Model-free variable selection in sufficient dimension reduction via FDR control

Yixin Han¹, Xu Guo² & Changliang Zou³

¹Department of ISOM, Hong Kong University of Science and Technology
²Department of Mathematical Statistics, Beijing Normal University
³School of Statistics and Data Science, LPMC & KLMDASR, Nankai University

Abstract

Simultaneously identifying contributory variables and controlling the false discovery rate (FDR) in high-dimensional data is an important statistical problem. In this paper, we propose a novel model-free variable selection procedure in sufficient dimension reduction via data splitting technique. The variable selection problem is first connected with a least square procedure with several response transformations. We construct a series of statistics with global symmetry property and then utilize the symmetry to derive a data-driven threshold to achieve error rate control. This method can achieve finite-sample and asymptotic FDR control under some mild conditions. Numerical experiments indicate that our procedure has satisfactory FDR control and higher power compared with existing methods.

Keywords: Data splitting; False discovery rate; Model-free; Sufficient dimension reduction; Symmetry
1 Introduction

Sufficient dimension reduction (SDR) is a powerful technique to extract relevant information from high-dimensional data (Li, 1991; Cook and Weisberg, 1991; Xia et al., 2002; Li and Wang, 2007). We use $Y$ with support $\Omega_Y$ to denote the univariate response, and let $X = (X_1, \ldots, X_p)^\top \in \mathbb{R}^p$ be the $p$-dimensional vector of all covariates. The basic idea of SDR is to replace the predictor vector with its projection onto a subspace of the predictor space without loss of information on the conditional distribution of $Y$ given $X$. If a predictor subspace $C \subseteq \mathbb{R}^p$ satisfies $Y \perp \perp X | \mathcal{P}_C X$, where “$\perp \perp$” stands for independence and $\mathcal{P}(\cdot)$ represents the projection matrix with respect to the standard inner product, then $C$ is called a dimension reduction subspace.

In practice, a large number of features in high-dimensional data are collected but only a small portion of them are truly associated with the response. However, while grasping important features or patterns in the data, the reduction subspace from SDR usually consists of all original variables which makes it difficult to interpret. In this paper, we aim at developing a model-free variable selection procedure to screen out truly non-contributing variables with certain error rate control. Let $F(Y | X)$ be the conditional distribution function of $Y$ given $X$. The index set of the active and inactive variables are defined respectively as

\[ \mathcal{A} = \{j : F(Y | X) \text{ functionally depends on } X_j, j = 1, \ldots, p\}, \]

\[ \mathcal{A}^c = \{j : F(Y | X) \text{ does not functionally depend on } X_j, j = 1, \ldots, p\}. \]

Many popular variable selection procedures were developed under the paradigm of linear
models/generalized linear models. See the review of Fan and Lv (2010) and the book of Fan et al. (2020) for a fuller list of references. In contrast, model-free variable selection can be achieved by SDR since it does not require the full knowledge of the underlying model, thus researchers do not need to dispose of the model misspecification.

SDR methods with variable selection aim to find the active set $\mathcal{A}$ such that

$$Y \perp \perp X_{\mathcal{A}^c} | X_{\mathcal{A}},$$

where $X_{\mathcal{A}} = \{X_j : j \in \mathcal{A}\}$ denotes the vector containing all active covariates and $X_{\mathcal{A}^c}$ is the complementary set of $X_{\mathcal{A}}$. Condition (1) implies that $X_{\mathcal{A}}$ contains all the relevant information in term of predicting $Y$. Li et al. (2005) proposed a method that combines sufficient dimension reduction and variable selection. A general shrinkage estimation framework of inverse regression for variable selection was discussed in Bondell and Li (2009). Chen et al. (2010) proposed a coordinate-independent sparse estimation that can simultaneously achieve sparse SDR and screen out irrelevant variables efficiently. Wu and Li (2011) focused on model-free variable selection with a diverging number of predictors. A marginal coordinate hypothesis is proposed by Cook (2004) for model-free variable selection under low-dimensional setting, and then is promoted by Shao et al. (2007) and Yu and Dong (2016). Yu et al. (2016a) constructed marginal coordinate tests for sliced inverse regression (SIR) and Yu et al. (2016b) suggested a trace-pursuit-based utility for ultrahigh-dimensional feature selection. See Li et al. (2020) and Zhu (2020) for some comprehensive review.

However, those existing approaches do not consider the uncertainty quantification of vari-
able selection, i.e., the global error rate control in the selected subset of important covariates. Recently, in the linear regression setting, Barber and Candès (2015) introduced the knockoff filter, a new variable selection procedure controlling the false discovery rate (FDR, Benjamini and Hochberg, 1995) with fixed-designs whenever there are at least as many observations as variables. The knockoff procedure operates by constructing “knockoff copies” of each of the $p$ covariates with a certain knowledge of the covariates or responses. The signs of test statistics constructed via knockoff would satisfy (or roughly) joint exchangeability and thus can yield accurate FDR control in finite samples. The extensions on high-dimensional situations or random designs can be found in Candes et al. (2018) among many others. In a different direction, Du et al. (2021) and Guo et al. (2022) proposed using data splitting strategies, named as symmetrized data aggregation (SDA), to construct a series of statistics with global symmetry property and then utilize the symmetry to derive a data-driven threshold to achieve error rate control. Similar to the knockoff method, the SDA is also free of $p$-values and its construction does not rely on contingent assumptions, which motivates us to employ it in the present problem.

In this paper, we propose a model-free variable selection procedure which could achieve an effective FDR control. We first recast the problem of conducting variable selection in sufficient dimension reduction into making inference on regression coefficients in a set of linear regressions with several response transformations. A variable selection procedure is subsequently developed via error rate control for low-dimensional and high-dimensional settings, respectively. This novel data-driven selection procedure can control the FDR while being combined with different existing SDR methods for model-free variable selection by choos-
ing different response transformation functions. The proposed procedure is computationally efficient and is convenient to implement since it only involves a one-time split of the data and the calculation of the product of two dimension reduction matrices obtained from two splits. This method can achieve finite-sample and asymptotic FDR control under some mild conditions. Numerical experiments indicate that our procedure has satisfactory FDR control and higher power compared with existing methods.

The rest of this paper is organized as follows. In section 2, we reframe the variable selection in SDR formulation as a multiple testing problem. In section 3, we firstly propose a low-dimensional variable selection procedure with error rate control and then discuss its extension in high-dimensional situations. The finite-sample and asymptotic theories for FDR control are developed in Section 4. Simulation studies and a real-data investigation are conducted in Section 5 to demonstrate the superior performance of the proposed method. Section 6 concludes the paper. The main theoretical proofs are given in Appendix. More detailed proofs and some additional numerical results are delineated in the Supplementary Material.

