires a computational complexity of $O(p)$. Motivated by this observation, in this paper, we will take an “inverse” modeling perspective to attack the problem of interaction detection without imposing model-specific assumptions on the relationship between the response and the predictors.

The rest of the paper is organized as follows. In Section 2.1, we first introduce an “inverse” model for the conditional distribution of $X$ given slices of $Y$. SIR method is presented as a maximum likelihood estimate of the model. A likelihood ratio statistics for selecting relevant predictors is derived in Section 2.2, which is shown to be asymptotically equivalent to the COP statistics. We further augment the model to consider higher order interactions between predictors in Section 2.3. A two-stage stepwise procedure for variable selection and interaction detection that we call Siri is developed and its asymptotic behavior under a diverging number of predictors and sample size is investigated in Section 2.4. In Section 2.5, we further show that the likelihood ratio statistics derived in Siri can be used as an independence screening criterion. Various implementation issues including the choices of slicing schemes and thresholds are discussed in Section 2.6. Simulations and real data examples are reported in Section 3 and 4. Additional remarks in Section 5 conclude the paper. Proofs of the theorems are provided in Appendix A.

2 A sliced inverse regression model with variable selection and interaction detection

Let $Y \in \mathbb{R}$ be a univariate response variable and $X = (X_1, X_2, \ldots, X_p)^T \in \mathbb{R}^p$ be a vector of $p$ continuous predictor variables. $\{(x_i, y_i)\}_{i=1}^n$ are independent observations of $(X, Y)$, where $x_i = (x_{i1}, \ldots, x_{ip})^T$. For discrete (categorical or ordinal) response, we can naturally group $\{y_i\}_{i=1}^n$ into a finite number of classes. For continuous response, which is mostly concerned in this paper, the range of $\{y_i\}_{i=1}^n$ can be divided into $H$ disjoint intervals, called “slices”, which are denoted as $S_1, S_2, \ldots, S_H$. Let $Z$ denote the slice membership of response $Y$, i.e. $Z = h$ if $Y \in S_h$, $1 \leq h \leq H$. For a fixed slicing scheme, we assume that $n_h = |S_h| = s_hn$, where $s_h$’s are fixed and $\sum_{h=1}^H s_h = 1$. The $p$ predictor variables are partitioned into three groups: group 1 contains variables that have different means but the same covariance across $H$ slices, group 2 contains variables that have different means and covariances given group 1 across slices and group 0 contains variables that have the same distribution given group 1 and 2 across slices. Let $I = (I_1, \ldots, I_p)$ indicate the membership of the variables with $I_j = 0, 1, 2$ and $G_k = \{j : I_j = k\}$. $p_k = |G_k|$ ($k = 0, 1, 2$) is the number of variables in each group. Our goal is to infer the variables that have different distributions across slices and thus are associated with the response (that is, the set $G_{1,2} = G_1 \cup G_2 = \{j : I_j > 0\}$).

Let $X_A$ denote random variables in index set $A$. We further assume that variables in group 1 and 2 follow multivariate distributions and a general form of their distributions conditioning on slice $S_h$ can be written as

$$X_{G_{1,2}} | Z = h \sim N\left(\mu_{G_{1,2}}^{(h)}, \Sigma_{G_{1,2}}^{(h)}\right), 1 \leq h \leq H.$$  \hfill (2)
where \( \mu_{G_1}^{(h)} \) is a \((p_1 + p_2)\)-dimension column vector and \( \Sigma_{G_1}^{(h)} \) is a \((p_1 + p_2)\) covariance matrix. Given variables in group 1 and group 2, we assume that the conditional distribution of variables in group 0 is

\[
X_{G_0}^T X_{G_1}, Z = h \sim N \left( \mu_{G_0|G_1}, \Sigma_{G_0|G_1} \right), 1 \leq h \leq H, \tag{3}
\]

where \( \mu_{G_0|G_1} = (\beta_{G_0|G_1})^T \tilde{X}_{G_1} \) with \( \tilde{X}_{G_1} = (1, X_{G_1}^T)^T \), \( \beta_{G_0|G_1} \) is a \((p_1 + p_2 + 1)\) by \( p_0 \) matrix and \( \Sigma_{G_0|G_1} \) is a \( p_0 \) by \( p_0 \) covariance matrix. Note that the above definitions of variables in \( G_{1,2} \) and \( G_0 \) are not unique in the sense that any variable in \( G_0 \) can be treated as a special case of the model (2) with the same mean and variance across slices. Thus we impose the following minimum property on \( G_{1,2} \).

**Definition 1.** Under model (2) and (3), \( G_{1,2} \) satisfies the minimum property if there does not exist a variable \( j \in G_{1,2} \) such that the conditional distribution of \( j \) given \( G_{1,2} - \{j\} \) is the same across slices (i.e., variable \( j \) can also be classified into group 0).

We further assume the following condition on the covariance matrix of variables.

**Condition 1.** Suppose \( \lambda_{\min}(M) \) and \( \lambda_{\max}(M) \) represent, respectively, the smallest and largest eigenvalues of a positive definite matrix \( M \). For the conditional distribution of predictors \( X \in \mathbb{R}^p \) given slice \( S_h \), we assume that

\[
\tau_{\min} \leq \lambda_{\min}(\text{Var}(X|Z = h)) < \lambda_{\max}(\text{Var}(X|Z = h)) \leq \frac{\tau_{\max}}{2},
\]

and that

\[
\lambda_{\max}(\text{Var}(E(X|Z))) \leq \frac{\tau_{\max}}{2},
\]

where \( h = 1, 2, \ldots, H \) and \( \tau_{\max} > \tau_{\min} > 0 \).

Since \( \text{Var}(X) = \text{Var}(E(X|Z)) + E(\text{Var}(X|Z)) \), under Condition 1,

\[
\tau_{\min} \leq \lambda_{\min}(\text{Var}(X)) < \lambda_{\max}(\text{Var}(X)) \leq \tau_{\max}.
\]

Then, We have the following proposition on the identifiability of \( G_{1,2} \) and \( G_0 \):

**Proposition 1.** Under model assumption (2) and (3) and Condition 1, the group of relevant variables, \( G_{1,2} \), that satisfies the minimum property is unique.

We will discuss the partition of \( G_{1,2} \) into \( G_1 \) and \( G_2 \) in the following subsections.

### 2.1 Sliced inverse regression via maximum likelihood

First, we will consider the inclusion of variables into group 1, while group 2 is empty, that is, \( G_{1,2} = G_1 \). In particular, for variables in group 1, we assume that

\[
X_{G_1}, Z = h \sim N \left( \mu_{G_1}^{(h)}, \Sigma_{G_1} \right), 1 \leq h \leq H, \tag{4}
\]

where \( \mu_{G_1}^{(h)} \) is a \( p_1 \)-dimensional vector and \( \Sigma_{G_1} \) is a \( p_1 \) by \( p_1 \) covariance matrix. For the purpose of simultaneous dimension reduction and variable selection, we further assume that \( \mu_{G_1}^{(h)} \) can
be embedded in a \( q \)-dimensional affine space: \( \alpha + \mathcal{V}^q \), where \( \mathcal{V}^q \) is a \( q \)-dimensional subspace of \( \mathbb{R}^{p_1} \) \( (q \leq p_1) \) and \( \alpha \in \mathbb{R}^{p_1} \). \( \mu_{G_i}^{(h)} \) can be also represented as \( \alpha + \Gamma h \), where \( \Gamma \) is a \( p_1 \) by \( q \) matrix the column of which contains the bases of the subspace \( \mathcal{V}^q \) and \( \gamma_h \) is a \( q \)-dimensional vector. Since \( (\Omega O^T)(O \gamma_h) = \Gamma \gamma_h \), for any \( q \) by \( q \) orthogonal matrix \( O \), \( \Gamma \) and \( \gamma_h \) are only identifiable up to orthogonal transformation. Note that this model is a special case of the principle fitted component model proposed by Cook (2007).

Given variables in group 1, we assume that the conditional distribution of variables in group 0 follows model (3) with \( G_{1:2} = G_1 \). Let \( x_A \) denote a \( n \) by \( |A| \) matrix of observed variables in index set \( A \), and let \( x_{hj}^A, 1 \leq j \leq n_h \), denote a \( |A| \)-dimensional column vector of observed variables in index set \( A \) and slice \( S_h, 1 \leq h \leq H \). Define \( B_A = \text{Var}(E(X_A|Z)) \), \( W_A = E(\text{Var}(X_A|Z)) \) and \( \Omega_A = B_A + W_A \), and their corresponding sample estimates

\[
\hat{B}_A = \frac{1}{n} \sum_{h=1}^{H} n_h \left( \bar{x}_h^A - \bar{x}_c^A \right) \left( \bar{x}_h^A - \bar{x}_c^A \right)^T, \quad \text{and} \\
\hat{W}_A = \frac{1}{n} \sum_{h=1}^{H} n_h \sum_{j=1}^{n_h} \left( x_{hj}^A - \bar{x}_h^A \right) \left( x_{hj}^A - \bar{x}_h^A \right)^T, \tag{5}
\]

where \( \bar{x}_c^A = \frac{1}{n} \sum_{h=1}^{H} n_h \bar{x}_j^A \), \( \bar{x}_h^A = \frac{1}{n_h} \sum_{j=1}^{n_h} x_{hj}^A \) and \( \hat{\Omega}_A = \hat{B}_A + \hat{W}_A \).

Szretter and Yohai (2009) showed that the maximum likelihood estimate based on model (4) corresponds to solution of the sliced inverse regression procedure. The results are summarized in the following proposition:

**Proposition 2.** (Szretter and Yohai, 2009) Let \( A = G_1 \), and let \( C \) be the orthogonal matrix \( [c_1, \ldots, c_{p_1}] \), where \( c_j \) is an eigenvector of \( \hat{B}_A^{-1/2} \hat{W}_A \hat{B}_A^{-1/2} = C^T \Theta C \) corresponding to the eigenvalue \( \theta_j \), where \( \theta_1 \geq \theta_2 \geq \ldots \geq \theta_{p_1} \). \( \Theta \) is the diagonal matrix with \( \theta_1, \theta_2, \ldots, \theta_{p_1} \) in the diagonal. \( C_r \) is the matrix with the first \( r \) columns of \( C \).

(a) The maximum likelihood estimate of \( \Sigma_A \) in model (4) is

\[
\hat{\Sigma}_A = \hat{W}_A + \hat{B}_A^{-1/2} C_{p_1-q} C_{p_1-q}^T \hat{B}_A^{-1/2}. \tag{6}
\]

(b) Let \( u_i, 1 \leq i \leq p_1 \), be orthogonal eigenvectors of norm one of \( \hat{\Sigma}_A^{-1/2} \hat{B}_A \hat{\Sigma}_A^{-1/2} \) corresponding to the eigenvalues \( \omega_1 \geq \omega_2 \geq \ldots \geq \omega_{p_1} \). The maximum likelihood estimate of \( \mu^{(h)}_A \) is

\[
\hat{\mu}^{(h)}_A = \hat{\Sigma}_A^{-1/2} U_q U_q^T \hat{\Sigma}_A^{-1/2} \left( \bar{x}^A_h - \bar{x}_c^A \right) + \bar{x}_c^A, 1 \leq h \leq H,
\]

where \( U_q = [u_1, u_2, \ldots, u_q] \). Then, \( \hat{\mu}^{(h)}_A \) is the orthogonal projection, using the norm associated to \( \hat{\Sigma}_A \), of \( \bar{x}^A_h - \bar{x}_c^A \) on the \( q \)-dimensional affine subspace \( \bar{x}_c^A + \mathcal{V}^q \), where \( \mathcal{V}^q \) is spanned by \( \hat{\Sigma}_A^{-1/2} u_1, \hat{\Sigma}_A^{-1/2} u_2, \ldots, \hat{\Sigma}_A^{-1/2} u_q \).

(c) \( \hat{\Sigma}_A^{-1/2} u_j \) is an eigenvector of \( \hat{B}_A \hat{W}_A^{-1} \) corresponding to eigenvalue \( 1/\theta_{p_1-i+1}, 1 \leq i \leq p_1 \).

(d) \( \hat{\Sigma}_A \) can also be written as

\[
\hat{\Sigma}_A = \frac{1}{n} \sum_{h=1}^{H} n_h \sum_{j=1}^{n_h} \left( x_{hj}^A - \hat{\mu}^{(h)}_A \right) \left( x_{hj}^A - \hat{\mu}^{(h)}_A \right)^T.
\]
conditioning on the memberships of the other variables, the likelihood ratio for testing
where the second equality follows from

(e) The $j$th principle direction estimated by sliced inverse regression algorithm coincides with $\Phi^{-1}_A \hat{Z}_A^{1/2} u_i$, $1 \leq i \leq q$.

Proposition 2 shows the equivalence between the sliced inverse regression procedure and the maximum likelihood estimate based on model (4) of group 1 variables. Given the variables in group 1, the maximum likelihood estimate of model (3) can be obtained by multivariate regression estimates, i.e.

$$\hat{\beta}_{G_0|G_{1,2}} = \left( \bar{x}_{G_{1,2}}^T \bar{x}_{G_{1,2}} \right)^{-1} \bar{x}_{G_{1,2}}^T \bar{x}_{G_0},$$

(7)

and

$$\hat{\Sigma}_{G_0|G_{1,2}} = \frac{1}{n} \bar{x}_{G_0}^T \left( I_n - \bar{x}_{G_{1,2}} \left( \bar{x}_{G_{1,2}}^T \bar{x}_{G_{1,2}} \right)^{-1} \bar{x}_{G_{1,2}}^T \right) \bar{x}_{G_0},$$

(8)

where $\bar{x}_{G_k} = [1_n, x_{G_k}]$, $1_n$ is a $n$-dimensional column vector of 1, $I_n$ is a $n$ by $n$ identity matrix and $G_{1,2} = G_1$ for the purpose of discussion in this subsection.

2.2 Likelihood ratio tests for selecting variables

Suppose the $j$th variable $X_j$ is being considered for inclusion into group 1 or group 0, and $A_1$ contains indexes of variables that have already been selected into group 1, $A_{1+} = A_1 \cup \{j\}$, and $A^c = \{1, 2, \ldots, p\} - A$. Denote model (4) by $M_1$ and model (3) by $M_0$. Then,

$$\frac{P(X|G_1 = A_{1+}, G_0 = A_{1+}^c)}{P(X|G_1 = A_1, G_0 = A_1^c)} = \frac{P_{M_1}(X_{A_{1+}})}{P_{M_1}(X_{A_1})} \frac{P_{M_0}(X_{A_{1+}^c}|X_{A_{1+}})}{P_{M_0}(X_{A_1^c}|X_{A_1})} = \frac{P_{M_1}(X_{A_{1+}})}{P_{M_1}(X_{A_1})} \frac{P_{M_0}(X_j|X_{A_1})}{P_{M_0}(X_j|X_{A_1})},$$

where the second equality follows from

$$P_{M_0}(X_{A_{1+}^c}|X_{A_1}) = P_{M_0}(X_{A_{1+}^c}|X_{A_{1+}}) P_{M_0}(X_j|X_{A_1}).$$

Conditioning on the memberships of the other variables, the likelihood ratio for testing $H_0 : I_j = 0$ versus $H_1 : I_j = 1$ is given by

$$\hat{L}_{j|A_1} = \frac{P_{\hat{M}_1}(X_{A_{1+}})}{P_{\hat{M}_1}(X_{A_1}) P_{\hat{M}_0}(X_j|X_{A_1})},$$

where $\hat{M}_1$ and $\hat{M}_0$ are maximum likelihood estimates of model $M_1$ and $M_0$, respectively. The representation and the distribution of the test statistic is given by the following theorem:

**Theorem 3.** Conditioning on the memberships of the other variables, the log-likelihood ratio statistic for testing $H_0 : I_j = 0$ versus $H_1 : I_j = 1$ can be written as

$$\hat{l}_{j|A_1} = 2 \log(\hat{L}_{j|A_1}) = -n \left( \log \det \left( \hat{\Omega}_{A_{1+}}^{-1} \hat{\Sigma}_{A_{1+}} \right) - \log \det \left( \hat{\Omega}_A^{-1} \hat{\Sigma}_A \right) \right),$$

where $\hat{\Omega}_A$ and $\hat{\Sigma}_A$ are given by (5) and (6), respectively.
(a) Let \( \hat{\lambda}_i^A \) be the \( i \)th largest eigenvalue of \( \hat{\Omega}_A^{-1} \hat{B}_A \). Then the log-likelihood ratio statistic can also be written as

\[
\hat{l}_{j|A_i} = -n \left( \sum_{i=1}^q \log \left( 1 - \hat{\lambda}_i^{A_i} \right) - \sum_{i=1}^q \log \left( 1 - \hat{\lambda}_i^{A_i} \right) \right),
\]

where \( q \leq |A_i| \) is the dimension of subspace \( \mathcal{Y}^q \) defined in model (4).

(b) For any fixed slicing scheme, let \( \lambda_i^A \) be the \( i \)th largest eigenvalue of \( \Omega_A^{-1} B_A \). We further assume that \( \lambda_1 > \lambda_2 > \ldots > \lambda_q > 0 \). Then, for any \( j \in A_i \), given that \( G_1 \subset A_1 \),

\[
\hat{l}_{j|A_i} \approx \sum_{i=1}^q \frac{n \left( \hat{\lambda}_i^{A_i} - \lambda_i^{A_i} \right)}{1 - \lambda_i^{A_i}},
\]

which is asymptotically equivalent to the COP test statistic defined in Zhong et al. (2012) and has an asymptotic distribution of \( \chi^2_q \).

(c) Let \( \Sigma_{A^c|A} = \text{Var}(X_{A^c}|X_A) \) and \( D_{A^c|A} \) represent a diagonal matrix with the same diagonal elements as \( \Sigma_{A^c|A} \). Define \( \Psi_A = D_{A^c|A}^{-1/2} \Sigma_{A^c|A} D_{A^c|A}^{-1/2} \). Let \( A = A_1 \), and then under the same conditions in (b),

\[
\left( \hat{l}_{j|A} \right)_{j \in A^c} \overset{D}{\to} \left( \sum_{i=1}^q z_{ij}^2 \right)_{j \in A^c}, \quad \text{and} \quad \max_{j \in A^c} \left( \hat{l}_{j|A} \right) \overset{D}{\to} \max_{j \in A^c} \left( \sum_{i=1}^q z_{ij}^2 \right),
\]

where \( z_i = (z_{ij})_{j \in A^c} \) for \( i = 1, 2, \ldots, q \) are independent random vectors and each \( z_i \) follows a multivariate normal distribution with mean \( \mathbf{0} \) and covariance matrix \( \Psi_A \).

(d) For \( A = A_1 \) and \( j \in A_i \), as \( n \to \infty \),

\[
\hat{l}_{j|A} \quad n \to \quad \text{log} \left( 1 + \text{Var} (M_j) - \text{Cov} (M_j, X_A) \left[ \text{Var} (X_A) \right]^{-1} \text{Cov} (M_j, X_A)^T \right),
\]

where \( M_j = M_j (X_A, Z) = E (X_j | X_A, Z) \), \( V_j = V_j (X_A, Z) = \text{Var} (X_j | X_A, Z) \), and \( V_j \) is a constant that does not depend on \( X_A \) or \( Z \). Furthermore, as \( n \to \infty \),

\[
\hat{l}_{j|A} \quad n \to \quad 0 \quad \text{if and only if} \quad M_j = E (X_j | X_A, Z) = E (X_j | X_A).
\]

### 2.3 An augmented model with interaction detection

Next, we will consider the inclusion of variables into group 2 from group 0. Given the set of variables in group 1, \( X_{G_1} \), we assume that the conditional distribution of variables in group 2, \( X_{G_2} \), is

\[
X_{G_2} | X_{G_1}, Z = h \sim N (\mu_{G_2|G_1}^{(h)}, \Sigma_{G_2|G_1}^{(h)}), \quad 1 \leq h \leq H,
\]

where \( \mu_{G_2|G_1}^{(h)} = (\beta_{G_2|G_1}^{(h)})^T \bar{X}_{G_1} \) with \( \bar{X}_{G_1} = (1, X_{G_1}^T)^T \), \( \beta_{G_2|G_1}^{(h)} \) is a \( p_1 + 1 \) by \( p_2 \) matrix and \( \Sigma_{G_2|G_1}^{(h)} \) is a \( p_2 \) by \( p_2 \) covariance matrix. Given \( G_{1,2} = G_1 \cup G_2 \), we assume that the conditional distribution of variables in group 0 is given by (3).
Given variables in group $G$ and already been selected into group $\tilde{G}$, we have the following theorem on the characteristics of the test statistic.

The maximum likelihood estimate based on model (11) is given by multivariate regression estimates within each slice, i.e.

$$\hat{\beta}^{(h)}_{G_2|G_1} = \left( x^{(h)}_G x^{(h)}_G \right)^{-1} x^{(h)}_G x^{(h)}_{G_2},$$

and

$$\hat{\Sigma}^{(h)}_{G_2|G_1} = \frac{1}{n_h} x^{(h)}_{G_2} \left( I_{n_h} - x^{(h)}_{G_1} \left( x^{(h)}_{G_1} x^{(h)}_{G_1} \right)^{-1} x^{(h)}_{G_1} \right) x^{(h)}_{G_2},$$

where $x^{(h)}_{G_k} = [1_{n_h}, x^{(h)}_{G_k}]$, $x^{(h)}_{G_k}$ is a $n_h$ by $|G_k|$ matrix of observed variables in index set $G_k$ and slice $S_h$, $1_{n_h}$ is a $n_h$-dimensional column vector of 1 and $I_{n_h}$ is a $n_h$ by $n_h$ identity matrix. Given variables in group 1 and group 2, the maximum likelihood estimates of $\hat{\beta}_{G_0|G_{1:2}}$ and $\hat{\Sigma}_{G_0|G_{1:2}}$ in model (3) are given by (7) and (8), respectively.