Notations. Let \(\lambda_{\min}(B)\) and \(\lambda_{\max}(B)\) denote the smallest and largest eigenvalues of square matrix \(B = (b_{ij})\). Write \(\|B\|_2 = (\sum_i \sum_j b_{ij}^2)^{1/2}\) and \(\|B\|_\infty = \max_i \sum_j |b_{ij}|\). Denote \(\|\mu\|_1 = \sum_i |\mu_i|\) and \(\|\mu\|_2 = (\sum_i \mu_i^2)^{1/2}\) be the \(L_1\) and \(L_2\) norm of vector \(\mu\). Denote \(\mathbb{E}(X)\) and \(\text{cov}(X)\) be the expectation and covariance for random vector \(X\), respectively. Let \(A_n \approx B_n\) denote that two quantities \(A_n\) and \(B_n\) are asymptotically equivalent, in the sense that there is a constant \(C > 1\) such that \(B_n/C \leq A_n \leq B_nC\) with probability tending to 1. The “\(\gtrsim\)” and “\(\lesssim\)” are similarly defined.
2 Problem and model formulation

The variable selection in (1) can be framed as a multiple testing problem

\[ H_0' : j \in A^c \text{ versus } H_1' : j \in A. \]  

(2)

This is known as the marginal coordinate hypothesis described in Cook (2004) and Yu et al. (2016a). Some related works include Li et al. (2005); Shao et al. (2007); Yu and Dong (2016). This type of selection procedure usually uses some nonnegative marginal utility statistics \( W_j \)’s to measure the importance of \( X_j \)’s to \( Y \) in certain sense. However, the global error rate control with those methods remains an issue because the determination of selection thresholds generally involves the approximation to the distribution of \( W_j \), and the accuracy of asymptotic distributions heavily affects the error rate control.

As a remedy, we introduce a novel reformulation for (2). Let \( \Sigma = \text{cov}(X) > 0 \) and assume \( \mathbb{E}(X) = 0 \). For any function \( f(Y) \) satisfying \( \mathbb{E}\{f(Y)\} = 0 \), following Yin and Cook (2002) and Wu and Li (2011) one can show that

\[ \Sigma^{-1} \text{cov}(X, f(Y)) \in C, \]

under the linearity condition (Li, 1991), which is usually satisfied when \( X \) is elliptical distribution. The transformation \( f(\cdot) \) is used in a way different from its traditional role of being a mechanism for improving the goodness of model fitting. It serves as an intermediate tool for performing dimension reduction. Consequently, different transformation functions
correspond to different SDR methods (Dong, 2021). One can choose a series of transformation functions, $f_1(Y), \ldots, f_H(Y)$, whose forms do not depend on data. The $H (> d)$, a pre-specified integer, is usually called a working dimension and $d$ is the true structural dimension of the subspace $\mathcal{C}$. At the population level, define

$$\beta^0_h = \arg \min_{\beta_h} \mathbb{E} \left[ \left\{ f_h(Y) - X^\top \beta_h \right\}^2 \right], \ h = 1, \ldots, H. \quad (3)$$

Write $B_0 = (\beta^0_1, \ldots, \beta^0_H)$, then $\text{Span}(B_0) \subseteq \mathcal{S}_{Y\mid X}$, and $\text{Span}(B_0)$ represents the subspace spanned by the column vector of $B_0$. By the following usual protocol in the literature of sufficient dimension reduction, we take one step further by assuming the coverage condition $\text{Span}(B_0) = \mathcal{S}_{Y\mid X}$ whenever $\text{Span}(B_0) \subseteq \mathcal{S}_{Y\mid X}$. This condition often holds in practice; see Cook and Ni (2006) for a discussion.

For $j = 1, \ldots, p$, let $\beta^0_{hj}$ be the $j$th element of $\beta^0_h \in \mathbb{R}^p$, $h = 1, \ldots, H$. If the $j$th variable is unimportant, $B_j = 0 \in \mathbb{R}^H$, where $B_j = (\beta^0_{1j}, \beta^0_{2j}, \cdots, \beta^0_{Hj})^\top$ denotes the $j$th row of $B_0$. Further, $Y \perp \perp X \mid X_A$ implies that $\sum_{h=1}^{H} |\beta^0_{hj}| > 0$ for $j \in A$ and $\sum_{h=1}^{H} |\beta^0_{hj}| = 0$ for $j \in A^c$. In other words, if $j$ belongs to the active set $A$, response $Y$ must depend on $X_j$ through at least one of the $H$ linear combinations. If $j$ belongs to the inactive set $A^c$, none of the $H$ linear combinations involve $X_j$ (Yu et al., 2016a). This connection implies that we may try to select the active set $A$ through recovering the central space.

Accordingly, the testing problem (2) is equivalent to

$$H_{0j} : \sum_{h=1}^{H} |\beta^0_{hj}| = 0 \text{ versus } H_{1j} : \sum_{h=1}^{H} |\beta^0_{hj}| > 0. \quad (4)$$
Based on above discussion, selecting active variables in a model-free framework is equivalent to select important variables in *multiple response linear model*. Assume that there are independent and identical distributed data \( D = \{X_i, Y_i\}_{i=1}^{2n} \). The least square optimization in (3) is solved by

\[
\hat{\beta}_h = \arg \min_{\beta_h} \left( 2n \sum_{i=1}^{2n} \left\{ f_h(Y) - X^\top \beta_h \right\}^2 \right), \ h = 1, \ldots, H, \tag{5}
\]

where \( \hat{\beta}_1, \ldots, \hat{\beta}_H \) are the estimators of \( \beta_1^0, \ldots, \beta_H^0 \). Denote \( W_j \)'s as the marginal statistics based on the sample \( D \) associated with the variants of \( \hat{\beta}_1, \ldots, \hat{\beta}_H \). Its explicit form would be given in next section. A selection procedure with a threshold \( L \) is formed as

\[
\hat{A}(L) = \{ j : W_j \geq L, \text{ for } 1 \leq j \leq p \}, \tag{6}
\]

where \( \hat{A}(L) \) is the estimate of \( A \) with threshold \( L \). Obviously, \( L \) plays an important role in variable selection to control the model complexity. We will construct an appropriate threshold by controlling the FDR to achieve model-free variable selection in SDR.

Denote \( p_0 = |A^c|, p_1 = |A| \) and assume that \( p_1 \) is dominated by \( p \), i.e., \( p_1 = o(p) \). The false discovery proportion (FDP) associated with the selection procedure (6) is

\[
\text{FDP} \left( \hat{A}(L) \right) = \frac{\# \{ j : j \in \hat{A}(L) \cap A^c \}}{\# \{ j : j \in \hat{A}(L) \} \lor 1},
\]

where \( a \lor b = \max \{ a, b \} \) and \( \# \{ \} \) stands for the cardinality of an event. The FDR is defined as the expectation of the FDP, i.e., \( \text{FDR}(L) = \mathbb{E}(\text{FDP}(L)) \). Our main goal is to find a
data-driven threshold $L$ that controls the asymptotic FDR at a target level $\alpha$

$$\limsup_{n \to \infty} \text{FDR} \left( \hat{A}(L) \right) \leq \alpha.$$ 

3 Variable selection via FDR control

In this section, we first provide a data-driven threshold selection in Subsection 3.1 for controlling the FDR of variable selection procedure via data splitting technique in a model-free context when $p < n$, and the high-dimensional version is postponed in Subsection 3.2.

3.1 Low-dimensional procedure

We first split the full data $\mathcal{D} = \{X_i, Y_i\}_{i=1}^{2n}$ into two independent parts $\mathcal{D}_1 = \{X_{1i}, Y_{1i}\}_{i=1}^{n}$ and $\mathcal{D}_2 = \{X_{2i}, Y_{2i}\}_{i=1}^{n}$ with equal size, which are respectively used to estimate the dimension reduction spaces as $\hat{B}_1$ and $\hat{B}_2$ by (3), where $\hat{B}_1 = (\hat{\beta}_1^{(1)}, \ldots, \hat{\beta}_H^{(1)})$ and $\hat{B}_2 = (\hat{\beta}_1^{(2)}, \ldots, \hat{\beta}_H^{(2)})$ from (5). One can find the unequal size data splitting investigation in Du et al. (2021). On split $\mathcal{D}_k$, $k = 1, 2$, the least square estimator of $B_j$ in (5) is

$$\hat{B}_{kj}^\top = e_j^\top \left( \sum_{i=1}^{n} X_{ki}X_{ki}^\top \right)^{-1} \sum_{i=1}^{n} X_{ki}f_{ki}^\top,$$

where $f = (f_1(Y), \ldots, f_H(Y))^\top$ and $e_j$ is the $p$-dimensional unit vector with the $j$th element being 1. The information in the two parts is then combined by the inner product to form a
symmetrized ranking statistic

$$W_j = \frac{\hat{B}_{1j} \hat{B}_{2j}}{s_{1j}s_{2j}}, \quad j = 1, \ldots, p,$$

(7)

where

$$s_{kj}^2 = e_j^\top (\sum_{i=1}^n X_{ki}X_{ki}^\top)^{-1} e_j, \quad k = 1, 2.$$ 

For an active variable, if $\sum_{h=1}^H |\beta_{hj}|$ is large (under $\mathbb{H}_{1j}$), then both $\hat{B}_{1j}$ and $\hat{B}_{2j}$ have the same sign and tend to have large absolute values, thereby leading to a positive and large $W_j$ (Du et al., 2021). For a null feature, $W_j$ is symmetrically distributed around zero. It implies that $W_j$ owns the marginal symmetry property (Barber and Candès, 2015; Du et al., 2021) for all inactive variables such that it can be adopted to determine active or inactive variables. This motivates us to choose a data-driven threshold $L$ as the following to control the FDR at level $\alpha$

$$L = \inf \left\{ t > 0 : \frac{\# \{ j : W_j \leq -t \}}{\# \{ j : W_j \geq t \} \lor 1} \leq \alpha \right\},$$

(8)

and then our decision rule is given by $\hat{A}(L) = \{ j : W_j \geq L, 1 \leq j \leq p \}$. If $\hat{A}(L)$ is empty, we simply set $L = +\infty$. The fraction in (8) is an estimate of the FDP since $\# \{ j : W_j \leq -t \}$ is a good approximation to $\# \{ j : W_j \geq t, j \in \mathcal{A}^c \}$ by the marginal symmetry of $W_j$ under null.

The core of our procedure is to construct marginal symmetric statistics by data splitting technique, so as to obtain a data-driven threshold to realize variable selection. It is a model-free extension for the SDA filter (Du et al., 2021). Therefore, we refer our method to Model-Free Symmetrized Data Aggregation (MFSDA).

Since the estimators, $\hat{B}_1$ and $\hat{B}_2$, only are two approximations of $B$ and they are not derived by eigenvalue decomposition (Li, 1991; Cook and Weisberg, 1991), there is no concern
about that $\hat{\mathbf{B}}_1$ and $\hat{\mathbf{B}}_2$ may not be in a same subspace and our ultimate goal is to identify the active variables rather than to recover the subspace itself. It implies that variable selection achieved through (3) requires no dimension reduction basis estimation and thus is dispensable for the knowledge of the structural dimension $d$ either. Therefore, the proposed method can be adapted to a family of inverse slice regression estimators by choosing different $f(Y)$. In a nutshell, our method can be widely used due to its simplicity, computational efficiency, and generality. It is summarized as follows.

\textbf{Algorithm 1} Data-driven threshold for variable selection via MFSDA

\textbf{Step 1} (Initialization) Specify $H$, $f$ and $\alpha$;

\textbf{Step 2} (Data splitting) Randomly split the data into two independent parts $\mathcal{D}_1$ and $\mathcal{D}_2$ with equal size. Obtain the dimension reduction estimates $\hat{\mathbf{B}}_1$ and $\hat{\mathbf{B}}_2$ by (3);

\textbf{Step 3} (Ranking statistics) Construct the test statistics $W_j$ by (7) and then rank them;

\textbf{Step 4} (Thresholding) Compute the threshold $L$ in (8), and obtain selected variable set $\hat{\mathcal{A}}(L)$.

The total computational complexity of Algorithm 1 is of order $O(2nHp^2 + p \log p)$ so that this algorithm can be easily implemented. Practically, our method involves data splitting that may lead to some information loss concerning the full data (Du et al., 2021). Fortunately, we obtain a data-driven threshold by the marginal symmetry property of $W_j$ under null, which does not need to find the null asymptotic distribution anymore. Here we use a toy example to illustrate the advantage of data splitting. The data generation details can be founded in Section 5. In Figure 1, we can observe that the data splitting method (left panel) places most active variables above zero, and many inactive variables are symmetrically distributed around zero. This is an extremely important property for our selection procedure while the
full estimation (middle panel) and half data estimation (right panel) methods both fail to achieve this symmetry.

Figure 1: Scatter plot of proposed $W_j$ with red points and black dots denoting active and inactive variables, respectively. Left panel: the proposed $W_j$ in (7); Middle panel: $W_j = \sum_{h=1}^{H} |\hat{\beta}_{hj}|$ with $(\hat{\beta})_{hj} = \hat{\beta}_{hj}$, which is the least square estimator on the full data; Right panel: replace $\hat{\beta}_2$ with $\hat{\beta}_1$ in (7).

3.2 High-dimensional procedure

When the dimension $p$ is very large in practice, the above procedure does not work since the ordinary least square procedure cannot be directly implemented. Note that our data splitting procedure can be essentially extended to the version of regularization form. Inspired by the idea of SDA filter proposed by Du et al. (2021), we then put forward the following selection procedure for high-dimensional data.