Suppose the $j$th variable $X_j$ is being considered for inclusion into group 2 or group 0 given the current set of group 1 variables, $A_1$. Let $A_2$ be the index set of variables that have already been selected into group 2, $A_{2+} = A_2 \cup \{ j \}$, $A_{1:2+} = A_1 \cup A_{2+}$ and $A_{1:2} = A_1 \cup A_2$. Denote model (11) by $\mathcal{M}_2$ and model (3) by $\mathcal{M}_0$. Then,

$$\frac{P \left( X|G_2 = A_{2+}, G_1 = A_1, G_0 = A_{1:2+} \right)}{P \left( X|G_2 = A_2, G_1 = A_1, G_0 = A_{1:2} \right)} = \frac{P_{\mathcal{M}_2} (X_j|X_{A_{1:2}})}{P_{\mathcal{M}_0} (X_j|X_{A_{1:2}})}.$$

Conditioning on the memberships of the other variables, the likelihood ratio for testing $H_0 : I_j = 0$ versus $H_1 : I_j = 2$ is given by

$$\hat{L}^{\text{aug}}_{j|A_{1:2}} = \frac{P_{\mathcal{M}_2} (X_j|X_{A_{1:2}})}{P_{\mathcal{M}_0} (X_j|X_{A_{1:2}})},$$

where $\mathcal{M}_2$ and $\mathcal{M}_0$ are maximum likelihood estimates of model $\mathcal{M}_1$ and $\mathcal{M}_0$, respectively. Then, we have the following theorem on the characteristics of the test statistic.

**Theorem 4.** Conditioning on the memberships of the other variables, the log-likelihood ratio statistic for testing $H_0 : I_j = 0$ versus $H_1 : I_j = 2$ can be written as

$$\hat{\ell}^{\text{aug}}_{j|A_{1:2}} = 2 \log (\hat{L}^{\text{aug}}_{j|A_{1:2}}) = -n \left( \sum_{h=1}^{H} s_h \log \left[ \hat{\sigma}_j^{(h)} \right]^2 - \log \hat{\sigma}_j^2 \right), \quad (12)$$

where

$$\hat{\sigma}_j^2 = \frac{1}{n} x_j^T \left( I_n - \bar{x}_{A_{1:2}} \left( \bar{x}_{A_{1:2}}^T \bar{x}_{A_{1:2}} \right)^{-1} \bar{x}_{A_{1:2}}^T \right) x_j,$$

and

$$\left[ \hat{\sigma}_j^{(h)} \right]^2 = \frac{1}{n_h} x_j^{(h)}^T \left( I_{n_h} - \bar{x}_{A_{1:2}}^{(h)} \left( \bar{x}_{A_{1:2}}^{(h)} x_{G_{1:2}} \right)^{-1} x_{A_{1:2}}^{(h)} \right) x_j^{(h)}.$$

(a) Let $A = A_{1:2}$. For any $j \in A^c$, given that $G_{1:2} \subset A = A_{1:2}$ and $A_{1:2+} = A_{1:2} \cup \{ j \}$,

$$\hat{\ell}^{\text{aug}}_{j|A} \sim n \left( \log \left( 1 + \frac{Q_0}{\sum_{h=1}^{H} Q_h} \right) - \frac{1}{n} \sum_{h=1}^{H} Q_h \log \left( \frac{Q_h}{\sum_{h=1}^{H} Q_h} \right) \right),$$

where
where $Q_0 \sim \chi^2_{(H-1)d}$, $Q_h \sim \chi^2_{n_h - d}$, $h = 1, \ldots, H$, are mutually independent, and $d = |A| + 1$.

(b) For any fixed slicing scheme and under the same condition in (a), $\hat{l}_{j|A}^{\text{aug}}$ has an asymptotic distribution of $\chi^2_{(H-1)(d+1)}$.

(c) Let $\Sigma_{A^c|A} = \text{Var}(X_{A^c}|X_A)$ and $D_{A^c|A}$ represent a diagonal matrix with the same diagonal elements as $\Sigma_{A^c|A}$. Define $\Psi_A = D_{A^c|A}^{-1/2} \Sigma_{A^c|A} D_{A^c|A}^{-1/2}$ and $\tilde{\Psi}_A = [\tilde{\Psi}_{ij}]$. Let $A = A_{1:2}$, and then under the same conditions in (b),

$$
(\hat{l}_{j|A}^{\text{aug}})_{j \in A^c} \overset{D}{\to} \left( \sum_{i=1}^{(H-1)d} z_{ij}^2 + \sum_{i=1}^{(H-1)} z_{ij}^2 \right)_{j \in A^c}, \quad \text{and} \quad \max_{j \in A^c} (\hat{l}_{j|A}^{\text{aug}}) \overset{D}{\to} \max_{j \in A^c} \left( \sum_{i=1}^{(H-1)d} z_{ij}^2 + \sum_{i=1}^{(H-1)} z_{ij}^2 \right),
$$

where $z_i = (z_{ij})_{j \in A^c}$ for $i = 1, 2, \ldots, (H-1)d$ and $\tilde{z}_i = (\tilde{z}_{ij})_{j \in A^c}$ for $i = 1, 2, \ldots, (H-1)$ are independent random vectors, each $z_i$ follows a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\Psi_A$, and each $\tilde{z}_i$ follows a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\tilde{\Psi}_A$.

(d) For $A = A_{1:2}$ and $j \in A^c$, as $n \to \infty$,

$$
\frac{\hat{l}_{j|A}^{\text{aug}}}{n} \overset{p}{\to} \log \left( 1 + \frac{\text{Var}(M_j) - \text{Cov}(M_j, X_A)[\text{Var}(X_A)]^{-1} \text{Cov}(M_j, X_A)^T}{E(V_j)} \right) + \log (EV_j) - E \log (V_j),
$$

where $M_j = M_j(X_A, Z) = E(X_j|X_A, Z)$ and $V_j = V_j(X_A, Z) = \text{Var}(X_j|X_A, Z)$. Furthermore, as $n \to \infty$,

$$
\frac{\hat{l}_{j|A}^{\text{aug}}}{n} \overset{p}{\to} 0 \text{ if and only if } M_j = E(X_j|X_A, Z) = E(X_j|X_A) \text{ and } V_j = \text{Var}(X_j|X_A, Z) = \text{Var}(X_j|X_A).
$$

### 2.4 Siri: a two-stage stepwise procedure for variable selection

To combine the test statistics in Section 2.2 and Section 2.3, we propose a two-stage variable selection procedure. In the first stage, we aim to select variables in $G_1$ with forward selection and backward elimination using the likelihood ratio test derived under model (4) in Section 2.2. Next, given variables selected in the first stage, we continue to select variables in $G_2$ in the second stage using the likelihood ratio test under the augmented model (11) and (3). Note that model (4) is nested in model (11) by restricting $\Sigma_{A^c|A}^{(h)}$ to be the same across slices and $\mu_{G_2|G_1}^{(h)}$ embedded in a lower dimensional-space. The two models compensate each other in terms of bias-variance trade-off. Model (4) is simpler and more powerful when the response is driven by some linear combinations of covariates, while model (11) is useful in detecting more complex relationships such as heteroscedastic effect and interaction terms among covariates. Also note that if we set group 1 variable to be empty (that is, $q = 0$ in the previous section), then we will be assuming a general form of the model as in (2) and skipping the first stage. Below we summarize detailed steps of two stages:

- **Stage 1.** set the dimension $q$ in model (4) and the threshold values $u_u$ and $u_d$, randomly select $q + 1$ predictors into $A_1$, and then iterate the following steps until no more addition or deletion of predictors can be performed:
1. in the addition step, given current set of selected predictors in $\mathcal{A}_1$, find $j_a$ such that $\hat{\ell}_{j_a|\mathcal{A}_1} = \max_{j \in \mathcal{A}_1^j} \hat{\ell}_{j|\mathcal{A}_1}$, where $\hat{\ell}_{j|\mathcal{A}_1}$ is given by (9); and if $\hat{\ell}_{j_a|\mathcal{A}_1} > u_a$, add $j_a$ to $\mathcal{A}_1$, i.e. let $\mathcal{A}_1 = \mathcal{A}_1 + \{j_a\}$.

2. in the deletion step, given current set of selected predictors in $\mathcal{A}_1$ and $\mathcal{A}_{1/j} = \mathcal{A}_1 - \{j\}$, find $j_d$ such that $\hat{\ell}_{jd|\mathcal{A}_{1/j}} = \min_{j \in \mathcal{A}_1} \hat{\ell}_{j|\mathcal{A}_{1/j}}$; and if $\hat{\ell}_{jd|\mathcal{A}_{1/j}} < u_d$, delete $j_d$ from $\mathcal{A}_1$, i.e. let $\mathcal{A}_1 = \mathcal{A}_1 - \{j_d\}$.

- **Stage 2.** set the threshold values $v_a$ and $v_d$, let $\mathcal{A}_1$ be the predictors selected in Stage 1 and let $\mathcal{A}_2 = \emptyset$, and then iterate the following steps until no more addition or deletion of predictors can be performed:

1. in the addition step, given $\mathcal{A}_1$ and current set of selected predictors in $\mathcal{A}_2$, find $j_a$ such that $\hat{\ell}^{\text{aug}}_{j_a|\mathcal{A}_1,2} = \max_{j \in \mathcal{A}_1^j} \hat{\ell}^{\text{aug}}_{j|\mathcal{A}_1,2}$, where $\hat{\ell}^{\text{aug}}_{j|\mathcal{A}_1,2}$ is given by (12); and if $\hat{\ell}^{\text{aug}}_{j_a|\mathcal{A}_1,2} > v_a$, add $j_a$ to $\mathcal{A}_2$, i.e. let $\mathcal{A}_2 = \mathcal{A}_2 + \{j_a\}$.

2. in the deletion step, given $\mathcal{A}_1$, current set of selected predictors in $\mathcal{A}_2$ and $\mathcal{A}_{1,2/j} = \mathcal{A}_{1,2} - \{j\}$, find $j_d$ such that $\hat{\ell}^{\text{aug}}_{jd|\mathcal{A}_{1,2/j}} = \min_{j \in \mathcal{A}_2} \hat{\ell}^{\text{aug}}_{j|\mathcal{A}_{1,2/j}}$; and if $\hat{\ell}^{\text{aug}}_{jd|\mathcal{A}_{1,2/j}} < v_d$, delete $j_d$ from $\mathcal{A}_2$, i.e. let $\mathcal{A}_2 = \mathcal{A}_2 - \{j_d\}$.

We called this two-stage procedure sliced inverse regression with interaction detection, or Siri for short.

### 2.4.1 Selection consistency of the first stage

We first consider the case when variables are either from group 1 or group 0. Two more conditions are needed for the consistency results we state below.

**Condition 2.** Under model (4), for variable $j \in \mathcal{G}_1$, the conditional distribution of $j$ given the other variables in $\mathcal{G}_1$ and slice $S_h$ can be written as

$$X_j|X_{\mathcal{G}_1\setminus\{j\}}, Z = h \sim N \left( \eta_j^{(h)} + \beta_j^{(h)} X_{\mathcal{G}_1\setminus\{j\}}, \delta_j^{(h)} \right), \ 1 \leq h \leq H.$$  

Let $\eta_j^Z = \eta_j^{(h)}$ when $Z = h$. We assume that there exists $\kappa > 0$ and $\xi_0 > 0$ such that

$$\text{Var} \left( \eta_j^Z \right) \geq \kappa n^{-\xi_0},$$

for any $j \in \mathcal{G}_1$.

**Condition 3.** $\lim_{n \to \infty} (p) = \infty$ and $p = o(n^{\rho_0})$ with $\rho_0 > 0$ and $2\rho_0 + 2\xi_0 < 1$, where $\xi_0$ is defined in Condition 2.

We have the following theorem on the selection consistency of the first stage:

**Theorem 5.** Given a fixed slicing scheme, we assume Condition 1, 2 and 3 and model (4) for variables in group 1 and model (3) with $\mathcal{G}_2 = \emptyset$ for variables in group 0. Let $\mathcal{A}_1$ be the set of currently selected predictors in group 1. Then there exists a constant $\nu$ such that

$$P \left( \min_{\mathcal{A}_1 : \mathcal{A}_1 \cap \mathcal{G}_1 \neq \emptyset} \max_{j \in \mathcal{A}_1 \cap \mathcal{G}_1} \hat{\ell}_{j|\mathcal{A}_1} \geq \nu n^{1-\xi_0} \right) \to 1,$$

(13)
as $n \to \infty$, and
\[
P \left( \max_{A_1 \cup G} \max_{\hat{A}_1 \cap G} \hat{I}_{\hat{A}_1} < Cn^{-\xi_0} \right) \to 1 \tag{14}
\]
for any positive constant $C$ with $n \to \infty$.

Remark 1. If we choose $u_a = \nu n^{1-\xi_0}$ and $u_d = \nu n^{1-\xi_0}/2$, result (13) in Theorem 5 guarantees that, asymptotically, the addition step in Stage 1 will not stop selecting variables until all the true predictors in group 1 have been included. Moreover, once all the true predictors in group 1 have been included, according to result (14), all the redundant variables in group 0 will be removed from the selected variables.

### 2.4.2 Selection consistency of the second stage

We first define a set of variables that can be “detected” in the second stage:

**Definition 2.** A set of variables, denoted as $\mathcal{D}^{(k)}$, is said to be $k$th-wave detectable ($k \geq 1$) with respect to $\xi_0 > 0$ if $\mathcal{D}^{(k)} \cap \left( \bigcup_{l=1}^{k-1} \mathcal{D}^{(l)} \right) = \emptyset$, and for any set $\mathcal{A}$ satisfying $\left( \bigcup_{l=1}^{k-1} \mathcal{D}^{(l)} \right) \subset \mathcal{A}$ and $\mathcal{A}^c \cap \mathcal{D}^{(k)} \neq \emptyset$, either there exists $\kappa_1 > 0$ such that
\[
\max_{j \in \mathcal{A} \cap \mathcal{D}^{(k)}} \left[ \frac{\text{Var}(M_j) - \text{Cov}(M_j, \mathbf{X}_\mathcal{A}) [\text{Var}(\mathbf{X}_\mathcal{A})]^{-1} \text{Cov}(M_j, \mathbf{X}_\mathcal{A})^T}{\mathbb{E}(V_j)} \right] \geq \kappa_1 n^{-\xi_0}, \tag{15}
\]
or there exists $\kappa_2 > 0$ such that
\[
\max_{j \in \mathcal{A} \cap \mathcal{D}^{(k)}} \left[ \log(\mathbb{E}V_j) - \mathbb{E}\log(V_j) \right] \geq \kappa_2 n^{-\xi_0}
\]
where $M_j = E(X_j|\mathbf{X}_\mathcal{A}, Z)$ and $V_j = \text{Var}(X_j|\mathbf{X}_\mathcal{A}, Z)$. Let $\mathcal{D}^{(0)} = \emptyset$ and $\mathcal{D} = \left( \bigcup_{l=0}^{\infty} \mathcal{D}^{(l)} \right)$ denote the set of detectable variables.

Remark 2. Under model (4) and model (3) with $\mathcal{G}_2 = \emptyset$, if Condition 2 is satisfied, then variables in group 1 are first-wave detectable. Note that for $j \in \mathcal{G}_1$, $V_j = \text{Var}(X_j|\mathbf{X}_\mathcal{A}, Z) = E(V_j)$ is a constant that does not depend on $\mathbf{X}_\mathcal{A}$ or $Z$, and in this case we only concern about the condition (15) in the definition above. In general, according to Theorem 4,

$\text{Var}(M_j) - \text{Cov}(M_j, \mathbf{X}_\mathcal{A}) [\text{Var}(\mathbf{X}_\mathcal{A})]^{-1} \text{Cov}(M_j, \mathbf{X}_\mathcal{A})^T = 0,$

and

$\log(\mathbb{E}V_j) - \mathbb{E}\log(V_j) = 0,$

if and only if $M_j = E(X_j|\mathbf{X}_\mathcal{A}, Z) = E(X_j|\mathbf{X}_\mathcal{A})$ and $V_j = \text{Var}(X_j|\mathbf{X}_\mathcal{A}, Z) = \text{Var}(X_j|\mathbf{X}_\mathcal{A})$, which are satisfied by group 0 variables under model (3) given that $\mathcal{G}_1 \cup \mathcal{G}_2 \subset \mathcal{A}$. If there is a set $\mathcal{A}$ such that the conditions in Definition 2 does not hold for variable $j$, then either $X_j$ is highly correlated with some variables in set $\mathcal{A}$ or conditioning on $\mathbf{X}_\mathcal{A}$, the differences between conditional means of $X_j$ and the differences between conditional variances of $X_j$ are minimal across different slices. Consider a simple example when there are only two variables...
and two slices, if the conditional distributions of variables \((X_1, X_2)\) given slice \(Z = 1\) and \(Z = 2\) are
\[
\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \mid Z = 1 \sim \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \quad \text{and} \quad \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \mid Z = 2 \sim \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix},
\]
then neither variables can be identified by the stepwise procedures that select one variable at a time and \((X_1, X_2)\) are not “detectable”.

**Theorem 6.** Given a fixed slicing scheme, we assume Condition 1, Condition 3 and model assumptions (4), (11) and (3) for variables in group 1, group 2 and group 0. Let \(A_{1,2}\) be the set of currently selected predictors in group 1 and group 2. If all the variables in group 1 and group 2 are detectable, i.e. \(G_{1,2} = D\), then there exists a constant \(\nu\) such that
\[
P \left( \min_{A_{1,2} = A_{1,2}^{\prime} \cap G_{1,2} \neq \emptyset} \max_{j \in A_{1,2}^{\prime}} \hat{\mu}_{1,2}^{aug} j \mid A_{1,2} \geq \nu n^{1-\xi_0} \right) \to 1, \tag{16}
\]
as \(n \to \infty\), and
\[
P \left( \max_{A_{1,2} = A_{1,2}^{\prime} \cap G_{1,2} \neq \emptyset} \max_{j \in A_{1,2}^{\prime}} \hat{\mu}_{1,2}^{aug} j \mid A_{1,2} \leq C n^{1-\xi_0} \right) \to 1, \tag{17}
\]
for any positive constant \(C\) with \(n \to \infty\).

**Remark 3.** If we choose \(\nu_a = \nu n^{1-\xi_0}\) and \(\nu_d = \nu n^{1-\xi_0}/2\), result (16) in Theorem 6 guarantees that, asymptotically, the addition step in Stage 2 will not stop selecting variables until all the true predictors in group 1 and group 2 have been included. Moreover, once all the true predictors in group 1 and group 2 have been included, according to result (17), all the redundant variables in group 0 will be removed from the selected variables.

### 2.5 Sure independence screening with Siri

When the number of predictors is extremely large, the performance of Siri can be compromised. Fan and Lv (2008) proposed a two-step approach, sure independence screening (SIS), to attack so-called ultrahigh dimensionality. The first step is to perform screening to reduce the dimensionality from ultrahigh to high or moderately high, and then, in the second step, variable selection methods are applied to identify the true predictors. The same approach can be used for the variable selection procedure described in the previous section. To be more specific, let
\[
\hat{l}_{1,2}^{aug} = -n \left( \sum_{h=1}^{H} s_h \log \left[ \hat{\sigma}_{j}^{(h)} \right] - \log \hat{\sigma}_{j}^{2} \right), 1 \leq j \leq p
\]
which is calculated by Siri during the first iteration of the second stage, and
\[
\bar{l}_{j}^{aug} = \lim_{n \to \infty} \frac{\hat{l}_{1,2}^{aug}}{n} = \log \left( \sigma_{j}^{2} \right) - \sum_{h=1}^{H} s_h \log \left[ \hat{\sigma}_{j}^{(h)} \right]^{2}
\]
\[
= \log \left[ 1 + \frac{\text{Var} \left( E(X_j | Z) \right)}{E \left( \text{Var}(X_j | Z) \right)} \right] + \log \left[ \frac{E \left( \text{Var}(X_j | Z) \right)}{E \left( \text{Var}(X_j | Z) \right)} \right] - E \left[ \log \left( \text{Var}(X_j | Z) \right) \right],
\]
13
where

\[ \hat{\sigma}_j^2 = \frac{1}{n} x_j^T \left( I_n - \mathbf{1}_n \mathbf{1}_n^T \right) x_j, \]

\[ \left[ \hat{\sigma}_j^{(h)} \right]^2 = \frac{1}{n_h} x_j^{(h)T} \left( I_{n_h} - \mathbf{1}_{n_h} \mathbf{1}_{n_h}^T \right) x_j^{(h)}, \]

and \( \sigma_j^2 = \text{Var}(X_j) \), \( \left[ \sigma_j^{(h)} \right]^2 = \text{Var}(X_j|Z = h) \). We define

\[ \tilde{M}_{\xi,\kappa} = \{ j : \tilde{r}_{j \text{aug}} \geq \kappa n^{1-\xi}, \text{ for } 1 \leq j \leq p \}, \]

where \( \kappa \) and \( \xi \) are pre-specified threshold values. Under the following conditions, by appropriate choosing \( \kappa \) and \( \xi \), the sure screening property holds for \( \tilde{M}_{\xi,\kappa} \), while the size of \( \tilde{M}_{\xi,\kappa} \) is controlled.