To extract information from $D_1$, we replace the least square solution in (3) with LASSO selector (Tibshirani, 1996) as follows

$$\hat{\beta}^{(1)}_{\lambda_h} = \arg \min_{\beta_h} \left[ n^{-1} \sum_{i=1}^{n} \left\{ f_h(Y_i) - X_i^T \beta_h \right\}^2 + \lambda_h \|\beta_h\|_1 \right], \quad h = 1, \ldots, H, \quad (9)$$
where $\lambda_h > 0$ is a tuning parameter. The caveat is that, although LASSO estimator does not provide guarantees on the FDR control of the selected variables, it serves a useful tool here that simultaneously take into account the sparsity and dependency structures as described in Du et al. (2021). Here we do not need to simultaneously penalize $H$ slices of coefficients to establish $B$, because our variable selection is not implemented through $B$ as we have shown in Figure 1. This is quite different from the traditional model-free variable selection, such as Wu and Li (2011).

Let $S = \{ j : \sum_{h=1}^{H} |\hat{\beta}_{\lambda,h,j}^{(1)}| > 0 \}$ be the subset of variables selected by (9), where $\hat{\beta}_{\lambda,h,j}^{(1)}$ is the $j$th element of $\hat{\beta}_{\lambda,h}$. We then use $D_2$ to obtain the least square estimates $\hat{B}_{2S}$ in (3) for coordinates in the narrowed subset $S$. Denote the estimates from $D_1$ and $D_2$ be $\hat{B}_1$ and $\hat{B}_2$, where

$$
\hat{B}_{2j} = \begin{cases} 
\hat{B}_{2Sj}, & j \in S; \\
0, & \text{otherwise.} 
\end{cases}
$$

Accordingly, the ranking statistics in high-dimensional setting are constructed as

$$
W_j = \frac{\hat{B}_{1j}^\top \hat{B}_{2j}}{s_{1Sj} s_{2Sj}}, \quad j = 1, \ldots, p,
$$

where $s_{kSj}^2 = e_j^\top (n^{-1} \sum_{i=1}^n X_{kSi} X_{kSi}^\top)^{-1} e_j$, $k = 1, 2$, with narrowed subset $S$. The statistics $W_j$ has similar properties to the proposed one in (7), which is (asymptotically) symmetric with mean zero for $j \in A^c$ and is large positive value for $j \in A$ without imposing the
relationship between $Y$ and $X$. Therefore, we propose to choose a threshold $L_+$

$$L_+ = \inf \left\{ t > 0 : \frac{1 + \# \{ j : W_j \leq -t \}}{\# \{ j : W_j \geq t \} \lor 1} \leq \alpha \right\}, \quad (10)$$

and select the active variables by $\hat{A}(L_+) = \{ j : W_j \geq L_+, 1 \leq j \leq p \}$ in high-dimensional setting.

The proposed $L_+$ in (10) shares a similar spirit to Model-X knockoff (Candes et al., 2018) or SDA filter (Du et al., 2021) to obtain an accurate FDR control. However, in the high-dimensional variable selection problems, we usually can not collect enough information on $(Y, X)$, and the knockoff copies are generally not available when $p > n$. Fortunately, the proposed method does not require any prior information on the distribution of $(Y, X)$ or the asymptotic distribution of statistics, and thus it is more suitable for high-dimensional problems.

4 Theoretical results

In this section, we entirely focus on controlling the FDR. Firstly, we need to impose a mild restriction on the response transformation function $f(Y)$.

**Assumption 1** (Response transformation). Function $f(Y)$ satisfies $\mathbb{E} \{ f(Y) \} = 0$ and $\text{var} \{ f(Y) - X^T \beta_0 \} < \infty$.

Assumption 1 ensures that our work differs from most model-based selection methods. It transforms a general model into a multivariate response linear problem to further achieve the
model-free variable selection (Wu and Li, 2011). Our first theorem is a finite sample theory for FDR control.

**Theorem 4.1** (Finite-sample FDR control). Suppose Assumption 1 hold. Assume that the statistics \( W_j, 1 \leq j \leq p \), are well-defined. For any \( \alpha \in (0, 1) \), the FDR of our model-free selection procedure satisfies

\[
\text{FDR} \leq \min_{\epsilon \geq 0} \left\{ \alpha (1 + 4\epsilon) + \Pr \left( \max_{j \in A^c} \Delta_j > \epsilon \right) \right\},
\]

where \( \Delta_j = |\Pr (W_j > 0 \mid |W_j|, W_{-j}) - 1/2| \) and \( W_{-j} = (W_1, \ldots, W_p)^T \setminus W_j \).

Theorem 4.1 holds no matter the unknown relationship between variables \( X \) and response \( Y \). This result can be established using the techniques developed in Barber et al. (2020). The quantity \( \Delta_j \) is interpreted as a measure to investigate the effect of both the asymmetry of \( X_j \) and the dependence between \( W_j \) and \( W_{-j} \) on FDR. Under the asymmetric cases, \( \Delta_j \) can still be expected to be small since \( \beta_{hj}^{(1)} \) and \( \beta_{hj}^{(2)} \) all converges to the normal distribution if \( n \) is not too small. Theorem 4.1 implies that tight control of \( \Delta_j \)’s under asymmetric cases also leads to effective FDR control.

For asymptotic FDR control of the proposed procedure, we need the following technical assumptions, which are not the weakest one, but facilitate the technical proofs in the Supplementary Material. Let \( \|A - \Sigma_S\|_\infty = O_p(a_{np}) \) with \( a_{np} \rightarrow 0 \), where \( A = n^{-1} \sum_{i=1}^n X_{2S_i}X_{2S_i}^T \) and \( \Sigma_S = \mathbb{E}(X_SX_S^T \mid D_1) \). Define \( v_n = \max \{\|\Sigma_S\|_\infty, \|\Sigma_S^{-1}\|_\infty\} \) and \( B_{0S} = \{B_j : j \in S\} \). Denote \( d_n = |A|, q_n = |S| \) and \( q_m = |S \cap A^c| \). Assume that \( q_n \) is uniformly bounded above by some non-random sequence \( \bar{q}_n \).
Assumption 2 (Sure screening property). As $n \to \infty$, $\Pr(A \subseteq S) \to 1$.

Assumption 3 (Moments). Let $\varepsilon = f - B_{0S}^T X_S \in \mathbb{R}^H$. Conditioning on $S$, there exists a positive diverging sequence $K_n$ and a constant $\varpi > 2$ such that

$$\max_{1 \leq h \leq H} \max_{1 \leq i \leq n} \mathbb{E}(\|\Sigma_{S}^{-1} X_{2Si} \varepsilon_{ih}\|_\infty^\varpi) \leq K_n^\varpi,$$

for $i \in D_2$. Assume that as $n \to \infty$, $q_n^{1/\varpi + \gamma + 1/2} K_n / n^{1/2 - \gamma - \varpi^{-1}} \to 0$ for some small $\gamma > 0$.

Assumption 4 (Design matrix). There exist positive constants $\kappa$ and $\bar{\kappa}$ such that

$$\kappa \leq \lim \inf_{n \to \infty} \lambda_{\min}(X_{2S}^T X_{2S}) < \lim \sup_{n \to \infty} \lambda_{\max}(X_{2S}^T X_{2S}) \leq \bar{\kappa},$$

hold with probability one.

Assumption 5 (Estimation accuracy). Assume that $\|\hat{B}_{1j} - B_j\|_\infty = O_p(c_{np})$ uniformly holds for $j \in S$, where $\hat{B}_1$ is an estimator of $B$ from $D_1$, $c_{np} \to 0$ and $1/(\sqrt{n}c_{np}) = O(1)$.

Assumption 6 (Signals). Denote $C_B = \{ j \in A : \|B_j\|_2^2 / \max\{c_{np}^2, \bar{q}_n \log \bar{q}_n / n\} \to \infty \}$. Let $\eta_n := |C_B| \to \infty$ as $(n, p) \to \infty$.

Assumption 7 (Dependence). Let $\rho_{jl}$ denotes the conditional correlation between $W_j$ and $W_l$ given $D_1$. Assume that for each $j$ and some $C > 0$, $\#\{ l \in A^c : |\rho_{jl}| \geq C(\log n)^{-2-\nu} \} \leq r_p$, where $\nu > 0$ is some small constant, and $r_p / \eta_n \to 0$ as $(n, p) \to \infty$.

Remark 1. Assumption 2 has been used in Barber and Candès (2019); Du et al. (2021); Meinshausen et al. (2009) to ensure that $\hat{B}_{2j}$ is unbiased for $j \in S$. Assumptions 3 and 4 are
commonly used in the content of variable selection. The rate $c_{np}$ in Assumption 5 ensures that $\hat{B}_1$ is a reasonable estimator of $B$ from $D_1$. For the LASSO selector, $c_{np} = d_n \sqrt{\log p / n}$ typically satisfies the Assumption 5. Assumption 6 implies that the number of informative covariates with identifiable effect sizes is not too small as $(n, p) \to \infty$. Assumption 7 allows $W_j$ to be correlated with all others but requires that the correlation coefficients need to converge to zero at a log-rate. This condition is similar the weak dependence structure given in Fan et al. (2012).

Our main asymptotically theoretical result for both FDP and FDR control is given by the following theorem.

**Theorem 4.2** (Asymptotic FDR control). Suppose Assumptions 1–7 and hold. For any $\alpha \in (0, 1)$, $c_{np} a_{np} v_n \check{q}_n \sqrt{n (\log \check{q}_n)^{3/2+\gamma}} \to 0$ for a small $\gamma > 0$, the FDP of the MFSDA procedure with threshold $L$ satisfies

$$
\text{FDP}(L) := \frac{\# \{ j : W_j \geq L, j \in \mathcal{A}_c \}}{\# \{ j : W_j \geq L \} \lor 1} \leq \alpha + o_p(1),
$$

and $\lim \sup_{(n,p) \to \infty} \text{FDR} \leq \alpha$.

Theorem 4.2 implies that the variable selection procedure with the data-driven threshold $L$ can control the FDR at target level asymptotically. Further investigations are needed to better understand the condition $c_{np} a_{np} v_n \check{q}_n \sqrt{n (\log \check{q}_n)^{3/2+\gamma}} \to 0$. The conventional estimate of $\| A - \Sigma_S \|_\infty$ indicates that $a_{np} = O_p (v_n \sqrt{\log \check{q}_n / n})$. With $c_{np} = d_n \sqrt{\log p / n}$ of LASSO selector, the condition degenerates to $d_n v_n^2 \check{q}_n / \sqrt{n} \to 0$ if $p$ is of a polynomial rate of $n$. The above condition basically imposes restrictions on the rate of $d_n$, $v_n$ and $\check{q}_n$. Accordingly,
the screening stage on split $\mathcal{D}_1$ must satisfy $\bar{q}_n = o(n^{1/2})$ if we assume that $d_n$ and $\nu_n$ are bounded. Thus a sufficient requirement for the condition in Theorem 4.2 is $\bar{q}_n = o(n^{1/4})$ due to the potential assumption $d_n \leq \bar{q}_n$ holds. This is a reasonable rate in the problem with a diverging number of parameters, such as Fan and Peng (2004) and Wu and Li (2011).

5 Numerical results

We evaluate the performance of our proposed procedure on several simulated datasets and a real-data example under low-dimensional and high-dimensional settings.

5.1 Implantation details

The following methods as benchmarks are considered for comparison with our MFSDA. The first one is the marginal coordinate test in SIR (Cook, 2004), which aims at controlling the error rate for each coordinate. Then we apply the BH procedure (Benjamini and Hochberg, 1995) to the $p$-values to make a global error rate control. This method is implemented using function “dr” and “drop1” in R package *dr*. The second method is the Model-X knockoff (Candes et al., 2018), which also is a model-free and data-driven variable selection procedure as the proposed method. This method is implemented by the function “create.gaussian” in R package *knockoff* using the lasso signed maximum feature important statistics. The two methods are termed as MSIR-BH and MX-Knockoff, respectively.

The FDR level is set as $\alpha = 0.2$. All the simulation results are based on 500 replications. The performance of the proposed MFSDA is evaluated along with above benchmarks through the comparisons of FDR, the true positive rate (TPR), $P_a = \Pr(\mathcal{A} \subseteq \hat{\mathcal{A}}(L))$ and the average
computing time.

To fairly compare with BH, we adopt threshold $L$ in low-dimensional numerical analysis, and thus we are technically controlling a slightly modified version of FDR (Candes et al., 2018). The FDR is nevertheless effectively controlled in all simulations except in extremely low power cases, and even then violations are small. In a high-dimensional case, we can use $L_+$ to achieve more conservative/exact FDR control for MFSDA and MX-Knockoff.

5.2 Simulations

5.2.1 Low-dimensional studies

We generate the covariates $X$ following three distributions: multivariate normal distribution $\mathcal{N}(0, \Sigma)$ with $\Sigma = (\sigma_{ij}) = \rho^{i-j}, 1 \leq i, j \leq p$; multivariate $t(5)$ distribution with covariance $\Sigma$; a mixed distribution which consists of $\{X_j\}_{j=1}^{p/3}$ are from $\mathcal{N}(0, \Sigma[p/3])$, $\{X_j\}_{j=[p/3]+1}^{2p/3}$ are from $\mathcal{N}(0, I[p/3])$, and $\{X_j\}_{j=[2p/3]+1}^{p}$ are i.i.d from a $t(5)$ distribution. The error term $\eta$ is standard normal distribution which is independent with $X$. We fix $(p,p_1) = (20, 10)$. For a scalar $c$, write $c_p = (c, \ldots, c)$ be the $p$-dimensional row vector of $c$’s. Three models have been considered:

- **Scenario 1a**: $Y = \beta_1^T X + 3\eta$, where $\beta = (1_{p_1}, 0_{p-p_1})^T$.