**Condition 4.** For \( j \in \mathcal{G}_{1,2} \), there exists \( \xi_0 > 0 \) such that either

\[ \frac{\text{Var}(E(X_j|Z))}{E(\text{Var}(X_j|Z))} \geq \kappa_1 n^{-\xi_0}, \]

or

\[ \log \left[ E(\text{Var}(X_j|Z)) \right] - E \left[ \log (\text{Var}(X_j|Z)) \right] \geq \kappa_2 n^{-\xi_0} \]

with \( \kappa_1 > 0 \) and \( \kappa_2 > 0 \).

Here, \( \xi_0 \) controls the strengths of true predictors so that they do not decreases to 0 too fast. This condition also rules out the situation in which an important variable has the same marginal distribution across slices, but jointly associated with the response when other variables are considered. In the following condition, we consider the case when the total number of predictors grows exponentially in sample size, while the number of true predictors is “sparse” and grows moderately with sample size.

**Condition 5.** \( \lim_{n \to \infty} (p) = \infty \) and \( \log(p) = o(n^{\gamma_0}) \) with \( \gamma_0 > 0 \) and \( \gamma_0 + 2\xi_0 < 1 \), where \( \xi_0 \) is defined in Condition 4. Furthermore, the number of true predictors of the response \( |\mathcal{G}_{1,2}| \leq cn^{\eta_0} \) with \( \eta_0 > 0 \), \( \eta_0 + \xi_0 < 1 \) and a positive constant \( c \).

Then, we can establish the following sure screening property:

**Theorem 7.** Under Condition 1, Condition 4, Condition 5, and model assumption (2) and (3), (a) there exists \( \kappa_0 > 0 \) such that

\[ P \left( \mathcal{G}_{1,2} \subset \tilde{M}_{\xi_0,\kappa_0} \right) \to 1 \]

as \( n \to \infty \). (b) Furthermore, there exists \( C > 0 \) such that

\[ P \left( \left| \tilde{M}_{\xi_0,\kappa_0} \right| \leq C n^{\xi_0+\eta_0} \right) \to 1, \]

with \( \xi_0 + \eta_0 < 1 \) as \( n \to \infty \).
Under the same condition in Theorem 7, if we rank the variables according to \( \hat{\beta}_{lj}^{\text{aug}}, 1 \leq j \leq p \), then the first \( n - 1 \) or \( [n/\log(n)] \) variables have a high probability to include the true predictors (almost surely as \( n \to \infty \)). Thus, for the problem of ultrahigh dimensionality, we can first use a sure independence screening stage to reduce the dimensionality from \( p \) to a scale below sample size, say \( [n/\log(n)] \). Then, we apply the two-stage procedure of Siri to choose a more refined model. We call the above procedure SIS-Siri.

As aforementioned, an important predictor that has the same marginal distribution across slices cannot be picked up by the independence screening method. Consider another simple example when there are only two variables and two slices, if the conditional distributions of variables \( (X_1, X_2) \) given slice \( Z = 1 \) and \( Z = 2 \) are

\[
\left( \begin{array}{c} X_1 \\ X_2 \end{array} \right) \bigg| Z = 1 \sim \left( \begin{array}{c} 1 \\ 0 \end{array} \right), \left( \begin{array}{c} 1 \\ 1 \end{array} \right) \right), \quad \text{and} \quad \left( \begin{array}{c} X_1 \\ X_2 \end{array} \right) \bigg| Z = 2 \sim \left( \begin{array}{c} -1 \\ 0 \end{array} \right), \left( \begin{array}{c} 1 \\ 1 \end{array} \right) \right),
\]

then variable \( X_2 \) has the same distribution in \( Z = 1 \) and \( Z = 2 \) but the conditional distributions of \( X_2 \) given \( X_1 \) are different in two slices, that is, \( X_2 \in \mathcal{G}_{1,2} \). However, an independence screening method can only pick up variable \( X_1 \), but not \( X_2 \). Fan and Lv (2008) advocated an iterative procedure that alternate between a large-scale screening and a moderate-scale variable selection. The same idea can be applied to enhance the performance of SIS-Siri. The whole variable selection procedure, called ISIS-Siri, iterates as follows:

- In the first iteration, rank all the variables according to \( \{\hat{\beta}_{lj}^{\text{aug}}, 1 \leq j \leq p\} \) and reduce the number of variables by taking highest ranked \( [n/\log(n)] \) variables as set \( \mathcal{R}_1 \). Then, run the two-stage procedure of Siri on \( \mathcal{R}_1 \) to select a subset of variables \( \mathcal{S}_1 \), and let \( \mathcal{A}_1 = \mathcal{S}_1 \).

- In the \( i \)th \((i \geq 2)\) iteration, given \( \mathcal{A}_i \), the current set of variables that have been selected, rank the remaining variables according to \( \{\hat{\beta}_{lj}^{\text{aug}}, j \notin \mathcal{A}_i\} \), where \( \hat{\beta}_{lj}^{\text{aug}} \) is calculated according to (12). Then, take the highest ranked \( ([n/\log(n)] - |\mathcal{A}_i|) \) variables as set \( \mathcal{R}_i \), on which run the second stage of Siri to select variables \( \mathcal{S}_i \), and let \( \mathcal{A}_{i+1} = \mathcal{A}_i \cup \mathcal{S}_i \). Stop the procedure when \( \mathcal{S}_i = \emptyset \) or \( |\mathcal{A}_{i+1}| > n/\log(n) \), and output \( \mathcal{A}_{i+1} \) as selected variables.

### 2.6 Implementation issues

#### 2.6.1 Choice of slicing scheme

Suppose \( S = (S_1, S_2, \ldots, S_H) \) is the true slicing scheme for the sliced inverse regression model in Section 2, and \( Z \) denotes the slice membership of slice scheme \( S \), i.e. \( Z = h \) if \( Y \in S_h \). We say that a slicing scheme \( S' = (S'_1, S'_2, \ldots, S'_{H'}) \) is a refinement of \( S \), which is denoted by \( S' \prec S \), if, for any \( S'_{h'} \subset S' \), there is an \( S_h \subset S \) such that \( S'_{h'} \subset S_h \).

For any slicing scheme \( S' \) with membership denoted by \( Z' \), as \( n \to \infty \), the average of the
log-likelihood ratio statistics defined in Theorem 3 is given by

$$\bar{l}_{j|A,S} = \lim_{n \to \infty} \frac{\hat{l}_{j|A_1}}{n}$$

$$= \log \left[ \text{Var} (X_j) - \text{Cov} (X_j, X_A) \text{Var} (X_A)^{-1} \text{Cov} (X_j, X_A)^T \right] - \log \left[ \text{E} \left( \text{Var} (X_j | Z') \right) - \text{E} \left( \text{Cov} (X_j, X_A | Z') \right) \text{Var} (X_A)^{-1} \text{Cov} (X_j, X_A | Z')^T \right]$$

and the average of the augmented log-likelihood ratio statistics defined in Theorem 4 is given by

$$\bar{l}^{\text{aug}}_{j|A,S} = \lim_{n \to \infty} \frac{\hat{l}^{\text{aug}}_{j|A_1}}{n}$$

$$= \log \left[ \text{Var} (X_j) - \text{Cov} (X_j, X_A) \text{Var} (X_A)^{-1} \text{Cov} (X_j, X_A)^T \right] - \text{E} \left( \log \left[ \text{Var} (X_j | Z') - \text{Cov} (X_j, X_A | Z') \text{Var} (X_A)^{-1} \text{Cov} (X_j, X_A | Z')^T \right] \right)$$

For the true slicing scheme $S$ or a slicing scheme $S'$ that is a refinement of $S$, i.e. $S' \preceq S$, under model assumptions in Section 2.2, we have

$$\bar{l}_{j|A,S'} \preceq \bar{l}_{j|A,S} = \bar{l}_{j|A_1}$$

$$= \log \left[ \text{Var} (X_j) - \text{Cov} (X_j, X_A) \text{Var} (X_A)^{-1} \text{Cov} (X_j, X_A)^T \right] - \log \left( \text{Var} (X_j | X_A, Z) \right)$$

where $\text{Var}(X_j | X_A, Z)$ is a constant that does not depend on $X_A$ or $Z$. Similary, under the augmented model in Section 2.3,

$$\bar{l}^{\text{aug}}_{j|A,S'} \preceq \bar{l}^{\text{aug}}_{j|A,S} = \bar{l}^{\text{aug}}_{j|A_1}$$

$$= \log \left[ \text{Var} (X_j) - \text{Cov} (X_j, X_A) \text{Var} (X_A)^{-1} \text{Cov} (X_j, X_A)^T \right] - \text{E} \left( \log \left( \text{Var} (X_j | X_A, Z) \right) \right)$$

For a slicing scheme $S'$ that is “coarser” than the true slicing scheme $S$, i.e. $S \preceq S'$, we have the following theorem.

**Theorem 8.** Suppose $S'$ is a slicing scheme such that $S \preceq S'$, where $S$ is the true slicing scheme.

(a) Under model assumptions (4) and (3) for variables in group 1 and group 0,

$$\bar{l}_{j|A,S} \geq \bar{l}_{j|A,S'}$$

where the equality holds if $G_1 \subset A$.

(b) Under model assumptions (4), (11) and (3) for variables in group 1, group 2 and group 0,

$$\bar{l}^{\text{aug}}_{j|A,S} \geq \bar{l}^{\text{aug}}_{j|A,S'}$$

where the equality holds if $G_1 \cup G_2 \subset A$. 

16
Theorem 8 suggests that the power of the two-stage stepwise procedure in selecting true predictors tends to increase if a slicing scheme uses 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 categorical). When a slicing scheme uses a larger number of slices, the number of observations in each slice will decrease, 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.

2.6.2 Choice of thresholds and the number of dimensions

Section 2.2, 2.3 and 2.4 characterize the asymptotic distributions and behaviors of the two-stage procedure under different model assumptions, and provide some theoretical guidelines for choosing the thresholds \((u_a, u_d)\) and \((v_a, v_d)\). In practice, however, these theoretical results should be used with much caution because of the following concerns. First, the distributions that we obtained in Section 2.2 and 2.3 are for a single addition or deletion step and under various assumptions. Second, the consistency results in Section 2.4 are valid in asymptotic sense and the rate of increase in dimension relative to sample size is usually unknown. In this section, we propose to use a cross-validation (CV) procedure for selecting thresholds and the number of dimensions \(q\) in model (4).

Let \(\chi^2_{\alpha,k}\) be the \(100\alpha\)th quantile of \(\chi^2_k\) and \(\{\alpha_i(n,p)\}_{1 \leq i \leq m}\) be a prespecified grid on a subinterval in \((0,1)\) given \(p\) predictor variables and \(n\) observations. For each value of \(\alpha_i\) and the number of dimensions \(q\) with \(0 \leq q \leq M\) (note that \(q = 0\) means that we will skip the first stage), we will consider two pairs of thresholds in the first and the second stage, \((u_a = \chi^2_{\alpha_i,q}, u_d = \chi^2_{\alpha_i-0.05, q})\) and \((v_a = f\chi^2_{\alpha_i, (H-1)(d+1)}, v_d = f\chi^2_{\alpha_i-0.05, (H-1)(d+1)};\) where \(d = |A_{1,2}|\) is the current number of variables that have been selected and \(f = \frac{n}{n-H(d+1)}\) is a finite sample correction factor. We follow the general \(K\)-fold CV scheme to select the best \(\alpha_i\) and \(q\). We randomly divide the original data into \(K\) equal-sized subsets and then apply the Siri procedure to any \(K-1\) subsets to generate the estimation and variable selection results. The remaining subset of the data is used to test the model and to generate a performance measurement. The performance measurements are averaged and the result is used as the CV score. We choose the pair of \(\alpha_i\) and \(q\) that maximizes the CV score.

We will consider two performance measures: classification error (CE) and mean absolute error (AE). Suppose there are \(n\) training samples and \(m\) testing samples. The \(j\)th observation \((j = 1, \ldots, m)\) in the testing set has response \(y_j\) and slice membership \(z_j\) (boundaries of slices are fixed based on training samples). Let \(p_j(h) = P_{\hat{M}}(Z = h|X = x_j, I = \hat{I})\) denote the estimated probability that observation \(j\) is from slice \(h\), where \(\hat{M}\) denotes the maximum likelihood estimate of model parameters and \(\hat{I} = (\hat{I}_1, \ldots, \hat{I}_p)\) denotes the inferred group membership of variables based on training samples. The classification error is defined as

\[
CE = \frac{1}{m} \sum_{j=1}^{m} I[z_j \neq \text{argmax}_h p_j(h)],
\]
where \( I[A] \) is the indicator function of an event \( A \). Let

\[
\bar{y}_h = \frac{\sum_{i=1}^{n} I[z_i = h] y_i}{\sum_{i=1}^{n} I[z_i = h]}
\]

denote the average response of training samples in slice \( h \). The mean absolute error is defined as

\[
AE = \frac{1}{m} \sum_{j=1}^{m} \left| y_j - \sum_{h=1}^{H} p_j(h) \bar{y}_h \right|
\]

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 are not explored here. We will compare CE and AE as performance measures through simulation studies and call the corresponding methods Siri-A and Siri-C, respectively.

### 3 Simulation studies

In this section, we study the performance of the Siri-based variable selection procedures by simulations. In order to facilitate fair comparisons with other existing methods that are motivated from the “forward” model perspective, the example presented here are all generated under the “forward” model, which differs from the “inverse” model assumptions of Siri. The setting of the simulation can also used to demonstrate the performance of Siri when some of its model assumptions are violated, especially the normal distribution assumption on relevant predictor variables within each slice.

In the following examples, we implemented a fixed slicing scheme with 5 slices of equal size (i.e. \( H = 5 \)) for the Siri-based methods. 10-fold CV procedures with grid \( \{\alpha_i(n, p)\}_{1 \leq i \leq 5} = \{1 - p^{-1}, 1 - 0.5p^{-1}, 1 - 0.1p^{-1}, 1 - 0.05p^{-1}, 1 - 0.01p^{-1}\} \) were used for choosing threshold values and the number of dimensions \( q \), and the range for selecting \( q \) is from 0 to 2, where \( q = 0 \) means that we will skip the first stage in the two-stage procedure of Siri. For Siri-based methods, results from CV procedures based on two different performance measures (CE and AE) are compared.

The variable selection methods that we compare the Siri-based procedure with include the Lasso, ISIS-SCAD, COP and DC-SIS+hierNet, where DC-SIS (Li et al., 2012), a feature screening procedure via distance correlation, is combined with hierNet, a Lasso-like procedure with hierarchical interaction constraints, to detect interactions between predictor variables. The R packages glmnet, SIS, COP and hierNet are used to run the Lasso, SIS-SCAD, COP and hierNet, respectively. We use the code that was provided by the original authors to run DC-SIS. For the 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 that are involved in ISIS-SCAD
is also selected by CV. For DC-SIS+hierNet, we first used DC-SIS procedure to reduce the dimensionality from $p$ to $[n / \log(n)]$ before applying hierNet. Both COP and Siri involve slicing the range of the response variable and choosing thresholds and number of dimensions, for which we use the same scheme described above except that the range of selecting $q$ is from 1 to 4 for COP.

3.1 Index models

For the first simulation, we compare the performance of different methods on index models. The predictor variables $X = (X_1, X_2, \ldots, X_p)^T$ are generated from a $p$-variate normal distribution with mean 0 and covariances $\text{Cov}(X_i, X_j) = \rho^{|i-j|}$ for $1 \leq i, j \leq p$. We will consider the following three scenarios:

Scenario 1.1 : $Y = \beta^T X + \sigma \epsilon$, \hspace{1cm} $n = 200, \sigma = 1.0, \rho = 0.5$, \hspace{1cm} $\beta = (3, 1.5, 2, 2, 2, 2, 2, 0, \ldots, 0)$,

Scenario 1.2 : $Y = \frac{\sum_{j=1}^{3} X_j}{0.5 + (1.5 + \sum_{j=2}^{3} X_j)^2} + \sigma \epsilon$, \hspace{1cm} $n = 200, \sigma = 0.2, \rho = 0.0$,

Scenario 1.3 : $Y = \frac{\sigma \epsilon}{1.5 + \sum_{j=1}^{8} X_j}$, \hspace{1cm} $n = 1000, \sigma = 0.2, \rho = 0.0$,

where the number of predictors $p = 1000$ and $\epsilon$ is independent of $X$ and follows $N(0,1)$. Scenario 1.1 is a linear model which involves 8 true predictors and 992 irrelevant predictors. Scenario 1.2 is a multi-index model with 4 true predictors, which was used in Li (1991) and Zhong et al. (2012). The relationship between $Y$ and two projections $X_1 + X_2 + X_3$ and $X_2 + X_3 + X_4$ are non-linear and $Y$ and $X$ are independent given two projections. Scenario 1.3 is a single-index heteroscedastic rational model with 8 true predictors.

For each scenario, we randomly generate 100 data sets each with $n$ observations and applied Siri(-A and -C), ISIS-Siri(-A and -C), COP, the Lasso, ISIS-SCAD and DC-SIS+hierNet to each data set. Two quantities, the average number of irrelevant predictors falsely selected as true predictors (which is denoted by FP) and the average number of true predictors falsely excluded as irrelevant predictors (which is denoted 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 990 and from 0 to 10 respectively, with small values indicating good performance in variable selection. The FP- and FN-values of the methods tested together with their corresponding standard errors (in brackets) are reported in Table 1.

From Table 1, under Scenario 1.1, ISIS-Siri-A and Siri-A (CV using mean absolute errors), have the lowest FP-values with FP=0.01 and FP=0.14, respectively, and ISIS-Siri-C (CV using classification errors) has the third lowest FP-values with FP=0.26. The other methods tend to have more false positive results. In terms of FNs, methods dervied from linear models all have FN=0, while COP and Siri-based methods have FN-values around 0.08. This is because Sir-based methods are developed for variable selection under models that are more general than the linear model.
Under Scenario 1.2, Lasso achieves the lowest FN-value (0.08), but it almost always missed the true predictor \( X_4 \) in the non-linear relationship, which are the same for the other methods developed for linear models. ISIS-Siri-A has the lowest FP-value (0.07) and the second lowest FN-value (0.13), while Siri-A and ISIS-Siri-C have slightly higher FPs and comparable FNs.

For the heteroscedatic rational model in Scenario 1.3, all the methods based on linear models fail to detect most of the true predictors. Siri-A and ISIS-Siri-A achieve the lowest FP-value (0.43), but have relatively high FP-values (4.60 and 4.82, respectively). On the other hand, Siri-C and ISIS-Siri-C have have the lowest FN-values, (0.42 and 0.51, respectively) with reasonably low FP-values around 2.00. The performance of COP is between two extremes with a FP-value of 1.26 and a FN-value of 3.32. Because the rational model in Scenario 1.3 contains a discontinuous point at \( \sum_{j=1}^{8} X_j = -1.5 \), the classification error used in Siri-C is a more robust measure than the absolute error used in Siri-A. As a result Siri-A and ISIS-Siri-A are more sensitive to the inclusion of irrelevant predictors and thus more conservative in selecting predictors into the model, which explains the better performance of Siri-C and ISIS-Siri-C in picking up true predictors.

Table 1: False positive (FP) and false negative (FN) values of different variable selection methods under Scenario 1.1, 1.2 and 1.3 in Section 3.1.

<table>
<thead>
<tr>
<th>Method</th>
<th>Scenario 1.1</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FP (0.992)</td>
<td>FN (0.8)</td>
<td>FP (0.996)</td>
<td>FN (0.4)</td>
<td>FP (0.992)</td>
</tr>
<tr>
<td>Lasso</td>
<td>0.59 (0.10)</td>
<td>0.00 (0.00)</td>
<td>0.08 (0.03)</td>
<td>1.07 (0.03)</td>
<td>0.00 (0.00)</td>
</tr>
<tr>
<td>ISIS-SCAD</td>
<td>0.35 (0.07)</td>
<td>0.00 (0.00)</td>
<td>0.60 (0.08)</td>
<td>1.02 (0.01)</td>
<td>5.08 (0.65)</td>
</tr>
<tr>
<td>DC-SIS+hierNet</td>
<td>0.59 (0.10)</td>
<td>0.00 (0.00)</td>
<td>8.65 (0.36)</td>
<td>0.93 (0.03)</td>
<td>7.66 (0.48)</td>
</tr>
<tr>
<td>COP</td>
<td>0.69 (0.12)</td>
<td>0.06 (0.03)</td>
<td>1.84 (0.16)</td>
<td>0.98 (0.01)</td>
<td>1.26 (0.13)</td>
</tr>
<tr>
<td>Siri-A</td>
<td>0.14 (0.05)</td>
<td>0.07 (0.03)</td>
<td>0.24 (0.08)</td>
<td>0.07 (0.03)</td>
<td>0.43 (0.08)</td>
</tr>
<tr>
<td>Siri-C</td>
<td>1.02 (0.16)</td>
<td>0.08 (0.03)</td>
<td>1.03 (0.15)</td>
<td>0.07 (0.03)</td>
<td>2.00 (0.16)</td>
</tr>
<tr>
<td>ISIS-Siri-A</td>
<td>0.01 (0.01)</td>
<td>0.09 (0.04)</td>
<td>0.13 (0.04)</td>
<td>0.07 (0.03)</td>
<td>0.43 (0.08)</td>
</tr>
<tr>
<td>ISIS-Siri-C</td>
<td>0.26 (0.05)</td>
<td>0.08 (0.03)</td>
<td>0.55 (0.08)</td>
<td>0.09 (0.03)</td>
<td>2.02 (0.17)</td>
</tr>
</tbody>
</table>
3.2 Models with interactions

Next, we consider scenarios with models containing interactions:

Scenario 2.1: \( Y = X_1 X_2 + \sigma \epsilon, \quad n = 200, \)
Scenario 2.2: \( Y = X_1 + X_1 X_2 + X_1 X_3 + \sigma \epsilon, \quad n = 200, \)
Scenario 2.3: \( Y = X_1(X_2 + X_3) + \sigma \epsilon, \quad n = 200, \)
Scenario 2.4: \( Y = X_1 X_2 X_3 + \sigma \epsilon, \quad n = 200, 500 \) and \( 1000, \)
Scenario 2.5: \( Y = X_1^2 X_2 + \sigma \epsilon, \quad n = 200, \)
Scenario 2.6: \( Y = \frac{X_1}{X_2 + X_3} + \sigma \epsilon, \quad n = 200, \)

where \( X_1, X_2, \ldots, X_p \) \((p = 1000)\) are independent identically distributed \( N(0, 1) \) random variables, \( \sigma = 0.2 \) and \( \epsilon \) is \( N(0, 1) \) and independent of \( X \). Scenario 2.1 and Scenario 2.3 contain only non-hierarchical two-way interaction terms, while Scenario 2.2 is a linear model with hierarchical interactions. We simulate the three-way interaction model Scenario 2.4 in three settings with different sample sizes: \( n = 200, n = 500 \) and \( n = 1000 \). Scenario 2.5 has an quadratic interaction term and Scenario 2.6 is an example of a rational model with interactions.