- **Scenario 1b**: $Y = |\beta_1^T X| + \exp(3 + \beta_2^T X) + \eta$, where $\beta_1 = (1_5, 0_{p-5})^T$, $\beta_2 = (0_5, 1_5, 0_{p-p_1})^T$.

- **Scenario 1c**: $Y = \beta_1^T X + (\beta_2^T X + 3)^2 + \exp(\beta_3^T X) + \eta$, where $\beta_1 = (1_3, 0_{p-3})^T$; $\beta_2 = (0_3, 1_3, 0_{p-6})^T$; $\beta_3 = (0_6, 1_4, 0_{p-p_1})^T$. 

19
We consider three response transformation functions \( f_h(Y), h = 1, \ldots, H \), for the proposed MFSDA: (1) the slice indicator function in Li (1991) that \( f_h(Y) = 1 \) if \( Y \) is in the \( h \)th slice and 0 otherwise; (2) The CIRE-type response transformation in Cook and Ni (2006) that \( f_h(Y) = Y \) if \( Y \) is in the \( h \)th slice and 0 otherwise; (3) the normalized polynomial response transformation in Yin and Cook (2002) that \( f_h(Y) = Y^m \) if \( Y \) is in the \( h \)th slice and 0 otherwise. Here we fix \( m = 2 \). We name these three functions as Indicator, CIRE and Poly, respectively.

![Figure 2: FDR(%) of the proposed MFSDA against different H and f_h(Y) under Scenarios 1a—1c when n = 500, \( \rho = 0.5 \), and X is normal distribution. The gray solid line denotes the target FDR level.](image)

In Figure 2, it can be seen that our proposed procedure successfully controls FDR in an acceptable range of the target level no matter the number of working dimension \( H \) and the response transformation functions. The three response transformation functions exhibit similar patterns with FDR control, and we do not address which \( f(Y) \) is the “best” in this paper. In the rest of simulations, we focus on the slice indicator function and set \( H = 4 \) for the proposed MFSDA.

Next, we compare FDR and TPR of the proposed MFSDA under low-dimensional setting
with marginal SIR and MX-Knockoff in Table 1 and Table 2. Table 1 studies how the
proposed MFSDA and the benchmark methods are affected by the covariate distributions
and sample sizes. Table 2 displays the comparisons of covariate correlation for the three
methods. Under all scenarios, the FDR of MFSDA remains at the desired level consistently
and the TPR of MFSDA is higher than MSIR-BH and MX-Knockoff in most cases. The

Table 1: FDR and TPR (%) for three methods against different covariate distributions under Scenarios 1a–1c when \( \rho = 0.5 \).

| \( n \) | Scenario | Method        | \( \text{normal} \) |         | \( t_5 \) |         | \( \text{mixed} \) |         |
|-------|----------|----------------|---------------------|---------|---------|----------------|---------|
| 300   | 1a       | MFSDA          | 20.3                | 62.2    | 20.2    | 62.6          | 21.1    | 82.3 |
|       |          | MSIR-BH        | 11.2                | 59.9    | 10.3    | 53.8          | 12.1    | 86.7 |
|       |          | MX-Knockoff    | 23.2                | 98.6    | 22.5    | 99.1          | 22.8    | 99.7 |
|       | 1b       | MFSDA          | 20.8                | 68.0    | 20.4    | 56.6          | 21.2    | 79.3 |
|       |          | MSIR-BH        | 12.2                | 69.6    | 13.3    | 54.3          | 11.7    | 75.8 |
|       |          | MX-Knockoff    | 12.7                | 32.4    | 34.6    | 44.3          | 16.2    | 56.2 |
|       | 1c       | MFSDA          | 21.9                | 63.4    | 19.0    | 50.9          | 19.9    | 77.1 |
|       |          | MSIR-BH        | 11.8                | 60.1    | 13.2    | 48.8          | 11.3    | 74.9 |
|       |          | MX-Knockoff    | 14.0                | 39.6    | 35.8    | 45.5          | 19.1    | 78.2 |
| 400   | 1a       | MFSDA          | 21.3                | 79.2    | 21.9    | 77.2          | 21.7    | 92.4 |
|       |          | MSIR-BH        | 11.1                | 81.1    | 10.3    | 73.8          | 10.9    | 95.0 |
|       |          | MX-Knockoff    | 22.1                | 99.7    | 22.5    | 99.5          | 23.3    | 99.9 |
|       | 1b       | MFSDA          | 20.9                | 80.6    | 20.4    | 65.7          | 21.0    | 84.7 |
|       |          | MSIR-BH        | 10.7                | 78.9    | 12.6    | 62.9          | 11.3    | 82.8 |
|       |          | MX-Knockoff    | 13.7                | 36.6    | 36.0    | 48.3          | 15.7    | 57.0 |
|       | 1c       | MFSDA          | 19.9                | 72.2    | 19.2    | 60.8          | 20.9    | 82.1 |
|       |          | MSIR-BH        | 10.9                | 68.9    | 13.4    | 60.6          | 10.9    | 78.4 |
|       |          | MX-Knockoff    | 14.4                | 42.8    | 37.0    | 50.0          | 18.1    | 80.1 |
| 500   | 1a       | MFSDA          | 20.7                | 87.1    | 22.0    | 84.4          | 22.2    | 96.4 |
|       |          | MSIR-BH        | 10.4                | 91.4    | 10.1    | 84.6          | 10.8    | 98.2 |
|       |          | MX-Knockoff    | 21.0                | 99.9    | 22.4    | 100.0         | 22.8    | 100.0 |
|       | 1b       | MFSDA          | 21.8                | 85.6    | 19.1    | 70.4          | 21.5    | 89.7 |
|       |          | MSIR-BH        | 10.0                | 83.1    | 13.1    | 68.3          | 10.5    | 88.9 |
|       |          | MX-Knockoff    | 13.3                | 39.9    | 39.4    | 50.6          | 16.3    | 58.9 |
|       | 1c       | MFSDA          | 21.3                | 78.5    | 20.9    | 68.9          | 20.9    | 84.2 |
|       |          | MSIR-BH        | 10.5                | 74.8    | 14.1    | 68.1          | 11.2    | 80.9 |
|       |          | MX-Knockoff    | 13.6                | 46.0    | 38.8    | 52.4          | 19.1    | 82.2 |
Table 2: FDR and TPR (%) for three methods against different correlation $\rho$ under Scenarios 1a–1c when $n = 500$ and $X$ is from normal distribution.

| Scenario | Method      | $\rho = 0.2$ FDR | $\rho = 0.2$ TPR | $\rho = 0.5$ FDR | $\rho = 0.5$ TPR | $\rho = 0.8$ FDR | $\rho = 0.8$ TPR |
|----------|-------------|------------------|------------------|------------------|------------------|------------------|------------------|
| 1a       | MFSDA       | 22.6             | 99.1             | 20.7             | 87.1             | 14.6             | 27.7             |
|          | MSIR-BH     | 11.0             | 99.9             | 10.4             | 91.4             | 14.0             | 10.0             |
|          | MX-Knockoff | 22.8             | 100.0            | 21.0             | 99.9             | 20.6             | 95.6             |
| 1b       | MFSDA       | 22.9             | 99.5             | 21.8             | 85.6             | 21.6             | 30.2             |
|          | MSIR-BH     | 11.2             | 99.8             | 10.0             | 83.1             | 12.0             | 16.5             |
|          | MX-Knockoff | 15.9             | 54.9             | 13.3             | 39.9             | 10.1             | 20.7             |
| 1c       | MFSDA       | 21.6             | 89.9             | 21.3             | 78.5             | 20.2             | 34.9             |
|          | MSIR-BH     | 11.2             | 85.9             | 10.5             | 74.8             | 11.7             | 18.4             |
|          | MX-Knockoff | 17.5             | 69.3             | 13.6             | 46.0             | 11.1             | 22.6             |

The following discussion also can be made.