To test the performance of ISIS-SCAD with \( k \)-way interaction terms, we expand the set of predictor variables to include all the interaction terms up to \( k \)-way interactions among original predictors in the independence screening step. Then, we apply SCAD to top \([n/ \log(n)]\) expanded predictors and this method is called ISIS-SCAD\(_k\). Since DC-SIS has the ability to screen individual predictors even when the interactions between predictions are presented, we can use DC-SIS to reduce the number of predictors from \( p \) to \([n/ \log(n)]\) and then apply ISIS-SCAD to consider all the interaction terms up to \( k \)-way interactions among the selected \([n/ \log(n)]\) predictors. We call the corresponding methods DC-SIS+SCAD\(_k\). 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 true predictors. Because methods such as the Lasso and COP that are not specifically designed for detecting interactions are clearly at a disadvantage, we compare the performance of the Siri-based methods with DC-SIS+herNet, SIS+SCAD\(_2\), DC-SIS+SCAD\(_2\) and DC-SIS+herNet only. The FP- and FN-values (and their standard errors) of the methods tested under Scenario 2.1-2.3, Scenario 2.4-2.5 and Scenario 2.6 with different sample sizes are given in Table 2, Table 3 and Table 4, respectively. Note that FP- and FN-value 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 \( X_1, X_2 \) and \( X_1 X_2 \) all have non-zero coefficients in the regression model estimated by ISIS-SCAD\(_2\), we count the total number of predictors selected as 2, not 3.

As shown in Table 2, under Scenario 2.1-2.3, ISIS-SCAD\(_2\) correctly discovers most variables in the two-way interactions and does not pick up any irrelevant predictor. However, since ISIS-SCAD\(_2\) and herNet consider all the pairwise interactions between \( p \) predictor variables, they have computational complexity \( O(np^3) \) with \( p = 1000 \) and need much more
computational time compared with other methods (on average ISIS-SCAD<sub>2</sub> is about fifty
times slower than Siri-based methods). We can dramatically increase the computational
speed by using DC-SIS to screen variables before applying more refined variable selection
methods. The corresponding method DC-SIS+SCAD<sub>2</sub> and DC-SIS+tierNet as described
above have similar computational cost as Siri-based methods. On the other hand, true pre-
dictors may be incorrectly filtered by DC-SIS procedure, and thus methods with variable
screening by DC-SIS may have larger FN-values compared with the original methods with-
out screening, as shown by the underperformances of DC-SIS+SCAD<sub>2</sub> and DC-SIS+tierNet
under Scenario 2.3. The methods based on hierNet, which detects interaction terms under
hierarchical constraints, tend to have more FNs than the other methods, especially under
Scenarios with non-hierarchical interactions. Among methods with similar computational
costs, ISIS-Siri-A achieves the best performances in terms of both FNs and FPs for all three
scenarios in Table 2. The variable selection performances of ISIS-Siri-A are even comparable
with ISIS-SCAD<sub>2</sub> under Scenario 2.1 and 2.2, with much less computational cost.

Under Scenario 2.4 with three-way interactions, the computational cost prevents us to
directly apply ISIS-SCAD<sub>3</sub> or hierNet to consider all the three-way interaction terms and
we only compare the performance of ISIS-SCAD<sub>3</sub> and hierNet after variable screening by
DC-SIS. From Table 3, unsurprisingly, DC-SIS+SCAD<sub>3</sub> has the best performances under
different sample sizes since the data are generated exactly according to its model assumption.
However, as we increases the sample sizes, the performances of Siri-based methods, especially
ISIS-Siri-A, improve significantly compared with DC-SIS+tierNet. With sample size n = 1000,
ISIS-Siri-A is able select all the true predictors with very low FP-value.

Simulations in Scenario 2.1-2.4 are generated under the same model assumption as in
ISIS-SCAD<sub>2</sub> or ISIS-SCAD<sub>3</sub>, which gives them some advantage in these comparisons. Under
Scenario 2.5 and 2.6 in Table 4, when data are generated under different model assumptions,
Siri based methods, especially ISIS-Siri-A, significantly outperforms the other in detecting
true predictors with small FNs and reasonable FPs.

### 3.3 Variable screening performance

We further compare the performance of variable screening procedures in high dimensional
simulation studies. As in Scenario 1.1, the predictor variables \( \mathbf{X} = (X_1, X_2, \ldots, X_p)^T \) are
generated from a \( p \)-variate normal distribution with mean 0 and covariances \( \text{Cov}(X_i, X_j) = \rho^{|i-j|} \) for \( 1 \leq i, j \leq p \). We generate the response from the following scenarios:

- **Scenario 3.1**: \( Y = X_2 - \rho X_1 + 0.2X_{100} + \sigma \epsilon \),
- **Scenario 3.2**: \( Y = X_1X_2 + \sigma \epsilon^{2|X_{100}|} \epsilon \),
- **Scenario 3.3**: \( Y = \frac{X_{100}}{X_1 + X_2} + \sigma \epsilon \),

where sample size \( n = 200 \), \( \sigma = 0.2 \) and \( \epsilon \) is \( N(0, 1) \) and independent of \( \mathbf{X} \). For each scenario,
we consider four different cases with \( p = 2000 \) or 5000 and \( \rho = 0.0 \) or 0.5. Scenario 3.1 is a
linear model with three additive effects. The way \( X_1 \) is introduced is to make it marginally
Table 2: False positive (FP) and false negative (FN) values of different variable selection methods under Scenario 2.1, 2.2 and 2.3 in Section 3.2.

<table>
<thead>
<tr>
<th>Method</th>
<th>Scenario 2.1</th>
<th>Scenario 2.2</th>
<th>Scenario 2.3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FP (0.998)</td>
<td>FN (0.2)</td>
<td>FN (0.3)</td>
</tr>
<tr>
<td>ISIS-SCAD₂</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
</tr>
<tr>
<td>DC-SIS+SCAD₂</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>0.25 (0.09)</td>
</tr>
<tr>
<td>DC-SIS+hierNet</td>
<td>3.16 (0.28)</td>
<td>0.00 (0.00)</td>
<td>4.51 (0.34)</td>
</tr>
<tr>
<td>hierNet</td>
<td>2.38 (0.33)</td>
<td>0.00 (0.00)</td>
<td>6.93 (0.56)</td>
</tr>
<tr>
<td>Siri-A</td>
<td>0.01 (0.01)</td>
<td>0.00 (0.00)</td>
<td>0.03 (0.02)</td>
</tr>
<tr>
<td>Siri-C</td>
<td>0.66 (0.12)</td>
<td>0.00 (0.00)</td>
<td>0.39 (0.12)</td>
</tr>
<tr>
<td>ISIS-Siri-A</td>
<td>0.01 (0.01)</td>
<td>0.00 (0.00)</td>
<td>0.02 (0.01)</td>
</tr>
<tr>
<td>ISIS-Siri-C</td>
<td>0.76 (0.13)</td>
<td>0.00 (0.00)</td>
<td>0.29 (0.06)</td>
</tr>
</tbody>
</table>

uncorrelated with the response $Y$ (note that when $\rho = 0.0$, $X_1$ is not a true predictor). We also add another variable $X_{100}$ that has negligible correlation with $X_1$ and $X_2$ and a very small correlation with the response $Y$. Scenario 3.2 contains an interaction term $X_1X_2$ and a heteroscedastic noise determined by $X_{100}$. Scenario 3.3 is an example of a rational model with interactions.

We compare the performance of SIS-Siri and its iterative extension ISIS-Siri (discussed in Section 2.5) with sure independence screening with distance correlation (DC-SIS) proposed by Li et al. (2012), correlation learning based sure independence screening (SIS) and iterative sure independence screening combined with SCAD (ISIS-SCAD) both proposed by Fan and Lv (2008). We evaluate the performance using the proportion that true predictors are selected among the first $\lceil n/\log(n) \rceil$ predictors ranked by the procedure, with large values indicating good performance in variable screening. The performances of different methods are compared in terms of both individual predictors and all the predictors in the model.

From Table 3.3, under the linear model in Scenario 3.1, we can see that SIS and DC-SIS has better power than SIS-Siri in detecting variables that are weakly correlated with the response ($X_{100}$ in the example). As shown in Case 2 and 4, iterative procedures, such as ISIS-SCAD and ISIS-Siri, are more effective in detecting variables that are marginally uncorrelated with the response ($X_1$ in the example) than non-iterative procedures. Under Scenario 3.2, SIS and ISIS-SCAD are not able to pick up predictors in the interaction term and often miss the predictor in the heteroscedastic term. When there are moderate correlations between two variables in the interaction term ($X_1$ and $X_2$) in Case 2 and 4, DC-SIS is able to pick up them about half of the time. However, when the two variables are uncorrelated in Case 1 and 3, DC-SIS fails to detect them most of the time. ISIS-Siri and SIS-Siri have better performances than DC-SIS in detecting interactions for both $\rho = 0.0$ and $\rho = 0.5$. Under Scenario 3.3, when the relationship between the response and true predictors is rational,
Table 3: False positive (FP) and false negative (FN) values of different variable selection methods under Scenario 2.4 in Section 3.2.

<table>
<thead>
<tr>
<th>Method</th>
<th>Scenario 2.4 ((n = 200))</th>
<th></th>
<th>Scenario 2.4 ((n = 500))</th>
<th></th>
<th>Scenario 2.4 ((n = 1000))</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FP ((0, 997))</td>
<td>FN ((0, 3))</td>
<td>FP ((0, 997))</td>
<td>FN ((0, 3))</td>
<td>FP ((0, 997))</td>
<td>FN ((0, 3))</td>
</tr>
<tr>
<td>DC-SIS+SCAD (_3)</td>
<td>0.45 (0.12)</td>
<td>0.85 (0.12)</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
</tr>
<tr>
<td>DC-SIS+ hierNet</td>
<td>7.22 (0.64)</td>
<td>2.41 (0.08)</td>
<td>7.73 (1.17)</td>
<td>2.38 (0.08)</td>
<td>4.25 (1.17)</td>
<td>2.62 (0.06)</td>
</tr>
<tr>
<td>Siri-A</td>
<td>1.27 (0.19)</td>
<td>2.27 (0.05)</td>
<td>0.26 (0.09)</td>
<td>0.74 (0.07)</td>
<td>0.23 (0.7)</td>
<td>0.00 (0.00)</td>
</tr>
<tr>
<td>Siri-C</td>
<td>3.00 (0.25)</td>
<td>2.42 (0.06)</td>
<td>1.92 (0.18)</td>
<td>0.54 (0.06)</td>
<td>2.00 (0.19)</td>
<td>0.00 (0.00)</td>
</tr>
<tr>
<td>ISIS-Siri-A</td>
<td>0.98 (0.12)</td>
<td>2.27 (0.06)</td>
<td>0.36 (0.09)</td>
<td>0.70 (0.07)</td>
<td>0.21 (0.06)</td>
<td>0.00 (0.00)</td>
</tr>
<tr>
<td>ISIS-Siri-C</td>
<td>1.98 (0.16)</td>
<td>2.27 (0.07)</td>
<td>1.96 (0.17)</td>
<td>0.46 (0.05)</td>
<td>2.03 (0.19)</td>
<td>0.00 (0.00)</td>
</tr>
</tbody>
</table>

ISIS-Siri significantly outperforms the other methods in detecting the true predictors. The performances of different methods are only slightly affected as we increase the dimension from \(p = 2000\) and \(p = 5000\).

Table 4: False positive (FP) and false negative (FN) values of different variable selection methods Scenario 2.5 and 2.6 in Section 3.2.

<table>
<thead>
<tr>
<th>Method</th>
<th>Scenario 2.5</th>
<th>Scenario 2.6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FP ((0, 998))</td>
<td>FN ((0, 2))</td>
</tr>
<tr>
<td>ISIS-SCAD (_2)</td>
<td>0.04 (0.02)</td>
<td>1.09 (0.04)</td>
</tr>
<tr>
<td>DC-SIS+SCAD (_2)</td>
<td>2.38 (0.18)</td>
<td>0.51 (0.05)</td>
</tr>
<tr>
<td>DC-SIS+ hierNet</td>
<td>2.42 (0.44)</td>
<td>0.88 (0.05)</td>
</tr>
<tr>
<td>Siri-A</td>
<td>0.16 (0.05)</td>
<td>0.00 (0.00)</td>
</tr>
<tr>
<td>Siri-C</td>
<td>1.37 (0.17)</td>
<td>0.01 (0.01)</td>
</tr>
<tr>
<td>ISIS-Siri-A</td>
<td>0.08 (0.03)</td>
<td>0.00 (0.00)</td>
</tr>
<tr>
<td>ISIS-Siri-C</td>
<td>0.88 (0.11)</td>
<td>0.01 (0.01)</td>
</tr>
</tbody>
</table>

4 Real data examples

4.1 Leukemia classification

For the first example, we applied Siri-C to select features for the classification of a leukemia data set from high density Affymetrix oligonucleotide 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.
Table 5: The proportion that true predictors are selected by different screening procedures under Scenario 3.1-3.3 in Section 3.3.

<table>
<thead>
<tr>
<th>Method</th>
<th>Scenario 3.1</th>
<th>Scenario 3.2</th>
<th>Scenario 3.3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$X_1$</td>
<td>$X_2$</td>
<td>$X_{100}$</td>
</tr>
<tr>
<td><strong>Case 1</strong>: $p = 2000, \rho = 0.0$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SIS</td>
<td>-</td>
<td>1.00</td>
<td>0.63</td>
</tr>
<tr>
<td>ISIS-SCAD</td>
<td>-</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>DC-SIS</td>
<td>-</td>
<td>1.00</td>
<td>0.55</td>
</tr>
<tr>
<td>SIS-Siri</td>
<td>-</td>
<td>1.00</td>
<td>0.29</td>
</tr>
<tr>
<td>ISIS-Siri</td>
<td>-</td>
<td>1.00</td>
<td>0.30</td>
</tr>
<tr>
<td><strong>Case 2</strong>: $p = 2000, \rho = 0.5$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SIS</td>
<td>0.03</td>
<td>1.00</td>
<td>0.80</td>
</tr>
<tr>
<td>ISIS-SCAD</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>DC-SIS</td>
<td>0.02</td>
<td>1.00</td>
<td>0.71</td>
</tr>
<tr>
<td>SIS-Siri</td>
<td>0.01</td>
<td>1.00</td>
<td>0.36</td>
</tr>
<tr>
<td>ISIS-Siri</td>
<td>1.00</td>
<td>1.00</td>
<td>0.45</td>
</tr>
<tr>
<td><strong>Case 3</strong>: $p = 5000, \rho = 0.0$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SIS</td>
<td>-</td>
<td>1.00</td>
<td>0.47</td>
</tr>
<tr>
<td>ISIS-SCAD</td>
<td>-</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>DC-SIS</td>
<td>-</td>
<td>1.00</td>
<td>0.39</td>
</tr>
<tr>
<td>SIS-Siri</td>
<td>-</td>
<td>1.00</td>
<td>0.14</td>
</tr>
<tr>
<td>ISIS-Siri</td>
<td>-</td>
<td>1.00</td>
<td>0.14</td>
</tr>
<tr>
<td><strong>Case 4</strong>: $p = 5000, \rho = 0.5$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SIS</td>
<td>0.03</td>
<td>1.00</td>
<td>0.80</td>
</tr>
<tr>
<td>ISIS-SCAD</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>DC-SIS</td>
<td>0.05</td>
<td>1.00</td>
<td>0.71</td>
</tr>
<tr>
<td>SIS-Siri</td>
<td>0.01</td>
<td>1.00</td>
<td>0.33</td>
</tr>
<tr>
<td>ISIS-Siri</td>
<td>1.00</td>
<td>1.00</td>
<td>0.39</td>
</tr>
</tbody>
</table>

(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 the Siri, SIS-SCAD-LD and nearest shrunken centroids method are shown in Table 6. The results of SIS-SCAD-LD and the nearest shrunken centroids method were extracted from Fan and Lv (2008) and Tibshirani et al. (2012), respectively. The Siri 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. Of all the methods, the Siri used the smallest number of genes (8 genes) to achieve a comparable
Table 6: Leukemia classification results

<table>
<thead>
<tr>
<th>Method</th>
<th>Training error</th>
<th>Test error</th>
<th>Number of genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Siri-C</td>
<td>0/38</td>
<td>1/34</td>
<td>8</td>
</tr>
<tr>
<td>SIS-SCAD-LD</td>
<td>0/38</td>
<td>1/34</td>
<td>16</td>
</tr>
<tr>
<td>Nearest shrunken centroid</td>
<td>1/38</td>
<td>2/34</td>
<td>21</td>
</tr>
</tbody>
</table>

classification accuracy.

4.2 Study of gene expression regulation using next generation sequencing data

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, gene 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 et al. (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 $12408 \times 312$ matrix with $(i,j)$th entry representing the score of the $j$th TF on the $i$th gene’s promoter region.

In Zhong et al. (2012), the 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-A 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-A and COP are shown in Table 7. Among the top-10 TFs ranked by the Siri, 8 of them are TFASs. The Siri was also able to identify Nanog and Sox that are generally believed to be the master ESC regulators but were missed in the results of the COP procedure. A further study of the top-ranked TF motifs supplemented from the TRANSFAC and picked up by the Siri could help us better understand transcriptional regulatory networks in embryonic stem cells.
Table 7: The ranks of 12 TFASs (among 312 predictors) using different methods

<table>
<thead>
<tr>
<th>TF names</th>
<th>Ranks</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Siri</td>
</tr>
<tr>
<td>E2f1</td>
<td>1</td>
</tr>
<tr>
<td>Zfx</td>
<td>3</td>
</tr>
<tr>
<td>Mycn</td>
<td>4</td>
</tr>
<tr>
<td>Klf4</td>
<td>5</td>
</tr>
<tr>
<td>Myc</td>
<td>6</td>
</tr>
<tr>
<td>Esrrb</td>
<td>8</td>
</tr>
<tr>
<td>Oct4</td>
<td>9</td>
</tr>
<tr>
<td>Tcfcp2l1</td>
<td>10</td>
</tr>
<tr>
<td>Nanog</td>
<td>14</td>
</tr>
<tr>
<td>Stat3</td>
<td>17</td>
</tr>
<tr>
<td>Sox2</td>
<td>18</td>
</tr>
<tr>
<td>Smad1</td>
<td>32</td>
</tr>
</tbody>
</table>

5 Conclusions

This paper studies the problem of variable selection in high dimensions from an “inverse” modeling perspective. The contributions of the proposed procedure that we call Siri is twofold. First, it is effective and computational efficient in detecting interactions. Combined with independence screening, Siri can be used to attack 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 or parametric models, when the performance of model-specific methods are usually compromised. As a trade-off, Siri imposes various assumptions on the distribution of the predictors. As demonstrated by our simulation studies, Siri has competitive performance under some settings when the generative model is different from the model assumptions that are used to derive the procedure. However, we found that Siri is not very robust against extreme outliers in the predictor variables, and preprocessing of the data, such as quantile normalization, can be used when extreme outliers are presented.

Like other stepwise procedures such as linear stepwise regression and the COP, Siri may encounter issues that are typical to stepwise variable selection methods as discussed in Miller (1984). An alternative to the stepwise approach is to use a Bayesian variable selection method and MCMC sampling by simultaneously modeling the response and predictor variables. The model proposed in this paper can be generalized to consider multiple response variables and Bayesian factor analysis can be used to model the conditional distribution of the responses and relevant predictors in each “slice”. Furthermore, under the Bayesian model, we can avoid arbitrary choices of slicing schemes by allowing slices to vary, and a dynamic programming
approach can be used to effectively sample the boundaries of slices given priors. The Bayesian approach is currently under investigation and results will be reported in future.