(a) *MFSDA vs MSIR-BH.* The marginal SIR (Cook, 2004) with BH procedure (Benjamini and Hochberg, 1995) controls the FDR in all settings but can be very conservative. This is because that the BH procedure controls the FDR at level $\alpha p / p$ in a low-dimensional setting. By contrast, MFSDA performs accurate FDR control and better TPR in linear and nonlinear models (Candes et al., 2018). It seems that the power of the proposed MFSDA is slightly lower than MSIR-BH in very few cases since MFSDA involves data splitting to construct the statistic symmetry property. But as the sample size $n$ increases, our method will be more effective than others in a low-dimensional setting.

(b) *MFSDA vs MX-Knockoff.* MFSDA and MX-Knockoff both are model-free and data-driven variable selection methods, while MFSDA controls the FDR more accurately near the desired level as $n$ increases. By contrast, MX-Knockoff fails to control the FDR for $t(5)$ distribution in Column 2 of Table 1. Table 1 also shows that MX-Knockoff
works not well as our MFSDA in nonlinear cases. Table 2 implies that the MFSDA controls the FDR at the nominal level robustly, but the MX-Knockoff exhibits more conservative FDR and less TPR when correlation becomes higher across all scenarios.

5.2.2 High-dimensional studies

In high-dimensional settings, we consider the following benchmarks. The competitor MSIR-BH in low-dimension does not work in high-dimension since the p-value cannot directly be obtained. Thus, the sample-splitting method (Wasserman and Roeder, 2009), which conducts data screening using LASSO and then applies BH to the p-values calculated by marginal SIR (Cook, 2004). Since the commonly used Akaike information criterion such as in Du et al. (2021) causes inaccurate model deviance after slicing responses, cross-validation criterion is conducted to choose an overfitted model in screening stage. The MX-Knockoff is conducted by function `create.second.order` in R package `knockoff` to approximate an accurate precision matrix in high-dimensional setting (Candes et al., 2018). The next method is named as MFSDA-DB (debiased) method, which extends the least square solution in (3) to regularized version with $L_1$ penalty in (9) and makes a bias correction with R package `selectiveInference`. The fourth one is the marginal independence SIR proposed in Yu et al. (2016a). We choose two model sizes $\lfloor cn/\log(n) \rfloor$, $c = 0.05, 0.5$, as two simple competitors and denote them as IM-SIR1 and IM-SIR2, respectively.

We consider the following models when $p > n$.

- **Scenario 2a**: $Y = \exp(5 + \beta^\top X) + \eta$, where $\beta = (1_p, 0_{p-1})^\top$.

- **Scenario 2b**: $Y = 2\beta_1^\top X + 3 \exp(\beta_2^\top X) + \eta$, where $\beta_1 = (1_5, 0_{5-p})^\top, \beta_2 = (0_5, 1_5, 0_{p-1})^\top$. 
Scenario 2c: \( Y = \beta_1^\top X + |\beta_2^\top X + 5| + \exp(\beta_3^\top X) + \eta \), where \( \beta_1 = (1_3, 0_{p-3})^\top; \)
\( \beta_2 = (0_3, 1_3, 0_{p-6})^\top; \beta_3 = (0_6, 1_4, 0_{p-p_1})^\top. \)

Table 3: FDR, TPR, \( P_a(\%) \) and computing time (in second) for several methods against different \( X \) distributions under Scenarios 2a–2c when \((n, p, p_1, \rho) = (500, 1000, 10, 0.5)\).

| Scenario | Method      | normal |           | mixed |           |
|----------|-------------|--------|-----------|--------|-----------|
|          |             | FDR    | TPR       | \( P_a \) | time     | FDR    | TPR       | \( P_a \) | time     |
| 2a       | MFSDA       | 18.3   | 98.7      | 90.2   | 12.9      | 17.5   | 98.3      | 86.6     | 14.1     |
|          | MFSDA-DB    | 17.1   | 57.9      | 7.0    | 75.1      | 17.0   | 77.3      | 19.8     | 65.5     |
|          | MSIR-BH     | 4.5    | 32.8      | 0.0    | 11.4      | 4.2    | 32.9      | 0.0      | 12.6     |
|          | IM-SIR1     | 0.0    | 40.0      | 0.0    | 26.3      | 0.0    | 40.0      | 0.0      | 34.8     |
|          | IM-SIR2     | 75.0   | 100.0     | 100.0  | 26.3      | 75.0   | 100.0     | 100.0    | 34.5     |
|          | MX-Knockoff | 6.5    | 4.3       | 0.0    | 31.8      | 26.4   | 9.2       | 0.0      | 31.6     |
| 2b       | MFSDA       | 17.0   | 94.7      | 62.0   | 12.8      | 17.5   | 94.6      | 58.8     | 14.7     |
|          | MFSDA-DB    | 15.6   | 71.4      | 4.2    | 74.8      | 16.9   | 80.5      | 14.4     | 75.3     |
|          | MSIR-BH     | 5.3    | 34.9      | 0.0    | 11.5      | 4.6    | 34.6      | 0.0      | 13.3     |
|          | IM-SIR1     | 0.0    | 40.0      | 0.0    | 26.3      | 0.0    | 40.0      | 0.0      | 27.3     |
|          | IM-SIR2     | 75.0   | 100.0     | 100.0  | 26.3      | 75.0   | 100.0     | 100.0    | 27.1     |
|          | MX-Knockoff | 10.3   | 11.6      | 0.0    | 31.9      | 29.0   | 19.8      | 0.0      | 31.4     |
| 2c       | MFSDA       | 17.9   | 92.7      | 50.6   | 12.4      | 19.2   | 92.8      | 49.8     | 22.8     |
|          | MFSDA-DB    | 17.0   | 62.1      | 8.0    | 102.3     | 17.4   | 73.2      | 6.0      | 126.8    |
|          | MSIR-BH     | 6.5    | 33.4      | 0.0    | 12.2      | 6.3    | 31.1      | 0.0      | 21.8     |
|          | IM-SIR1     | 0.0    | 40.0      | 0.0    | 22.7      | 0.0    | 40.0      | 0.0      | 36.2     |
|          | IM-SIR2     | 75.0   | 100.0     | 100.0  | 22.9      | 75.0   | 100.0     | 100.0    | 35.2     |
|          | MX-Knockoff | 12.1   | 12.2      | 0.0    | 28.4      | 28.8   | 19.1      | 0.0      | 42.0     |