Acknowledgment We would like to 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.

A Proof of Theorems

A.1 Proof of Proposition 1

Proof of Proposition 1. Suppose there are two different sets of variables $G_{1,2}$ (with corresponding group 0 variables $G_0$) and $G'_{1,2}$ (with corresponding group 0 variables $G'_0$) both satisfying the minimum property. Let $A_1 = G_{1,2} \cap G'_{1,2}$, $A_2 = G_{1,2} \cap G'_0$ and $A_3 = G_0 \cap G'_{1,2}$. Then, we must have $A_2 \neq \emptyset$ and $A_3 \neq \emptyset$ because of minimum property. The conditional distribution of $A_2$ given $A_1$ and slice $h$ can be written as

$$X_{A_2 | X_{A_1}} = N \left( \alpha_{A_2 | A_1} + B_{A_2 | A_1} X_{A_1}, \Sigma_{12} = \Sigma_{A_2 | A_1} \right).$$

Let $\alpha^{(h)} = \alpha_{A_2 | A_1}$ and $B^{(h)} = B_{A_2 | A_1} = \left( B_{j}^{(h)} \right)_{j \in A_1}$. According to the minimum property, at least one of $\alpha^{(h)}$, $B_{j}^{(h)}$ ($j \in A_1$) and $\Sigma^{(h)}$ has to be different across slices. The conditional distribution of $X_{A_3 | X_{A_1}}$, given $X_{A_1 \cup A_2} = X_{g_{1,2}}$ is the same across slices,

$$X_{A_3 | X_{g_{1,2}}, Z} = N \left( a + b X_{A_1} + c X_{A_2}, \Sigma_3 = \Sigma_{g_{1,2}} \right).$$

The conditional distribution of $X_{A_3}$ given $X_{A_1}$ and slice $h$ is

$$X_{A_3 | X_{A_1}} = N \left( a + c \alpha^{(h)} + \left( b + c B^{(h)} \right) X_{A_1}, \Sigma_3 + c \Sigma_{2} c^T \right),$$

Since $A_3 \in G'_{1,2}$, according to the minimum property, at least one of $c \alpha^{(h)}$, $c B_{j}^{(h)}$ ($j \in A_1$) and $M^{(h)} = c \Sigma_{2} c^T$ has to be different across slices. Given $X_{A_1 \cup A_3} = X_{g_{1,2}}$ and slice $h$, the conditional distribution of $c X_{A_2}$ is

$$c X_{A_2 | X_{g_{1,2}}}, Z = N \left( \gamma^{(h)} + D^{(h)} X_{A_1} + C^{(h)} X_{A_3}, \Sigma^{(h)} \right),$$

where $C^{(h)} = M^{(h)} \left( \Sigma_3 + M^{(h)} \right)^{-1}$, $\gamma^{(h)} = -C^{(h)} a + \left( I_{|A_3|} - C^{(h)} \right) c \alpha^{(h)}$, $D^{(h)} = -C^{(h)} b + \left( I_{|A_3|} - C^{(h)} \right) c B^{(h)}$, and $\Sigma^{(h)} = M^{(h)} - M^{(h)} \left( \Sigma_3 + M^{(h)} \right)^{-1} M^{(h)}$.

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

$$\Sigma_3^{-\frac{1}{2}} \Sigma^{(h)} \Sigma_3^{-\frac{1}{2}} = N^{(h)} - N^{(h)} \left( I_{|A_3|} + N^{(h)} \right)^{-1} N^{(h)} = \Gamma^{(h)} A^{(h)} \left( I_{|A_3|} + A^{(h)} \right)^{-1} \left[ \Gamma^h \right]^{-1}.$$
Thus, $\Sigma_3^{-\frac{1}{2}}\Sigma(h)\Sigma_3^{-\frac{1}{2}}$ and $\Sigma(h)$ are different across slices.

Second, if $M^{(h)} = M$ ($h = 1, 2, \ldots, H$) are the same across slices. Then at least one of $cA(h)$ and $cB_j^{(h)}$ ($j \in A_1$) has to be different across slices. Without loss of generality, assume $cA(h)$ are different across slices, that is, trace $\left( \text{Var} \left( c\alpha^Z \right) \right) > 0$, where $\alpha^Z = \alpha^{(h)}$ when $Z = h$. Under Condition 1, $\lambda_{\max}(N) \leq \tau_{\max}^{(h)}$, $\lambda_{\min}(\Sigma_3) \geq \tau_{\min}$ and $\lambda_{\min}(\Sigma_3^{-1}) \geq \frac{1}{\tau_{\max}}$.

Then, $C^{(h)} = C = I_{A_3} - M (\Sigma_3 + M)^{-1} = \Sigma_3^{-\frac{1}{2}} \left( I_{A_3} - N (I_{A_3} + N)^{-1} \right) \Sigma_3^{-\frac{1}{2}}$, and

$$
\text{trace} \left( \text{Var} \left( \gamma^Z \right) \right) = \text{trace} \left( C \text{Var} \left( c\alpha^Z \right) C^T \right) \\
\geq \lambda_{\min}(\Sigma_3^{-1}) \lambda_{\min}(\Sigma_3) \lambda_2^{\min} \left( I_{A_3} - N (I_{A_3} + N)^{-1} \right) \text{trace} \left( \text{Var} \left( c\alpha^Z \right) \right) \\
\geq \left( \frac{\tau_{\max}}{\tau_{\max} + \tau_{\min}} \right)^2 \left( \frac{\tau_{\max}}{\tau_{\min}} \right) \text{trace} \left( \text{Var} \left( c\alpha^Z \right) \right) > 0
$$

Thus, $\gamma^{(h)}$ are different across slices.

Therefore, given $X_{G_{1,2}^t}$ and slice $h$, the conditional distribution of $cX_{A_3}$ depends on slice $h$, which is contradictory with the previous assumption that $A_2 \in G_0$. So we must have $G_{1,2} = G_{1,2}^t$.

### A.2 Proof of Theorem 3

**Proof of Theorem 3.** Since

$$
\log \left( P_{\tilde{\mathcal{M}}_1} (x_{A_1}) \right) - \log \left( P_{\tilde{\mathcal{M}}_1} (x_{A_1}) \right) = -\frac{n}{2} \log \left[ \det \left( 2\pi \Sigma_{A_1} \right) - \det \left( 2\pi \Sigma_{A_1} \right) \right] - \frac{n}{2},
$$

$$
\tilde{\sigma}_j = \frac{1}{n} x_j^T \left( I_n - \bar{x}_{A_1} \left( \bar{x}_{A_1}^T \bar{x}_{A_1} \right)^{-1} \bar{x}_{A_1}^T \right) x_j = \frac{\det \left( \hat{\Omega}_{A_1} \right)}{\det \left( \hat{\Omega}_{A_1} \right)},
$$

and

$$
\log \left( P_{\tilde{\mathcal{M}}_0} (x_j|x_{A_1}) \right) = -\frac{n}{2} \log (2\pi \tilde{\sigma}_j) - \frac{n}{2} = -\frac{n}{2} \left( \log \left[ \det \left( 2\pi \hat{\Omega}_{A_1} \right) \right] - \log \left[ \det \left( 2\pi \hat{\Omega}_{A_1} \right) \right] \right) - \frac{n}{2},
$$

$$
\hat{l}_{j|A_1} = 2 \log (\hat{L}_{j|A_1}) = -n \left( \log \left[ \det \left( \hat{\Omega}_{A_1}^{-1} \tilde{\Omega}_{A_1} \right) \right] - \log \left[ \det \left( \hat{\Omega}_{A_1}^{-1} \tilde{\Omega}_{A_1} \right) \right] \right).
$$

To prove (a), we just need to prove that

$$
\log \left[ \det \left( \hat{\Omega}_{A_1}^{-1} \tilde{\Sigma}_{A_1} \right) \right] = \sum_{i=1}^{q} \log \left( 1 - \hat{\lambda}_i^A \right),
$$

where $\hat{\lambda}_i^A$ is the $i$th largest eigenvalue of $\hat{\Omega}_{A_1}^{-1} \hat{B}_A$ corresponding to eigenvector $\eta_i$, $1 \leq i \leq |A|$. Since $\hat{B}_A \eta_i = \hat{\lambda}_i^A \hat{\Omega}_A \eta_i$ and $\hat{\Omega}_A = \hat{W}_A + \hat{B}_A$,

$$
\hat{\Omega}_A \eta_i = \frac{1}{\hat{\lambda}_i^A} \hat{B}_A \eta_i = \frac{1}{1 - \hat{\lambda}_i^A} \hat{W}_A \eta_i.
$$
Then, \[ \hat{W}_A \hat{B}_A^{-1/2} \hat{\Omega}_A \eta_i = \frac{1 - \hat{\lambda}_i^A}{\lambda_i^A} \hat{\Omega}_A \eta_i, \]
and
\[ (\hat{B}_A^{-1/2} \hat{W}_A \hat{B}_A^{-1/2}) \hat{B}_A^{-1/2} \hat{\Omega}_A \eta_i = \frac{1 - \hat{\lambda}_i^A}{\lambda_i^A} \hat{B}_A^{-1/2} \hat{\Omega}_A \eta_i, \]
Thus, the eigenvalues of \( \hat{B}_A^{-1/2} \hat{W}_A \hat{B}_A^{-1/2} \) are given by \( \theta_{|A|-i+1} = \frac{1 - \hat{\lambda}_i^A}{\lambda_i^A}, 1 \leq i \leq |A|. \) Let \( c_i \) be an eigenvector of \( \hat{B}_A^{-1/2} \hat{W}_A \hat{B}_A^{-1/2} = C \Theta C^T \) corresponding to the eigenvalue \( \theta_i \). We will prove that the eigenvalues of \( \hat{\Sigma}_A^{-1} \hat{B}_A \) are
\[ \omega_i = \begin{cases} 1/\theta_{|A|-i+1} = \frac{\hat{\lambda}_i^A}{1 - \hat{\lambda}_i^A} & \text{if } 1 \leq i \leq q; \\ 1/1+\theta_{|A|-i+1} = \hat{\lambda}_i^A & \text{if } q+1 \leq i \leq |A|. \end{cases} \]
with corresponding eigenvectors given by \( b_i = \hat{B}_A^{-1/2} c_{|A|-i+1} \). Therefore,
\[ \log \left[ \det \left( \hat{\Sigma}_A^{-1} \hat{B}_A \right) \right] = \log \left[ \det \left( \hat{\Omega}_A^{-1} \hat{B}_A \right) \right] - \log \left[ \det \left( \hat{\Sigma}_A^{-1} \hat{B}_A \right) \right] = \sum_{i=1}^{q} \log \left( 1 - \hat{\lambda}_i^A \right). \]
According to (6) in Proposition 2,
\[ \hat{B}_A^{-1/2} \hat{\Sigma}_A \hat{B}_A^{-1/2} c_{|A|-i+1} = \hat{B}_A^{-1/2} \left( \hat{W}_A + \hat{B}_A^{-1/2} C_{|A|-q} C_{|A|-q}^T \hat{B}_A^{-1/2} \right) \hat{B}_A^{-1/2} c_{|A|-i+1} \]
\[ = (\hat{B}_A^{-1/2} \hat{W}_A \hat{B}_A^{-1/2} \hat{\Sigma}_A \hat{B}_A^{-1/2} \hat{B}_A^{-1/2} c_{|A|-i+1} = 0 \text{ for } 1 \leq i \leq q \text{ and } 1 \text{ for } q \leq i \leq |A|. \]
Thus,
\[ \hat{B}_A^{-1/2} \hat{\Sigma}_A \hat{B}_A^{-1/2} c_{|A|-i+1} = \frac{1}{\omega_i} c_{|A|-i+1}, \]
and
\[ \hat{\Sigma}_A^{-1} \hat{B}_A b_i = \omega_i b_i, 1 \leq i \leq |A|. \]
To prove (b) and (c), note that
\[ \hat{\ell}_{j|A_1} = -n \left( \sum_{i=1}^{q} \log \left( 1 - \hat{\lambda}_i^{A_1+} \right) - \sum_{i=1}^{q} \log \left( 1 - \hat{\lambda}_i^{A_1} \right) \right) = n \sum_{i=1}^{q} \log \left( 1 + \frac{\hat{\lambda}_i^{A_1+} - \hat{\lambda}_i^{A_1}}{1 - \hat{\lambda}_i^{A_1+}} \right). \]
Given that \( G_1 \subset A_1 \) and \( A_1+ = A_1 \cup \{j\} \) and the conditions in (b), Zhong et al. (2012) have showed that
\[ \frac{n \left( \hat{\lambda}_i^{A_1+} - \hat{\lambda}_i^{A_1} \right)}{1 - \hat{\lambda}_i^{A_1+}}, i = 1, 2, \ldots, q, \]
are asymptotically independent and identically distributed as $\chi_i^2$. Thus,

$$\frac{\lambda_i^{A_{1+}} - \hat{\lambda}_i^{A_{1+}}}{1 - \lambda_i^{A_{1+}}} \xrightarrow{p} 0, \ i = 1, \ldots, q,$$

and

$$\hat{l}_{j|A_1} = n \sum_{i=1}^{q} \log \left( 1 + \frac{\lambda_i^{A_{1+}} - \hat{\lambda}_i^{A_{1+}}}{1 - \lambda_i^{A_{1+}}} \right) \approx \sum_{i=1}^{q} \frac{n(\lambda_i^{A_{1+}} - \hat{\lambda}_i^{A_{1+}})}{1 - \lambda_i^{A_{1+}}}.$$

Given $A_1$, the variables in $A_i^c$ follows a multivariate normal distribution with the same mean and variance across slices. Then, the proofs of (b) and (c) directly follows from the Theorem 1 and Theorem 2 in Zhong et al. (2012).

For (d), since $\lambda_i^A \xrightarrow{p} \lambda_i^A$, for $i = 1, 2, \ldots, p$,

$$\frac{\hat{l}_{j|A_1}}{n} \xrightarrow{p} - \sum_{i=1}^{q} \log \left( 1 - \lambda_i^{A_{1+}} \right) + \sum_{i=1}^{q} \log \left( 1 - \lambda_i^{A_1} \right)$$

$$= - \log \left( \det \left( \Omega_{A_{1+}}^{-1} W_{A_{1+}} \right) \right) + \log \left( \det \left( \Omega_{A_1}^{-1} W_{A_1} \right) \right)$$

$$= \log \left( \det \left( \Omega_{A_1} \right) \right) - \log \left( \det \left( \Omega_{A_{1+}} \right) \right) - \log \left( \det \left( W_{A_{1+}} \right) \right)$$

$$= \log \left( \frac{\text{Var}(X_j) - \text{Cov}(X_j, X_A) [\text{Var}(X_A)]^{-1} \text{Cov}(X_j, X_A)^T \right) - \log \left( \text{Var}(X_j | X_A, Z) \right)$$

$$= \log \left( 1 + \frac{\text{Var}(M_j) - \text{Cov}(M_j, X_A) [\text{Var}(X_A)]^{-1} \text{Cov}(M_j, X_A)^T \right)}{V_j}.$$

The last equality follows from $\text{Var}(X_j) = \text{Var}(M_j) + E(V_j) = \text{Var}(M_j) + V_j$ and $\text{Cov}(X_j, X_A) = \text{Cov}(M_j, X_A)$, where $M_j = E(X_j | X_A, Z)$, $V_j = \text{Var}(X_j | X_A, Z)$ and $V_j$ is a constant that does not depend on $X_A$ or $Z$. Note that

$$\text{Var}(M_j) \geq \text{Cov}(M_j, X_A) [\text{Var}(X_A)]^{-1} \text{Cov}(M_j, X_A)^T$$

where the equality holds if and only if $M_j = E(X_j | X_A, Z)$ is a linear combination of $X_A$ that does not depend on $Z$, that is, $M_j = E(X_j | X_A, Z) = E(X_j | X_A)$ under normality assumption.

### A.3 Proof of Theorem 4

**Proof of Theorem 4.** (a) Let $P_0 = \tilde{x}_{A} (\tilde{x}_{A}^T \tilde{x}_{A})^{-1} \tilde{x}_{A}^T$, $P_h = \text{diag} \left( 0, \ldots, \tilde{x}_{A}^{(h)} (\tilde{x}_{A}^{(h)})^T \tilde{x}_{A}^{(h)})^{-1} \tilde{x}_{A}^{(h)}^T, \ldots, 0 \right)$, and $R_0 = I_n - P_0$, $R_h = \text{diag}(0, \ldots, I_{n_h}, \ldots, 0) - P_h$. Then,

$$\left[ \tilde{\sigma}_j^{(h)} \right]^2 = \frac{1}{n_h} \tilde{x}_j^T R_h x_j = \frac{1}{n_h} \sigma_j^2 Q_h = \frac{1}{n} \frac{\sigma_j^2 Q_h}{s_h}, \text{ for } h = 1, \ldots, H.$$
and
\[ \hat{\sigma}_j^2 = \frac{1}{n} x_j^T R_0 x_j = \frac{1}{n} x_j^T \left( \sum_{h=1}^{H} R_h \right) x_j + \frac{1}{n} x_j^T \left( \sum_{h=1}^{H} P_h - P_0 \right) x_j = \frac{1}{n} \sum_{h=1}^{H} \sigma_j^2 Q_h + \frac{1}{n} \sigma_j^2 Q_0, \]

where \( \sigma_j^2 \) is the conditional variance of \( X_j \) given \( X_A \) for \( j \in A^c \).

The likelihood ratio test statistic can be written as
\[
\hat{l}_{\text{aug}}^j_{|A} = -n \left( \sum_{h=1}^{H} s_h \log \left( \hat{\sigma}_j^{(h)} \right)^2 - \log \hat{\sigma}_j^2 \right)
\]
\[
= n \left( \log \left( \frac{1}{n} \sum_{h=1}^{H} \sigma_j^2 Q_h + \frac{1}{n} \sigma_j^2 Q_0 \right) - \sum_{h=1}^{H} s_h \log \left( \frac{1 \sigma_j^2 Q_h}{n s_h} \right) \right)
\]
\[
= n \left( \log \left( 1 + \frac{Q_0}{\sum_{h=1}^{H} Q_h} \right) - \sum_{h=1}^{H} s_h \log \left( \frac{Q_h/s_h}{\sum_{h=1}^{H} Q_h} \right) \right).
\]

Note that both \( \left( \sum_{h=1}^{H} P_h - P_0 \right) \) and \( R_h \)'s are orthogonal to \( \tilde{x}_A \). Given that \( j \in A^c \), according to Cochran’s theorem, \( Q_0 = x_j^T (\sum_{h=1}^{H} P_h - P_0) x_j \) and \( Q_h = x_j^T R_h x_j / \sigma_j^2 \) are independent, and
\[ Q_0 \sim \chi^2_{(H-1)d}, \quad \text{and} \quad Q_h \sim \chi^2_{n_h - d}, \quad \text{for} \quad h = 1, \ldots, H, \quad \text{and} \quad d = |A| + 1. \]

(b) Let \( n_h' = n_h - d \) and \( n' = n - H d \). 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
\[ \frac{Q_0}{n'} \to 1, \quad \frac{\sum_{h=1}^{H} Q_h}{n'} \to 1, \quad \text{and} \quad \frac{Q_0}{\sum_{h=1}^{H} Q_h} \to 0, \]
we have
\[ U_j \equiv n \log \left( 1 + \frac{Q_0}{\sum_{h=1}^{H} Q_h} \right) \to \frac{Q_0}{\sum_{h=1}^{H} Q_h} \to Q_0 \sim \chi^2_{(H-1)d}. \]

Since
\[ \frac{Q_h/s_h}{\sum_{h=1}^{H} Q_h} = \frac{n_h'/n_h}{n'/n} \frac{Q_h/n'}{Q_h/n} \to 1, \quad \text{for} \quad h = 1, \ldots, H, \]
and
\[ \frac{\sum_{h=1}^{H} Q_h}{n} = \frac{n'}{n} \frac{\sum_{h=1}^{H} Q_h}{n'} \to 1, \]

32
we have

\[
\tilde{U}_j \equiv -n \sum_{h=1}^{H} s_h \log \left( \frac{Q_h / s_h}{\sum_{h=1}^{H} Q_h} \right) = -n \sum_{h=1}^{H} s_h \log \left( 1 + \frac{Q_h / s_h}{\sum_{h=1}^{H} Q_h} - 1 \right)
\]

\[
\simeq -n \sum_{h=1}^{H} s_h \left( \frac{Q_h / s_h}{\sum_{h=1}^{H} Q_h} - 1 \right) + \frac{1}{2} \sum_{h=1}^{H} n_h \left( \frac{Q_h / s_h}{\sum_{h=1}^{H} Q_h} - 1 \right)^2
\]

\[
= \frac{1}{2} \sum_{h=1}^{H} n_h \left( \frac{Q_h / n_h - \sum_{h=1}^{H} Q_h / n}{\sum_{h=1}^{H} Q_h / n} \right)^2 = \frac{1}{2} \sum_{h=1}^{H} n_h \left( \frac{Q_h / n_h - \sum_{h=1}^{H} Q_h / n}{\sum_{h=1}^{H} Q_h / n} \right)^2
\]

\[
\simeq \frac{1}{2} \sum_{h=1}^{H} n_h \left( \frac{Q_h / n_h}{\sqrt{2n_h}} - \frac{\sum_{h=1}^{H} \sqrt{n_h} Q_h}{\sqrt{2n_h}} \right)^2.
\]

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

\[
q_h = \sqrt{\frac{\sum_{i=1}^{n_h} z_{hij}^2}{n_h \frac{\sum_{i=1}^{n_h} z_{hij}^2}{\sqrt{2n_h}}}},
\]

where \( z_{hij} \) are independent and \( z_{hij} \sim N(0,1) \) for \( i = 1, \ldots, n_h' \) and \( h = 1, \ldots, H \). Thus, according to central limit theorem, \( q_h \overset{D}{\rightarrow} N(0,1) \), for \( h = 1, \ldots, H \), and let \( q = (q_1, \ldots, q_H)^T \). Then,

\[
\tilde{U}_j \simeq \sum_{h=1}^{H} q_h^2 - \left( \sum_{h=1}^{H} \sqrt{n_h} q_h \right)^2 = q^T \left( I_H - JJ^T \right) q
\]

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

\[
\tilde{U}_j^\text{aug} = U_j + \tilde{U}_j \overset{D}{\rightarrow} \chi^2_H(1)(d+1).
\]

To prove (c), note that

\[
U_j \simeq Q_0^{(j)} = \frac{x_j^T (\sum_{h=1}^{H-1} P_h - P_0) x_j}{\sigma_j^2} = \sum_{i=1}^{(H-1)d} z_{ij}^2,
\]

where for \( i = 1, \ldots, (H-1)d, z_{ij} \sim N(0,1) \) are independent, and for \( j' \neq j \), \( \text{Cov}(z_{ij}, z_{ij'}) = \Psi_{jj'} \). For \( h \in \{1, \ldots, H\} \),

\[
Q_h^{(j)} = \frac{x_j^T R_h x_j}{\sigma_j^2} = \sum_{i=1}^{n_h'} z_{hij}^2,
\]

where for \( i = 1, \ldots, n_h', z_{hij} \sim N(0,1) \) are independent, and for \( j' \neq j \), \( \text{Cov}(z_{hij}, z_{hij'}) = \Psi_{jj'} \). Since \( \text{Cov}(z_{hij}^2, z_{hij'}^2) = 2\Psi_{jj'}^2 \), for \( h = 1, \ldots, H \) and \( j \neq j' \), we have

\[
\text{Cov} \left( q_h^{(j)}, q_h^{(j')} \right) = \frac{1}{2n_h} \text{Cov} \left( Q_h^{(j)}, Q_h^{(j')} \right) = \frac{n_h'}{n_h} \Psi_{jj'}^2 \rightarrow \Psi_{jj'}^2.
\]
Hence

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

where for \( i = 1, \ldots, (H-1) \), \( \tilde{z}_{ij} \sim N(0, 1) \) are independent, and for \( j' \neq j \), \( \text{Cov}(\tilde{z}_{ij}, \tilde{z}_{ij'}) = \Psi_{jj'} \). Therefore,

\[ \left( \hat{\sigma}_{j|A}^{\text{aug}} \right)_{j \in A^{\text{c}}} = \left( U_j + \tilde{U}_j \right)_{j \in A^{\text{c}}} \overset{D}{\longrightarrow} \left( \left( \sum_{i=1}^{(H-1)d} \tilde{z}_{ij}^2 + \sum_{i=1}^{(H-1)} \tilde{z}_{ij}^2 \right) \right)_{j \in A^{\text{c}}}. \]

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

\[ \left[ \hat{\sigma}_j^{(h)} \right]^2 \overset{p}{\longrightarrow} \left[ \sigma_j^{(h)} \right]^2 \]

\[ = \text{Var} (X_j | Z = h) - \text{Cov} (X_j, X_A | Z = h) \left[ \text{Var} (X_A | Z = h) \right]^{-1} \text{Cov} (X_j, X_A | Z = h)^T \]

\[ = \text{Var} (X_j | X_A, Z = h), \]

and as \( n \to \infty \),

\[ \hat{\sigma}_j^2 \overset{p}{\longrightarrow} \sigma_j^2 \]

\[ = \text{Var} (X_j) - \text{Cov} (X_j, X_A) \left[ \text{Var} (X_A) \right]^{-1} \text{Cov} (X_j, X_A)^T \]

\[ = \text{Var} (X_j) - \text{Cov} (E(X_j | X_A, Z), X_A) \left[ \text{Var} (X_A) \right]^{-1} \text{Cov} (E(X_j | X_A, Z), X_A)^T. \]

Let \( M_j = E(X_j | X_A, Z) \) and \( V_j = \text{Var} (X_j | X_A, Z) \). Then, \( \text{Var} (X_j) = \text{Var} (M_j) + E(V_j) \), and

\[ \frac{\hat{\sigma}_{j|A}^{\text{aug}}}{n} = \log \hat{\sigma}_j^2 - \sum_{h=1}^{H} s_h \log \left[ \hat{\sigma}_j^{(h)} \right]^2 \]

\[ \overset{p}{\longrightarrow} \log \sigma_j^2 - \sum_{h=1}^{H} s_h \log \left[ \sigma_j^{(h)} \right]^2 \]

\[ = \log \left( E(V_j) + \text{Var} (M_j) - \text{Cov} (M_j, X_A) \left[ \text{Var} (X_A) \right]^{-1} \text{Cov} (M_j, X_A)^T \right) - \sum_{h=1}^{H} s_h \log \left( \text{Var} (X_j | X_A, Z = h) \right). \]

Since \( \text{Var} (X_j | X_A, Z = h) \) is constant that does not depend on \( X_A \) under normality assumption,

\[ E \log (V_j) = \sum_{h=1}^{H} s_h \log \left( \text{Var} (X_j | X_A, Z = h) \right), \]

and thus

\[ \frac{\hat{\sigma}_{j|A}^{\text{aug}}}{n} \overset{p}{\longrightarrow} \log \left( E(V_j) + \text{Var} (M_j) - \text{Cov} (M_j, X_A) \left[ \text{Var} (X_A) \right]^{-1} \text{Cov} (M_j, X_A)^T \right) - E \log (V_j) \]

\[ = \log \left( 1 + \frac{\text{Var} (M_j) - \text{Cov} (M_j, X_A) \left[ \text{Var} (X_A) \right]^{-1} \text{Cov} (M_j, X_A)^T}{E(V_j)} \right) + \log (EV_j) - E \log (V_j). \]

34
Note that
\[ \text{Var}(M_j) \geq \text{Cov}(M_j, X_A) [\text{Var}(X_A)]^{-1} \text{Cov}(M_j, X_A)^T \]
where the equality holds if and only if \( M_j = E(X_j|X_A, Z) \) is a linear combination of \( X_A \) that does not depend on \( Z \), that is, \( M_j = E(X_j|X_A, Z) = E(X_j|X_A) \) under normality assumption. Furthermore, according to Jensen’s inequality
\[ \log(EV_j) \geq E \log(V_j), \]
where the equality holds if and only if \( V_j = \text{Var}(X_j|X_A, Z) = EV_j \), that is, \( \text{Var}(X_j|X_A, Z = h) \) are the same constant for \( h = 1, \ldots, H \). Since \( M_j = E(X_j|X_A) \), \( \text{Var}(X_j|X_A) = EV_j \),

\[ \text{A.4 Proof of Theorem 5} \]

To prove Theorem 5, we will need the following two lemmas.

**Lemma 1.** Under the same condition as Theorem 5, there exists \( \kappa' > 0 \) such that for any set of predictors \( A \) such that \( A^c \cap G_1 \neq \emptyset \),

\[ \max_{j \in A^c \cap G_1} \left[ \frac{\text{Var}(M_j) - \text{Cov}(M_j, X_A) [\text{Var}(X_A)]^{-1} \text{Cov}(M_j, X_A)^T}{V_j} \right] \geq \kappa' n^{-\xi_0}, \]

where \( M_j = E(X_j|X_A, Z) \) and \( V_j = \text{Var}(X_j|X_A, Z) \) is a constant that does not depend on \( X_A \) or \( Z \).

**Proof of Lemma 1.** Let \( B = A \cap G_1, C = A \cup G_1^c \), and \( D = A^c \cap G_1 \neq \emptyset \). Then under model (4),
\[ X_D|X_B, Z = h \sim N \left( \alpha_{D|B}^{(h)} + \beta_{D|B}^T X_B, \Sigma_1 = \Sigma_{D|B} \right). \]

Let \( \eta_D^{(h)} = (\eta_j^{(h)})^T \), where \( \eta_j^{(h)} \) is defined in Condition 2. Then, \( \eta_D^{(h)} = \Psi \alpha_{D|B}^{(h)} \), where \( \Psi \) is a \( |D| \) by \( |D| \) matrix that satisfies \( \Psi^T \Psi = \Sigma_1^{-1} \Delta_2 \Sigma_1^{-1} \), \( \Delta_2 = \text{diag}(\delta_j^2) \) are defined in Condition 2. Under Condition 1, \( \delta_j^2 \leq \tau_{\max} \) 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,
\[ \text{trace} \left( \text{Var} \left( \eta_D^{Z} \right) \right) = \text{trace} \left( \Psi \text{Var} \left( \alpha_{D|B}^{Z} \right) \Psi^T \right) \leq \left( \frac{\tau_{\max}}{\tau_{\min}} \right)^2 \text{trace} \left( \text{Var} \left( \alpha_{D|B}^{Z} \right) \right), \]

and
\[ \text{trace} \left( \text{Var} \left( \alpha_{D|B}^{Z} \right) \right) \geq |D| \left( \frac{\tau_{\min}}{\tau_{\max}} \right)^2 \kappa n^{-\xi_0}, \]

35
Let $S_\lambda \geq |D|$ be a constant that does not depend on slice $h$. Therefore, the conditional distribution of $X_D$ given $X_A = X_{B \cup C}$ can be written as

$$X_D | X_A, Z = h \sim N \left( \alpha_{(h)}^{(h)} + \beta_{(h)}^{(h)} X_A, \Sigma_{(h)}^{(h)} \right),$$

where $\alpha_{(h)}^{(h)} = \alpha_0 + M \alpha_{(h)}^{(h)}$, $M = \Sigma_1^{\frac{1}{2}} \left( I_{|D|} - N^T (I_{|C|} + NN^T)^{-1} N \right) \Sigma_1^{-\frac{1}{2}}$, $N = \Sigma_0^{-\frac{1}{2}} \Sigma_0 \Sigma_1^{-\frac{1}{2}}$ and $\alpha_0$ is a constant that does not depend on slice $h$. Since $\lambda_{\max} \left( NN^T \right) \leq \frac{\lambda_{\max}(\Sigma_0)}{\lambda_{\min}(\Sigma_0)} \leq \frac{\tau_{\max}}{\tau_{\min}}$, $\lambda_{\min}(\Sigma_1) \geq \tau_{\min}$ and $\lambda_{\min}(\Sigma_1^{-1}) \geq \frac{1}{\tau_{\max}}$,

$$\lambda_{\min} \left( I_{|D|} - N^T (I_{|C|} + NN^T)^{-1} N \right) \geq \frac{1}{1 + \frac{\tau_{\max}}{\tau_{\min}}} = \frac{\tau_{\min}}{\tau_{\max} + \tau_{\min}},$$

and

$$\text{trace} \left( \text{Var} \left( \alpha_{(h)}^{(h)} \right) \right) = \text{trace} \left( M \text{Var} \left( \alpha_{(h)}^{(h)} \right) M^T \right) \geq \lambda_{\min} \left( \Sigma_1^{-1} \right) \lambda_{\min} \left( \Sigma_1 \right) \lambda_{\min}^2 \left( I_{|D|} - N^T (I_{|C|} + NN^T)^{-1} N \right) \text{trace} \left( \text{Var} \left( \alpha_{(h)}^{(h)} \right) \right) \geq \left( \frac{\tau_{\min}}{\tau_{\max} + \tau_{\min}} \right)^2 \left( \frac{\tau_{\min}}{\tau_{\max}} \right) \text{trace} \left( \text{Var} \left( \alpha_{(h)}^{(h)} \right) \right) \geq |D| \left( \frac{\tau_{\min}}{\tau_{\max} + \tau_{\min}} \right)^2 \left( \frac{\tau_{\min}}{\tau_{\max}} \right)^3 \kappa n^{-\xi_0}.$$

Thus, there exists $j \in D = A^c \cap G_1$ such that

$$\text{Var} \left( \alpha_{j|h} \right) \geq \left( \frac{\tau_{\min}}{\tau_{\max} + \tau_{\min}} \right)^2 \left( \frac{\tau_{\min}}{\tau_{\max}} \right)^3 \kappa n^{-\xi_0}.$$

For such $j$, we have $M_j = \alpha_{j|h} + \beta_{j|h} X_A, V_j = \Sigma_{j|h} \leq \tau_{\max}$, and

$$\text{Var} \left( M_j \right) - \text{Cov} \left( M_j, X_A \right) \left[ \text{Var} \left( X_A \right) \right]^{-1} \text{Cov} \left( M_j, X_A \right)^T \text{Var} \left( M_j \right),$$

$$= \text{Var} \left( \alpha_{j|h} \right) - \text{Cov} \left( \alpha_{j|h}, E \left( X_A | Z \right) \right) \left[ \text{Var} \left( X_A \right) \right]^{-1} \text{Cov} \left( \alpha_{j|h}, E \left( X_A | Z \right) \right)^T.$$

Let $S_1 = \text{Var} \left( E \left( X_A | Z \right) \right)$ and $S_2 = \text{Var} \left( X_A | Z \right) = \text{Var} \left( X_A \right)$.

Since $\lambda_{\min} \left( S_2 \right) \geq \tau_{\min}$
and \( \lambda_{\max} (S_1) \leq \lambda_{\max} (\text{Var}(X_A)) \leq \tau_{\max}, \lambda_{\min} \left( S_1^{-\frac{1}{2}} S_2 S_1^{-\frac{1}{2}} \right) \geq \frac{\tau_{\min}}{\tau_{\max}} \) and

\[
\text{Cov} \left( \alpha_{i,j,A}^Z, E(X_A | Z) \right) [\text{Var}(X_A)]^{-1} \text{Cov} \left( \alpha_{i,j,A}^Z, E(X_A | Z) \right)^T \leq \frac{1}{1 + \lambda_{\min} \left( S_1^{-\frac{1}{2}} S_2 S_1^{-\frac{1}{2}} \right)} \text{Cov} \left( \alpha_{i,j,A}^Z, E(X_A | Z) \right) S_1^{-1} \text{Cov} \left( \alpha_{i,j,A}^Z, E(X_A | Z) \right)^T \leq \frac{\tau_{\max}}{\tau_{\min} + \tau_{\max}} \text{Var} \left( \alpha_{i,j,A}^Z \right).
\]

Therefore,

\[
\frac{\text{Var}(M_j) - \text{Cov}(M_j, X_A) [\text{Var}(X_A)]^{-1} \text{Cov}(M_j, X_A)^T}{V_j} \geq \frac{1}{\tau_{\max}} \frac{\tau_{\min}}{\tau_{\max} + \tau_{\min}} \text{Var} \left( \alpha_{i,j,A}^Z \right) \geq \frac{1}{\tau_{\max}} \left( \frac{\tau_{\min}}{\tau_{\max} + \tau_{\min}} \right)^3 \left( \frac{\tau_{\min}}{\tau_{\max}} \right)^3 \kappa n^{-\lambda_0}.
\]

\( \square \)

**Lemma 2.** Under the same condition as Theorem 5, for any set of predictors \( \mathcal{A} \), let \( \hat{\lambda}_i^A \) be the \( i \)th largest eigenvalue of \( \hat{\Omega}_A^{-1} \hat{B}_A \) and let \( \lambda_i^A \) be the \( i \)th largest eigenvalue of \( \Omega_A^{-1} B_A \). Then, for \( 0 < \epsilon < 1 \) and \( i = 1, 2, \ldots, q \), there exists positive constants \( C_1 \) and \( C_2 \) such that

\[
P \left( \max_{\mathcal{A} \subset \{1, \ldots, p\}} \left| \log \left( 1 - \hat{\lambda}_i^A \right) - \log \left( 1 - \lambda_i^A \right) \right| > \epsilon \right) \leq 2p(p + 1)C_1 \exp \left( -C_2n\frac{\tau_{\min}^4 \epsilon^2}{64\tau_{\max}^2 p^2} \right)
\]

**Proof of Lemma 2.** Since \( X_A \) 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 exists constant \( C_1 \) and \( C_2 \) such that

\[
P \left( \max_{\mathcal{A} \subset \{1, \ldots, p\}} \left| \tilde{\lambda}_i^A - \lambda_i^A \right| > \epsilon \right) \leq 2p(p + 1)C_1 \exp \left( -C_2n\frac{\tau_{\min}^2 \epsilon^2}{16p^2} \right)
\]

following similar arguments in the proof of Lemma 2 in Zhong et al. (2012). Since \( \Omega_A = W_A + B_A \), where \( \Omega_A = \text{Var}(X_A) \) and \( W_A = E(\text{Var}(X_A | Z)) = \text{Var}(X_A | Z) \) under model assumptions (4) and (3), \( \lambda_{\max}(\Omega_A) \leq \tau_{\max} \) and \( \lambda_{\min}(W_A) \geq \tau_{\min} \). Thus,

\[
\lambda_i^A = \max_{||\eta||=1} \frac{\eta^T B_A \eta}{||\eta||^2} = 1 - \min_{||\eta||=1} \frac{\eta^T W_A \eta}{||\eta||^2} = 1 - \min_{||\eta||=1} \frac{\eta^T W_A \eta}{\max_{||\eta||=1} \eta^T \Omega_A \eta} \leq 1 - \frac{\tau_{\min}}{\tau_{\max}}.
\]

37
and
\[ 1 - \lambda_i^A \geq 1 - \lambda_i \geq \frac{\tau_{\min}}{\tau_{\max}}, \text{ for } i = 1, 2, \ldots, q. \]

Therefore, for \( i = 1, 2, \ldots, q \) and \( 0 < \epsilon < 1 \), there exists constant \( C_1 \) and \( C_2 \) such that
\[
P \left( \max_{\mathcal{A} \subset \{1, \ldots, p\}} \left| \log (1 - \hat{\lambda}_i^A) - \log (1 - \lambda_i^A) \right| > \epsilon \right) \leq p \left( \max_{\mathcal{A} \subset \{1, \ldots, p\}} \frac{\lambda_i^A - \hat{\lambda}_i^A}{1 - \lambda_i^A} > \frac{\epsilon}{2} \right) \]
\[
\leq P \left( \max_{\mathcal{A} \subset \{1, \ldots, p\}} \left| \hat{\lambda}_i^A - \lambda_i^A \right| > \frac{\tau_{\min}^4}{2\tau_{\max}} \epsilon \right) \leq 2p(p + 1) C_1 \exp \left( -C_2 n \frac{\tau_{\min}^4 \epsilon^2}{64\tau_{\max}^2 p^2 q^2} \right).
\]

\[\square\]

**Proof of Theorem 5.** Let \( R_A = \sum_{i=1}^q \log \left(1 - \hat{\lambda}_i^A\right) - \sum_{i=1}^q \log \left(1 - \lambda_i^A\right) \).

Then, according to Lemma 2, for \( 0 < \epsilon < 1 \), there exists constant \( C_1 \) and \( C_2 \) such that
\[
P \left( \max_{\mathcal{A} \subset \{1, \ldots, p\}} |R_A| > \epsilon \right) \leq 2p(p + 1) q C_1 \exp \left( -C_2 n \frac{\tau_{\min}^4 \epsilon^2}{64\tau_{\max}^2 p^2 q^2} \right).
\]

Under Condition 3, \( p = o(n^p) \) and \( 2\rho_0 + 2\xi_0 < 1 \), and for any positive constant \( C_0 \),
\[
P \left( \max_{\mathcal{A} \subset \{1, \ldots, p\}} |R_A| > C_0 n^{-\xi_0} \right) \leq 2p(p + 1) q C_1 \exp \left( -C_2 n^{1-2\xi_0-2\rho_0} \frac{\tau_{\min}^4 C_0^2}{64\tau_{\max}^2 p^2 q^2} \right) \to 0
\]
as \( n \to \infty \).

From Theorem 3, for \( j \notin A_1 \) and \( A_{1+} = A_1 \cup \{j\} \)
\[
\frac{\hat{\ell}_j|_{A_1}}{n} = \frac{-\sum_{i=1}^q \log \left(1 - \hat{\lambda}_i^{A_{1+}}\right) + \sum_{i=1}^q \log \left(1 - \hat{\lambda}_i^{A_1}\right)}{n} = \frac{-\sum_{i=1}^q \log \left(1 - \lambda_i^{A_{1+}}\right) + \sum_{i=1}^q \log \left(1 - \lambda_i^{A_1}\right) - R_{A_{1+}} + R_{A_1}}{n} = \log \left(1 + \frac{\Var(M_j) - \Cov(M_j, X_{A_1}) [\Var(X_{A_1})]^{-1} \Cov(M_j, X_{A_1})^T}{V_j}\right) - R_{A_{1+}} + R_{A_1},
\]

where \( M_j = E (X_j | X_{A_1}, Z) \), \( V_j = \Var (X_j | X_{A_1}, Z) \), and \( V_j \) is a constant that does not depend on \( X_{A_1} \) or \( Z \) under model assumptions (4) and (3).

When \( A^c_1 \cap G_1 \neq \emptyset \), according to Lemma 1, there exists \( \kappa' > 0 \) such that
\[
\max_{j \in A^c_1 \cap G_1} T_{j|A_1} = \max_{j \in A^c_1 \cap G_1} \left[ \frac{\Var(M_j) - \Cov(M_j, X_{A_1}) [\Var(X_{A_1})]^{-1} \Cov(M_j, X_{A_1})^T}{V_j} \right] \geq \kappa' n^{-\xi_0},
\]

Then, for sufficiently large \( n \), there exists \( j \in A^c_1 \cap G_1 \) such that
\[
n \log \left(1 + T_{j|A_1}\right) \geq \frac{\kappa'}{2} n^{1-\xi_0},
\]

38
and

$$\hat{l}_{j|A_1} \geq n \log \left(1 + T_{j|A_1}\right) - n \left(|R_{A_1^+}| + |R_{A_1}|\right) \geq \frac{\kappa'}{2} n^{1-\xi_0} - n \left(|R_{A_1^+}| + |R_{A_1}|\right).$$

Since

$$P \left( \max_{A \subset \{1, \ldots, p\}} n |R_A| > \frac{\kappa'}{8} n^{1-\xi_0} \right) \to 0,$$

we have

$$P \left( \min_{A_1 : A_1 \cap \mathcal{G}_1 \neq \emptyset} \max_j \hat{l}_{j|A_1} \geq \frac{\kappa'}{4} n^{1-\xi_0} \right) \to 1,$$

as $n \to \infty$.

When variable $A_1^c \cap \mathcal{G}_1 = \emptyset$, for $j \in A_1^c \subset \mathcal{G}_0$, $M_j = E(X_j|X_{A_1}, Z) = E(X_j|X_{A_1})$ is a linear combination of $X_{A_1}$ under model (3), and

$$T_{j|A_1} = \frac{\text{Var}(M_j) - \text{Cov}(M_j, X_{A_1}) \left[\text{Var}(X_{A_1})\right]^{-1} \text{Cov}(M_j, X_{A_1})^T}{V_j} = 0.$$

Thus,

$$\hat{l}_{j|A_1} \leq n \left(|R_{A_1^+}| + |R_{A_1}|\right),$$

and

$$P \left( \max_{A_1 : A_1 \cap \mathcal{G}_1 = \emptyset} \max_j \hat{l}_{j|A_1} \geq C n^{1-\xi_0} \right) \leq P \left( \max_{A \subset \{1, \ldots, p\}} |R_A| \geq C n^{-\xi_0} \right) \to 0$$

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

### A.5 Proof of Theorem 6

We will need the following lemma to prove Theorem 6.

**Lemma 3.** Under the same condition as Theorem 6, for $0 < \epsilon < 1$, there exists positive constants $C_1$ and $C_2$ such that

$$P \left( \max_{A \subset \{1, \ldots, p\}} \max_{j \in A^c} \left| \log \sigma_j^2 - \log \sigma_j^2 \right| > \epsilon \right) \leq \frac{p(p + 1)}{2} C_1 \exp \left( -C_2 n \frac{\epsilon^2}{p^2 L^2} \right),$$

and

$$P \left( \max_{A \subset \{1, \ldots, p\}} \max_{j \in A^c} \sum_{h=1}^H s_h \log \left[ \sigma_j^{(h)} \right] - \sum_{h=1}^H s_h \log \left[ \sigma_j^{(h)} \right]^2 > \epsilon \right) \leq \frac{Hp(p + 1)}{2} C_1 \exp \left( -C_2 n \frac{\epsilon^2}{H^2 p^2 L^2} \right),$$

where $L = \frac{4}{\tau_{\min}} \left( 3 \left( \frac{r_{\max}}{\tau_{\min}} \right)^{3/2} + 1 \right)$. 39
Proof of Lemma 3. Let $V_A = \text{Var}(X_A) = (v_{j_1,j_2})_{j_1,j_2 \in A}$, $r_{j,A} = \text{Cov}(X_j, X_A)^T$ and $v_j = \text{Var}(X_j)$, and let $\hat{V}_A = (\hat{v}_{j_1,j_2})_{j_1,j_2 \in A}$, $\hat{r}_{j,A}$ and $\hat{v}_j$ be the corresponding estimates. Then, 

\[
\sigma_j^2 = v_j - r_{j,j}^T V_A^{-1} r_{j,A}, \quad \bar{\sigma}_j^2 = \hat{v}_j - \hat{r}_{j,j}^T \hat{V}_A^{-1} \hat{r}_{j,A}, \quad \text{and} \quad ||\bar{\sigma}_j^2 - \sigma_j^2|| \leq ||\hat{v}_j - v_j|| + ||\hat{r}_{j,j}^T \hat{V}_A^{-1} \hat{r}_{j,j} - \hat{r}_{j,j}^T V_A^{-1} r_{j,j}|| + ||\hat{r}_{j,j}^T V_A^{-1} \hat{r}_{j,j} - r_{j,j}^T V_A^{-1} r_{j,j}||.
\]

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

\[
\max_{\mathcal{A} \subset \{1,\ldots,p\}} \max_{j \in \mathcal{A}} ||\hat{v}_j - v_j|| \leq M,
\]

and 

\[
||\hat{r}_{j,j}^T \hat{V}_A^{-1} \hat{r}_{j,j} - r_{j,j}^T V_A^{-1} r_{j,j}|| = \left| \left( \hat{r}_{j,j} - r_{j,j} \right)^T V_A^{-1} \left( \hat{r}_{j,j} + r_{j,j} \right) \right|,
\]

\[
||\hat{r}_{j,j}^T \hat{V}_A^{-1} \hat{r}_{j,j} - \hat{r}_{j,j}^T \hat{V}_A^{-1} \hat{r}_{j,j}|| = \left( \hat{V}_A^{-1} \hat{r}_{j,j} - V_A^{-1} r_{j,j} \right) \left( V_A^{-1} \hat{r}_{j,j} - V_A^{-1} r_{j,j} \right) = \left| \eta_1 \left| \eta_2 \right| \eta_1^T \left( \hat{V}_A - V_A \right) \eta_2^* \right|,
\]

where $\eta_1 = \hat{V}_A^{-1} \hat{r}_{j,j}$, $\eta_2 = V_A^{-1} r_{j,j}$, and $\eta_1^* = \frac{\eta_1}{||\eta_1||}$, $\eta_2^* = \frac{\eta_2}{||\eta_2||}$.

Since 

\[
\tau_{\min} \leq \min_{\mathcal{A} \subset \{1,\ldots,p\}} \lambda_{\min} \{ V_A \} \leq \max_{\mathcal{A} \subset \{1,\ldots,p\}} \lambda_{\max} \{ V_A \} \leq \tau_{\max},
\]

using similar arguments to the proof of Lemma 1 in Wang (2009) with $p = o(n^{\rho_0})$ and $2\rho_0 + 2\xi_0 < 1$, we have 

\[
2^{-1} \tau_{\min} \leq \min_{\mathcal{A} \subset \{1,\ldots,p\}} \lambda_{\min} \{ \hat{V}_A \} \leq \max_{\mathcal{A} \subset \{1,\ldots,p\}} \lambda_{\max} \{ \hat{V}_A \} \leq 2 \tau_{\max}.
\]

Then, 

\[
||\hat{V}_A^{-1/2} \hat{r}_{j,j}||^2 = \hat{r}_{j,j}^T \hat{V}_A^{-1} \hat{r}_{j,j} \leq \hat{v}_j \leq 2 \tau_{\max}, \quad ||\hat{V}_A^{1/2}|| \leq \sqrt{2 \tau_{\max}}, \quad \text{and} \quad ||\hat{V}_A^{-1/2}|| \leq \frac{\sqrt{2}}{\tau_{\min}}.
\]

Thus, 

\[
||\eta_1|| = ||\hat{V}_A^{-1/2} \hat{r}_{j,j}|| \leq ||\hat{V}_A^{-1/2}|| ||\hat{V}_A^{-1/2} \hat{r}_{j,j}|| \leq 2 \sqrt{\frac{\tau_{\max}}{\tau_{\min}}},
\]

and 

\[
||\eta_2|| = ||V_A^{-1} \hat{r}_{j,j}|| \leq ||V_A^{-1}|| ||V_A^{-1/2}|| ||V_A^{-1/2} \hat{r}_{j,j}|| \leq 2 \frac{\tau_{\max}}{\tau_{\min}}.
\]

Furthermore, 

\[
||\eta_1^* (\hat{V}_A - V_A) \eta_2^*|| \leq \max_{j_1,j_2 \in \{1,\ldots,p\}} ||\hat{v}_{j_1,j_2} - v_{j_1,j_2}|| \sum_{j_1,j_2} ||\eta_{j_1}^*|| ||\eta_{j_2}^*|| = M \left( \sum_{j_1} ||\eta_{j_1}^*|| \right) \left( \sum_{j_2} ||\eta_{j_2}^*|| \right) \leq Mp.
\]

Hence, 

\[
||\hat{r}_{j,j}^T \hat{V}_A^{-1} \hat{r}_{j,j} - r_{j,j}^T V_A^{-1} r_{j,j}|| \leq 4 \left( \frac{\tau_{\max}}{\tau_{\min}} \right)^{3/2} M p.
\]

Since 

\[
||\hat{r}_{j,j} - r_{j,j}|| \leq \max_{j_1,j_2 \in \{1,\ldots,p\}} ||\hat{v}_{j_1,j_2} - v_{j_1,j_2}|| \sqrt{p} = M \sqrt{p}, \quad ||V_A^{-1} \hat{r}_{j,j}|| \leq \frac{2 \tau_{\max}}{\tau_{\min}}, \quad \text{and} \quad ||V_A^{-1} r_{j,j}|| \leq \frac{\sqrt{2}}{\tau_{\min}},
\]

\[
||\hat{r}_{j,j}^T V_A^{-1} (\hat{r}_{j,j} + r_{j,j})|| \leq ||\hat{r}_{j,j} - r_{j,j}|| \left( ||V_A^{-1} \hat{r}_{j,j}|| + ||V_A^{-1} r_{j,j}|| \right) \leq \left( \frac{2 \tau_{\max}}{\tau_{\min}} + \frac{\sqrt{2}}{\tau_{\min}} \right) M \sqrt{p}.
\]
Proof of Theorem 6. Let $R'_{j|A} = \log \hat{\sigma}_j^2 - \log \sigma_j^2$ and $R''_{j|A} = \sum_{h=1}^H s_h \log \hat{\sigma}_j^{(h)2} - \sum_{h=1}^H s_h \log \sigma_j^{(h)2}$. Then, according to Lemma 3, for $0 < \epsilon < 1$, there exists constant $C_1$ and $C_2$ such that

$$P \left( \max_{A \subset \{1, \ldots, p\}} \max_{j \in A^c} \left| R'_{j|A} \right| > \epsilon \right) \leq \frac{p(p+1)}{2} C_1 \exp \left( -C_2 n \frac{\epsilon^2}{H^2 p^2 L^2} \right),$$

$$P \left( \max_{A \subset \{1, \ldots, p\}} \max_{j \in A^c} \left| R''_{j|A} \right| > \epsilon \right) \leq \frac{p(p+1)}{2} C_1 \exp \left( -C_2 n \frac{\epsilon^2}{H^2 p^2 L^2} \right).$$

**Proof of Theorem 6.** Let $R'_{j|A} = \log \hat{\sigma}_j^2 - \log \sigma_j^2$ and $R''_{j|A} = \sum_{h=1}^H s_h \log \hat{\sigma}_j^{(h)2} - \sum_{h=1}^H s_h \log \sigma_j^{(h)2}$. Then, according to Lemma 3, for $0 < \epsilon < 1$, there exists constant $C_1$ and $C_2$ such that

$$P \left( \max_{A \subset \{1, \ldots, p\}} \max_{j \in A^c} \left| R'_{j|A} \right| > \epsilon \right) \leq \frac{p(p+1)}{2} C_1 \exp \left( -C_2 n \frac{\epsilon^2}{p^2 L^2} \right),$$

$$P \left( \max_{A \subset \{1, \ldots, p\}} \max_{j \in A^c} \left| R''_{j|A} \right| > \epsilon \right) \leq \frac{p(p+1)}{2} C_1 \exp \left( -C_2 n \frac{\epsilon^2}{H^2 p^2 L^2} \right).$$

\[ \square \]
and

\[ P\left( \max_{A \subseteq \{1, \ldots, p\}} \max_{j \in A^c} |R''_{j|A}| > \epsilon \right) \leq \frac{H p(p+1)}{2} C_1 \exp \left( -C_2 n^{-\frac{3}{2H^2p^2}} \right), \]

where \( L = \frac{4}{\min} \left( 3 \left( \frac{2\max}{R_{\min}} \right)^{3/2} + 1 \right) \). Under Condition 3, \( p = o(n^{\delta_0}) \) and \( 2p_0 + 2\xi_0 < 1 \), and for any positive constant \( C_0 \),

\[ P\left( \max_{A \subseteq \{1, \ldots, p\}} \max_{j \in A^c} |R''_{j|A}| > C_0 n^{-\xi_0} \right) \leq \frac{p(p+1)}{2} C_1 \exp \left( -C_2 n^{1-2\xi_0-2\rho_0} \frac{C_0^2}{T^2} \right) \to 0, \]

and

\[ P\left( \max_{A \subseteq \{1, \ldots, p\}} \max_{j \in A^c} |R''_{j|A}| > C_0 n^{-\xi_0} \right) \leq \frac{H p(p+1)}{2} C_1 \exp \left( -C_2 n^{1-2\xi_0-2\rho_0} \frac{C_0^2}{H^2L^2} \right) \to 0, \]

as \( n \to \infty \).

From Theorem 4,

\[ \frac{\hat{\rho}_{\text{aug}}_{j|12}}{n} \]

\[ = \log \sigma_j^2 - \sum_{h=1}^{H} s_h \log \left[ \sigma_j^{(h)} \right]^2 \]

\[ \geq \log \sigma_j^2 - \sum_{h=1}^{H} s_h \log \left[ \sigma_j^{(h)} \right]^2 + R'_{j|12} - R''_{j|12} \]

\[ \geq \log \left( 1 + \frac{\text{Var}(M_j) - \text{Cov}(M_j, X_{A_{12}}) [\text{Var}(X_{A_{12}})]^{-1} \text{Cov}(M_j, X_{A_{12}})^T}{E(V_j)} \right) \]

\[ + \log (EV_j) - E \log (V_j) + R'_{j|12} - R''_{j|12}, \]

where \( M_j = E(X_j | X_{A_{12}}, Z) \) and \( V_j = \text{Var}(X_j | X_{A_{12}}, Z) \).

Since \( G_{12} = D \), when \( A_{12}^c \cap G_{12} = A_{12}^c \cap D \neq \emptyset \), there exists \( k \geq 1 \) such that \( \cup_{i=0}^{k-1} D^{(i)} \subseteq A_{12} \) and \( A_{12}^c \cap D^{(k)} \neq \emptyset \). According to Definition 2, for some \( j \in A_{12}^c \cap D^{(k)} \subseteq A_{12}^c \cap G_{12} \), either there exists \( \kappa_1 > 0 \) such that

\[ T'_{j|12} = \frac{\text{Var}(M_j) - \text{Cov}(M_j, X_{A_{12}}) [\text{Var}(X_{A_{12}})]^{-1} \text{Cov}(M_j, X_{A_{12}})^T}{E(V_j)} \geq \kappa_1 n^{-\xi_0}, \]

or there exists \( \kappa_2 > 0 \) such that

\[ T''_{j|12} = \log (EV_j) - E \log (V_j) \geq \kappa_2 n^{-\xi_0}. \]

Note that if \( T'_{j|12} \geq \kappa_1 n^{-\xi_0} \), then for sufficiently large \( n \),

\[ n \log \left( 1 + T'_{j|12} \right) \geq \frac{\kappa_1}{2} n^{1-\xi_0}. \]

Thus,

\[ \frac{\hat{\rho}_{\text{aug}}_{j|12}}{n} \geq n \log \left( 1 + T'_{j|12} \right) + nT''_{j|12} - n \left( |R'_{j|12}| + |R''_{j|12}| \right) \]

\[ \geq \min \left( \frac{\kappa_1}{2}, \kappa_2 \right) n^{1-\xi_0} - n \left( |R'_{j|12}| + |R''_{j|12}| \right) \]

42
Thus, for any positive constant $\kappa$, we have

$$P\left(\max_{j \in \mathcal{A}^c} \max_{j \in \mathcal{A}^c} \left| R_{j|A}^2 \right| > \min\left(\frac{\kappa_1}{8}, \frac{\kappa_2}{4}\right) n^{-\xi_0}\right) \to 0,$$

and

$$P\left(\max_{j \in \mathcal{A}^c} \max_{j \in \mathcal{A}^c} \left| R_{j|A}''^2 \right| > \min\left(\frac{\kappa_1}{8}, \frac{\kappa_2}{4}\right) n^{-\xi_0}\right) \to 0,$$

we have

$$P\left(\min_{\mathcal{A}_{1:2}} \max_{j \in \mathcal{A}_{1:2}} \hat{j}_{j|\mathcal{A}_{1:2}} \leq \min\left(\frac{\kappa_1}{4}, \frac{\kappa_2}{2}\right) n^{1-\xi_0}\right) \to 1,$$

as $n \to \infty$.

When $\mathcal{A}_{1:2} \cap \mathcal{G}_{1:2} = \emptyset$, for any $j \in \mathcal{A}_{1:2} \subset \mathcal{G}_0$, under model (3), $M_j = E \left( X_j | X_{\mathcal{A}_1}, Z \right) = E \left( X_j | X_{\mathcal{A}_{1:2}} \right)$, which is a linear combination of $X_{\mathcal{A}_1}$, and $V_j = \text{Var} \left( X_j | X_{\mathcal{A}_{1:2}}, Z \right) = \text{Var} \left( X_j | X_{\mathcal{A}_{1:2}} \right)$ which is a constant. Then,

$$T_{j|\mathcal{A}_{1:2}}' = \frac{\text{Var} (M_j) - \text{Cov} (M_j, X_{\mathcal{A}_{1:2}}) \left[ \text{Var} (X_{\mathcal{A}_{1:2}}) \right]^{-1} \text{Cov} (M_j, X_{\mathcal{A}_{1:2}})^T}{E (V_j)} = 0,$$

and

$$T_{j|\mathcal{A}_{1:2}}'' = \log (EV_j) - E (V_j) = 0.$$

Thus,

$$\hat{j}_{j|\mathcal{A}_{1:2}} \leq \left\lfloor \frac{1}{n} \left( \left| R_{j|\mathcal{A}_{1:2}}' \right| + \left| R_{j|\mathcal{A}_{1:2}}'' \right| \right) \right\rfloor,$$

and

$$P\left(\max_{\mathcal{A}_{1:2}} \max_{\mathcal{A}_{1:2}} \hat{j}_{j|\mathcal{A}_{1:2}} < C n^{1-\xi_0}\right) \to 1,$$

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

\subsection*{A.6 Proof of Theorem 7}

\textit{Proof of Theorem 7.} (a) First, we will prove that

$$P\left( \max_{j \in \{1, \ldots, p\}} \left| \hat{j}_{j|\mathcal{Y}} - \hat{j}_{j|\mathcal{X}} \right| > C_0 n^{-\xi_0}\right) \to 0,$$

for any constant $C_0 > 0$ as $n \to \infty$.

\begin{align*}
P\left( \max_{j \in \{1, \ldots, p\}} \left| \hat{\sigma}_j^2 - \sigma_j^2 \right| > \epsilon \right) & \leq P\left( \max_{j \in \{1, \ldots, p\}} \left| \hat{\sigma}_j^2 - \sigma_j^2 \right| > \frac{\epsilon}{2} \right) \\
& \leq P\left( \max_{j \in \{1, \ldots, p\}} \left| \hat{\sigma}_j^2 - \sigma_j^2 \right| > \frac{\epsilon r_{\text{min}}}{2} \right) \leq \sum_{j \in \{1, \ldots, p\}} P\left( \left| \hat{\sigma}_j^2 - \sigma_j^2 \right| > \frac{\epsilon r_{\text{min}}}{2} \right) \\
& \leq p C_4 \exp\left( -C_2 n \frac{\epsilon r_{\text{min}}}{4} \right),
\end{align*}

43
where the last inequality follows from Bernstein inequality since \( X \) follows either a multivariate normal distribution or a finite mixture of multivariate normal distributions. Since \( \log(p) = o(n^{\gamma_0}) \) with \( \gamma_0 > 0 \) and \( \gamma_0 + 2\xi_0 < 1 \),

\[
P \left( \max_{j \in \{1, \ldots, p\}} \left| \log \hat{\sigma}_j^2 - \log \sigma_j^2 \right| > C_0 n^{-\xi_0} \right) \leq pC_1 \exp \left( -C_2 n^{1-2\xi_0} \frac{C_1^2 \tau_{\min}}{4} \right)
\]

\[
\leq C_1 \exp \left( -C_2 C_0^2 n^{1-2\xi_0} \frac{\tau_{\min}}{4} + n^{\gamma_0} \right) \to 0,
\]

for any constant \( C_0 > 0 \) as \( n \to \infty \). Similarly,

\[
P \left( \max_{j \in \{1, \ldots, p\}} \left| \log \hat{\sigma}_j^{(h)} - \log \sigma_j^{(h)} \right| > C_0 n^{-\xi_0} \right)
\]

\[
\leq \sum_{h=1}^H P \left( \max_{j \in \{1, \ldots, p\}} \left| \log \hat{\sigma}_j^{(h)} - \log \sigma_j^{(h)} \right| > C_0 n^{-\xi_0} \right)
\]

\[
\leq H pC_1 \exp \left( -C_2 n^{1-2\xi_0} \frac{C_0^2 \tau_{\min}}{4H^2} \right)
\]

\[
\leq HC_1 \exp \left( -C_2 C_0^2 n^{1-2\xi_0} \frac{\tau_{\min}}{4H^2} + n^{\gamma_0} \right) \to 0,
\]

for any constant \( C_0 > 0 \) as \( n \to \infty \). Thus,

\[
P \left( \max_{j \in \{1, \ldots, p\}} \left| \frac{n_{\text{aug}}}{n} - \frac{\hat{l}_j}{n} \right| > C_0 n^{-\xi_0} \right)
\]

\[
= P \left( \max_{j \in \{1, \ldots, p\}} \left( \log \hat{\sigma}_j^2 - \log \sigma_j^2 \right) + \sum_{h=1}^H s_h \left( \log \hat{\sigma}_j^{(h)} - \log \sigma_j^{(h)} \right) > C_0 n^{-\xi_0} \right) \to 0,
\]

for any constant \( C_0 > 0 \) as \( n \to \infty \).

Second, note that if \( \frac{\var{E(X_j|Z)}}{E(\var{X_j|Z})} \geq \kappa_1 n^{-\xi_0} \), then for sufficiently large \( n \),

\[
n \log \left[ 1 + \frac{\var{E(X_j|Z)}}{E(\var{X_j|Z})} \right] \geq \frac{\kappa_1}{2} n^{1-\xi_0}.
\]

So according to Condition 4, there exists \( \kappa_0 = \min\{\frac{\kappa_1}{4}, \frac{\kappa_2}{2}\} > 0 \) such that

\[
n \tilde{l}_{\text{aug}} \geq 2\kappa_0 n^{1-\xi_0}, \quad \text{for } j \in \mathcal{G}_{1.2}.
\]

The events \( \{ \mathcal{G}_{1.2} \not\subset \hat{\mathcal{M}}_{\xi_0, \kappa_0} \} \subset \{ \max_{j \in \{1, \ldots, p\}} \left| n_{\text{aug}} \tilde{l}_j - \hat{l}_j \right| > \kappa_0 n^{1-\xi_0} \} \). Thus,

\[
P \left( \mathcal{G}_{1.2} \not\subset \hat{\mathcal{M}}_{\xi_0, \kappa_0} \right) \leq P \left( \max_{j \in \{1, \ldots, p\}} \left| n_{\text{aug}} \tilde{l}_j - \hat{l}_j \right| > \kappa_0 n^{-\xi_0} \right) \to 0,
\]

as \( n \to \infty \).
(b) Define

\[ M_{\xi,\kappa} = \{ j : \tilde{\tau}_j^{\text{aug}} \geq \kappa n^{-\xi}, \text{ for } 1 \leq j \leq p \} \]

We will prove that there exists \( C > 0 \) such that \( |M_{\xi_0, \kappa_0}| \leq C n^{\xi_0+\kappa_0} \). Since

\[ \left| M_{\xi_0, \kappa_0} \right| \leq \sum_{j=1}^{p} \tilde{\tau}_j^{\text{aug}} , \]

we just need to prove that

\[ \sum_{j=1}^{p} \tilde{\tau}_j^{\text{aug}} \leq C' n^{\kappa_0}, \]

for some positive constant \( C' \). First,

\[ \sum_{j=1}^{p} \log \left[ 1 + \frac{\text{Var} \left( E(X_j|Z) \right)}{\text{Var} \left( X_j|Z \right)} \right] \leq \sum_{j=1}^{p} \text{Var} \left( E(X_j|Z) \right) \leq \frac{1}{\tau_{\min}} \sum_{j=1}^{p} \text{Var} \left( E(X_j|Z) \right) , \]

and

\[ \sum_{j=1}^{p} \text{Var} \left( E(X_j|Z) \right) = \sum_{j \in \mathcal{G}_0} \text{Var} \left( E(X_j|Z) \right) + \sum_{j \in \mathcal{G}_{12}} \text{Var} \left( E(X_j|Z) \right) . \]

According to model (3),

\[ E(X_{\mathcal{G}_0}|X_{\mathcal{G}_{12}}, Z) = \alpha + B_{\mathcal{G}_0|\mathcal{G}_{12}} X_{\mathcal{G}_{12}}, \]

where \( B_{\mathcal{G}_0|\mathcal{G}_{12}} = \text{Cov} (X_{\mathcal{G}_0}, X_{\mathcal{G}_{12}}|Z) [\text{Var} (X_{\mathcal{G}_{12}}|Z)]^{-1} \) is a \( p_0 \) by \( (p_1 + p_2) \) matrix that does not depend on \( X_{\mathcal{G}_{12}} \) or \( Z \). Therefore,

\[ E(X_{\mathcal{G}_0}|Z) = \alpha + B_{\mathcal{G}_0|\mathcal{G}_{12}} E(X_{\mathcal{G}_{12}}|Z) , \]

and

\[ \sum_{j \in \mathcal{G}_0} \text{Var} \left( E(X_j|Z) \right) = \text{trace} \left( \text{Var} \left( E(X_j|Z) \right) \right) \]

\[ = \text{trace} \left( B_{\mathcal{G}_0|\mathcal{G}_{12}} \text{Var} \left( E(X_{\mathcal{G}_{12}}|Z) \right) B_{\mathcal{G}_0|\mathcal{G}_{12}}^T \right) \]

\[ \leq \lambda_{\max} \left( B_{\mathcal{G}_0|\mathcal{G}_{12}} B_{\mathcal{G}_0|\mathcal{G}_{12}}^T \right) \text{trace} \left( \text{Var} \left( E(X_{\mathcal{G}_{12}}|Z) \right) \right) \]

\[ = \lambda_{\max} \left( B_{\mathcal{G}_0|\mathcal{G}_{12}} B_{\mathcal{G}_0|\mathcal{G}_{12}}^T \right) \sum_{j \in \mathcal{G}_{12}} \text{Var} \left( E(X_j|Z) \right) . \]

Since for any \( h = 1, \ldots, H \),

\[ \frac{\lambda_{\max} \left( B_{\mathcal{G}_0|\mathcal{G}_{12}} B_{\mathcal{G}_0|\mathcal{G}_{12}}^T \right)}{\lambda_{\min} \left( \text{Var} (X_{\mathcal{G}_{12}}|Z = h) \right)} \]

\[ \leq \frac{\lambda_{\max} \left( \text{Cov} (X_{\mathcal{G}_0}, X_{\mathcal{G}_{12}}|Z = h) [\text{Var} (X_{\mathcal{G}_{12}}|Z = h)]^{-1} \text{Cov} (X_{\mathcal{G}_0}, X_{\mathcal{G}_{12}}|Z = h)^T \right)}{\lambda_{\min} \left( \text{Var} (X_{\mathcal{G}_{12}}|Z = h) \right)} \]

\[ \leq \frac{\lambda_{\max} \left( \text{Var} (X_{\mathcal{G}_0}|Z = h) \right)}{\lambda_{\min} \left( \text{Var} (X_{\mathcal{G}_{12}}|Z = h) \right)} \leq \frac{\tau_{\max}}{\tau_{\min}} \]
\[ \sum_{j=1}^{p} \log \left[ 1 + \frac{\operatorname{Var}(E(X_j|Z))}{E(\operatorname{Var}(X_j|Z))} \right] \leq \frac{1}{\tau_{\min}} \sum_{j=1}^{p} \operatorname{Var}(E(X_j|Z)) \leq \frac{1}{\tau_{\min}} \left( 1 + \frac{\tau_{\max}}{\tau_{\min}} \right) \sum_{j \in \mathcal{G}_{1:2}} \operatorname{Var}(E(X_j|Z)). \]

Second, for \( j \in \mathcal{G}_0, \operatorname{Var}(X_j|X_{\mathcal{G}_{1:2}}, Z) \) is a constant, and

\[ \log [E(\operatorname{Var}(X_j|Z))] - E[\log(\operatorname{Var}(X_j|Z))] \leq E(\operatorname{Var}(X_j|Z)) E \left( \frac{1}{\operatorname{Var}(X_j|Z)} \right) - 1 \]

\[ \leq \frac{E(\operatorname{Var}(X_j|Z))}{\operatorname{Var}(X_j|X_{\mathcal{G}_{1:2}}, Z)} - 1 = \frac{E[\operatorname{Var}(E(X_j|X_{\mathcal{G}_{1:2}}, Z)|Z)]}{\operatorname{Var}(X_j|X_{\mathcal{G}_{1:2}}, Z)} \leq \frac{1}{\tau_{\min}} \operatorname{E}[\operatorname{Var}(E(X_j|X_{\mathcal{G}_{1:2}}, Z)|Z)] . \]

So

\[ \sum_{j \in \mathcal{G}_0} (\log [E(\operatorname{Var}(X_j|Z))] - E[\log(\operatorname{Var}(X_j|Z))]) \leq \frac{1}{\tau_{\min}} \sum_{j \in \mathcal{G}_0} E[\operatorname{Var}(E(X_j|X_{\mathcal{G}_{1:2}}, Z)|Z)] \]

\[ = \frac{1}{\tau_{\min}} \text{trace} (E(\operatorname{Var}(X_{\mathcal{G}_0}|X_{\mathcal{G}_{1:2}}, Z)|Z)) = \frac{1}{\tau_{\min}} \text{trace} (B_{\mathcal{G}_0|\mathcal{G}_{1:2}} E[\operatorname{Var}(X_{\mathcal{G}_{1:2}}|Z)] B_{\mathcal{G}_0|\mathcal{G}_{1:2}}^T) \]

\[ \leq \frac{1}{\tau_{\min}} \lambda_{\max}(B_{\mathcal{G}_0|\mathcal{G}_{1:2}} B_{\mathcal{G}_0|\mathcal{G}_{1:2}}^T) \sum_{j \in \mathcal{G}_{1:2}} E[\operatorname{Var}(X_j|Z)] \leq \frac{\tau_{\max}}{\tau_{\min}} \sum_{j \in \mathcal{G}_{1:2}} E[\operatorname{Var}(X_j|Z)] . \]

Therefore,

\[ \sum_{j=1}^{p} (\log [E(\operatorname{Var}(X_j|Z))] - E[\log(\operatorname{Var}(X_j|Z))]) \leq \frac{1}{\tau_{\min}} \left( 1 + \frac{\tau_{\max}}{\tau_{\min}} \right) \sum_{j \in \mathcal{G}_{1:2}} E[\operatorname{Var}(X_j|Z)] \]

Since

\[ \sum_{j \in \mathcal{G}_{1:2}} \operatorname{Var}(X_j|Z) \leq |\mathcal{G}_{1:2}| \tau_{\max}, \]

\[ \sum_{j=1}^{p} \hat{l}_{\text{aug}}^{\text{aug}} = \sum_{j=1}^{p} \left( \log \left[ 1 + \frac{\operatorname{Var}(E(X_j|Z))}{E(\operatorname{Var}(X_j|Z))} \right] \right) + \sum_{j=1}^{p} (\log [E(\operatorname{Var}(X_j|Z))] - E[\log(\operatorname{Var}(X_j|Z))]) \]

\[ \leq \frac{1}{\tau_{\min}} \left( 1 + \frac{\tau_{\max}}{\tau_{\min}} \right) \sum_{j \in \mathcal{G}_{1:2}} \operatorname{Var}(X_j|Z) \leq \frac{\tau_{\max}}{\tau_{\min}} \left( 1 + \frac{\tau_{\max}}{\tau_{\min}} \right) |\mathcal{G}_{1:2}| \leq \frac{\tau_{\max}}{\tau_{\min}} \left( 1 + \frac{\tau_{\max}}{\tau_{\min}} \right) c n^{\theta_0}. \]

Thus, there exists \( C > 0 \) such that

\[ |M_{\xi_0, \frac{\tau}{2}}| \leq \sum_{j=1}^{p} \hat{l}_{\text{aug}}^{\text{aug}} \leq \frac{\tau_{\max}}{\tau_{\min}} \left( 1 + \frac{\tau_{\max}}{\tau_{\min}} \right) 2c n^{\xi_0 + \theta_0} \leq C n^{\xi_0 + \theta_0}. \]

Then,

\[ P \left( \left| \hat{M}_{\xi_0, \kappa_0} \right| > C n^{\xi_0 + \theta_0} \right) \leq P \left( \left| \hat{M}_{\xi_0, \kappa_0} \right| > \left| M_{\xi_0, \frac{\tau}{2}} \right| \right), \]

and the event \{\( \left| \hat{M}_{\xi_0, \kappa_0} \right| > \left| M_{\xi_0, \frac{\tau}{2}} \right| \} \subseteq \{\text{there exists } j \text{ such that } n \hat{l}_{j}^{\text{aug}} < \frac{\kappa_0}{2} n^{1 - \xi_0} \text{ and } \hat{l}_{j}^{\text{aug}} \geq \kappa_0 n^{1 - \xi_0} \}

\[ \subseteq \{\max_{j \in \{1, \ldots, p\}} \left| n \hat{l}_{j}^{\text{aug}} - \hat{l}_{j}^{\text{aug}} \right| > \frac{\kappa_0}{2} n^{1 - \xi_0} \}. \]

Thus, according to the results in (a),

\[ P \left( \left| \hat{M}_{\xi_0, \kappa_0} \right| > C n^{\xi_0 + \theta_0} \right) \leq P \left( \max_{j \in \{1, \ldots, p\}} \left| \frac{\hat{l}_{j}^{\text{aug}}}{n} - \frac{\hat{l}_{j}^{\text{aug}}}{n} \right| > \frac{\kappa_0}{2} n^{1 - \xi_0} \right) \to 0, \]

as \( n \to \infty. \)

\[ \square \]
A.7 Proof of Theorem 8

Proof of Theorem 8. Since \( \text{Var} (X_j|Z') = E (\text{Var}(X_j|X_A, Z')|Z') + \text{Var} (E(X_j|X_A, Z')|Z') \),

\[
E (\text{Var}(X_j|Z')) = E (\text{Cov}(X_j, X_A|Z')) [E(\text{Var}(X_A|Z'))]^{-1} E (\text{Cov}(X_j, X_A|Z'))^T
\]

\[
= E (\text{Var}(X_j|X_A, Z')) + E [\text{Var}(E(X_j|X_A, Z')|Z')] - E [\text{Cov}(E(X_j|X_A, Z'), X_A|Z')] [E(\text{Var}(X_A|Z'))]^{-1} E [\text{Cov}(E(X_j|X_A, Z'), X_A|Z')]^T
\]

\[
\geq E (\text{Var}(X_j|X_A, Z'))
\]

where the equality holds if \( E(X_j|X_A, Z') \) is a linear combination of \( X_A \) that does not depend on \( Z' \). Since \( S \preceq S' \), the \( \sigma \)-algebra \( \sigma(Z') \subset \sigma(Z) \). Thus,

\[
\text{Var}(X_j|X_A, Z') = E(\text{Var}(X_j|X_A, Z')|X_A, Z') + \text{Var}(E(X_j|X_A, Z')|X_A, Z') \geq E(\text{Var}(X_j|X_A, Z')|X_A, Z'),
\]

and

\[
E(\text{Var}(X_j|X_A, Z')) \geq E(\text{Var}(X_j|X_A, Z)),
\]

where the equality holds if \( E(X_j|X_A, Z) \) is a linear combination of \( X_A \) that does not depend on \( Z \). Because \( Z \) is the true slicing scheme, under model assumptions (4) and (3) for variables in group 1 and group 0, \( \text{Var}(X_j|X_A, Z) \) is a constant that does not depend on \( X_A \) or \( Z \), i.e. \( E(\text{Var}(X_j|X_A, Z)) = \text{Var}(X_j|X_A, Z) \). Therefore,

\[
E(\text{Var}(X_j|Z')) - E(\text{Cov}(X_j, X_A|Z')) [E(\text{Var}(X_A|Z'))]^{-1} E(\text{Cov}(X_j, X_A|Z'))^T \geq \text{Var}(X_j|X_A, Z)
\]

and

\[
\bar{I}_{j,A,S} = \log \left[ \text{Var}(X_j) - \text{Cov}(X_j, X_A) [\text{Var}(X_A)]^{-1} \text{Cov}(X_j, X_A)^T \right] - \log (\text{Var}(X_j|X_A, Z)) \geq \log \left[ \text{Var}(X_j) - \text{Cov}(X_j, X_A) [\text{Var}(X_A)]^{-1} \text{Cov}(X_j, X_A)^T \right] - \log \left[ E(\text{Var}(X_j|Z')) - E(\text{Cov}(X_j, X_A|Z')) [E(\text{Var}(X_A|Z'))]^{-1} E(\text{Cov}(X_j, X_A|Z'))^T \right] \]

\[
= \bar{I}_{j,A,S'}.
\]

When \( G_1 \subset A \), under model assumptions (4) and (3), the variable \( j \) is in group 0 and has the same conditional distribution across difference slices given \( X_A \). So \( E(X_j|X_A, Z) \) and \( E(X_j|X_A, Z') \) are linear combinations of \( X_A \) that do not depend on \( Z \) or \( Z' \). Thus, the equalities hold in this case.

To prove (b), note that

\[
\text{Var}(X_j|Z') - \text{Cov}(X_j, X_A|Z') [\text{Var}(X_A|Z')]^{-1} \text{Cov}(X_j, X_A|Z')^T
\]

\[
= E(\text{Var}(X_j|X_A, Z')|Z') + \text{Var}(E(X_j|X_A, Z')|Z') - \text{Cov}(E(X_j|X_A, Z'), X_A|Z') [\text{Var}(X_A|Z')]^{-1} \text{Cov}(E(X_j|X_A, Z'), X_A|Z')^T
\]

\[
\geq E(\text{Var}(X_j|X_A, Z')|Z') .
\]

47
where the equality holds if $E(X_j|X_A, Z')$ is a linear combination of $X_A$ that does not depend on $Z'$. Since $S \preceq S'$, the $\sigma$-algebra $\sigma(Z') \subset \sigma(Z)$. Thus,

$$
\text{Var}(X_j|X_A, Z') = E(\text{Var}(X_j|X_A, Z)|X_A, Z') + \text{Var}(E(X_j|X_A, Z)|X_A, Z') \\
\geq E(\text{Var}(X_j|X_A, Z)|X_A, Z') ,
$$

and

$$
\text{Var}(X_j|Z') - \text{Cov}(X_j, X_A|Z') [\text{Var}(X_A|Z')^{-1}] \text{Cov}(X_j, X_A|Z')^T \\
\geq E(\text{Var}(X_j|X_A, Z')|Z') \geq E(\text{Var}(X_j|X_A, Z)|Z') .
$$

where the equalities hold if $E(X_j|X_A, Z)$ and $E(X_j|X_A, Z')$ are linear combinations of $X_A$ that do not depend on $Z$ or $Z'$. According to Jensen’s inequality,

$$
\log[E(\text{Var}(X_j|X_A, Z)|Z')] \geq E(\log[\text{Var}(X_j|X_A, Z)]|Z') .
$$

where the equality holds if $\text{Var}(X_j|X_A, Z)$ is a constant that does not depend on $X_A$ or $Z$. Therefore,

$$
E\left(\log[\text{Var}(X_j|Z') - \text{Cov}(X_j, X_A|Z') [\text{Var}(X_A|Z')^{-1}] \text{Cov}(X_j, X_A|Z')^T]\right) \\
\geq E(\log[E(\text{Var}(X_j|X_A, Z')|Z')]) \\
\geq E(E(\log[\text{Var}(X_j|X_A, Z)]|Z')) \\
= E(\log[\text{Var}(X_j|X_A, Z)]) ,
$$

and

$$
\bar{p}_{ji|A,S}^{\text{aug}} = \log[\text{Var}(X_j) - \text{Cov}(X_j, X_A) [\text{Var}(X_A)^{-1}] \text{Cov}(X_j, X_A)^T] - E(\log[\text{Var}(X_j|X_A, Z)]) \\
\geq \log[\text{Var}(X_j) - \text{Cov}(X_j, X_A) [\text{Var}(X_A)^{-1}] \text{Cov}(X_j, X_A)^T] - \\
E(\log[\text{Var}(X_j|Z') - \text{Cov}(X_j, X_A|Z') [\text{Var}(X_A|Z')^{-1}] \text{Cov}(X_j, X_A|Z')^T]) \\
= \bar{p}_{ji|A,S'}^{\text{aug}} .
$$

When $\mathcal{G}_1 \cup \mathcal{G}_2 \subset \mathcal{A}$, under the augmented model, the variable $j$ is in group 0 and has the same conditional distribution across difference slices given $X_A$. So $E(X_j|X_A, Z)$ and $E(X_j|X_A, Z')$ are linear combinations of $X_A$ that do not depend on $Z$ or $Z'$, and $\text{Var}(X_j|X_A, Z)$ is a constant that does not depend on $X_A$ or $Z$. Thus, the equalities hold in this case. \(\square\)

**References**

Zhang, W., Jiang, J., Loh, Y. H., Yeo, H. C., Yeo, Z. X., Narang, V., Govindarajan, K.
R., Leong, B., Shahab, A., Ruan, Y., Bourque, G., Sung, W. K., Clarke, N. D., Wei,
C. L. and Ng, H. H. (2008) Integration of external signaling pathways with the core
transcriptional network in embryonic stem cells. Cell, 133, 1106-1117.

C., Bruce, S. J., Lee, C. C., Ranade, S. S., Peckham, H. E., Manning, J. M., McKernan,


1, 302-332.

Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. and Lander,
E. S. (1999) Molecular classification of cancer: class discovery and class prediction by

ican Statistical Association, 86: 316-327.