Table 3 presents the comparison results for different covariate distributions to investigate the error rate control and detection power under high-dimensional settings. The FDRs of the proposed MFSDA are approximately controlled at the target level of \( \alpha \) with higher power, which is in line with our theory. A similar analysis also can be found in the Supplemental Material Table S1 with a larger sample size. Table 3 and Table S1 further demonstrate that the MFSDA is able to detect all active variables when \( n \) is large. As we can expect although MFSDA involves data splitting which may lose some data information, its power can still
be higher since the feature screening step significantly increases the signal-to-noise ratio. Besides, our method owns a lower computing time since we do not construct the asymptotic distribution for each dimension in marginal coordinate test (Cook, 2004) and generate the knockoff copies in Model-X knockoff (Candes et al., 2018). We have the following expositions.

(a) **MFSDA vs MFSDA-DB.** MFSDA-DB method is similar to RESS in Zou et al. (2020), which was developed for independent tests. The FDR of the MFSDA-DB method controls pretty well as the proposed MFSDA but it performs a lower power than MFSDA since MFSDA-DB uses bias correction instead of the screening stage which may not boost the signal-to-noise ratio. We know that the MFSDA-DB method is an extension of our low-dimensional procedure with a debiased lasso estimate but it needs to estimate the precision matrix which involves much more computational cost.

(b) **MFSDA vs MSIR-BH.** MSIR-BH method achieves a quite conservative FDR control than MFSDA. It adopts the sample splitting (Meinshausen et al., 2009) but they only construct the test statistics on $D_2$ which suffers from a serious power loss.

(c) **MFSDA vs IM-SIR.** The hard thresholding IM-SIR methods can detect more active variables only when model size $\lfloor cn/\log(n)\rfloor$ is greater than $p_1$. Table 3 implies that the hard-thresholding method can not control the FDR and have large power with user-specified model size.

(d) **MFSDA vs MX-Knockoff.** MX-Knockoff offers a variable selection solution without making any modeling assumptions in high-dimensional situation. But MX-Knockoff fails to control the FDR for mixed distribution which is consistent with the low-
dimensional conclusion. By contrast, MFSDA controls the FDR more accurately under mixed distribution.

![Graph showing FDR and TPR (%) curves against different covariate dimension $p$, different correlation $\rho$ and different signal number $p_1$ under Scenario 2c when $n = 500$ and $X$ generates from the mixed distribution. Left panel: $(p_1, \rho) = (10, 0.5)$; Middle panel: $(p, p_1) = (1000, 10)$; Right panel: $(p, \rho) = (1000, 0.5)$. The gray solid line denotes the target FDR level.]

Figure 3: FDR and TPR (%) curves against different covariate dimension $p$, different correlation $\rho$ and different signal number $p_1$ under Scenario 2c when $n = 500$ and $X$ generates from the mixed distribution. Left panel: $(p_1, \rho) = (10, 0.5)$; Middle panel: $(p, p_1) = (1000, 10)$; Right panel: $(p, \rho) = (1000, 0.5)$. The gray solid line denotes the target FDR level.

To further investigate the efficiency of our MFSDA procedure in high-dimensional setting against different covariate dimension $p$, covariate correlation $\rho$, and signal number $p_1$, the corresponding FDR and TPR are reported in Figure 3. The FDR of MFSDA is robust in an acceptable range of the target level no matter the dimension, correlation and signal number varied. Our MFSDA always achieves the most powerful TPR than other competitors except for IM-SIR2 since IM-SIR2 can not make a fair error rate control. The practical performance
between MFSDA and MX-Knockoff is quite different. It is seen that although controlling
the FDR below the target level with larger $p_1$ in Figure 3, MX-Knockoff suffers from a larger
power loss as signal number $p_1$ and correlation $\rho$ increases. Additional numerical results
with normal covariate distribution can be founded in the Supplementary Material Figure S1.
Figure S1 shows that the FDR of MX-Knockoff controls quite well as MFSDA under normal
distribution with different $p$, $\rho$, and $p_1$, but there also exists a power gap to our proposed
MFSDA.

5.3 Real data implementation

In this section, we apply our proposed MFSDA procedure to the children cancer data
for classifying small round blue cell tumors (SRBCT), which has been analyzed by Khan
et al. (2001) and Yu et al. (2016a). The SRBCT dataset looks at classifying four classes
of different childhood tumors sharing similar visual features during routine histology. Data
were collected from 83 tumor samples and the expression measurements on 2308 genes for
each sample are provided. In this dataset, we want to investigate the performance of the
FDR control between our MFSDA procedure and other existing methods.

We first present the scatterplots of the ranking statistics $W_j$ for MFSDA and MX-Knockoff
in Figure 4. As we expect, genes with larger $W_j$’s are possible to be selected as influential
genes (red dots), while the unselected genes (black dots) are roughly and symmetrically
distributed around zero line although there are plenty of $W_j$’s are exactly zero.

Next, we introduce some simulated variables as inactive genes to further investigate the
performance of our proposed method. In detail, 1000 noise variables $Z_1$ are from $\mathcal{N}(0,1)$
and another 1000 noise variables $Z_2$ are from $t(3)$, which are all independently and randomly generated. We name this dataset as SRBCT&Noise, and our main goal is to identify the important genes from the whole 4328 variables. Since $Z_1$ and $Z_2$ are actually inactive by construction, they can be viewed as truly false discoveries. The scatterplot of $W_j$’s and the number of discovery genes are shown in Figure 4 and Table 4. Yu et al. (2016a) analyzed that an average of 9 genes out of the 2308 total genes have been selected, which is very close to the number of variables selected by our method. We observed that MFSDA and MX-Knockoff identified the same number of important genes when noise variables were not added, but MX-Knockoff mistakenly selected a small number of truly noise variables, i.e., the added noise variables. Besides, the proposed MFSDA possesses a relatively smaller computing time.
Table 4: Number of discovered (ND) genes and computing times in SRBCT and SR-BCT&Noise datasets. Values in parentheses are the number of discovered noise variables.

| Method       | SRBCT  | SRBCT&Noise |
|--------------|--------|-------------|
|              | ND     | time(s)     | ND     | time(s)     |
| MFSDA        | 8      | 1.52        | 12(0)  | 1.92        |
| MSIR-BH      | 5      | 0.50        | 5(0)   | 0.72        |
| IM-SIR(c=0.05) | 1    | 130.20      | 1(0)   | 1514.23     |
| IM-SIR(c=1)  | 18     | 128.31      | 18(0)  | 1516.59     |
| MX-Knockoff  | 8      | 49.82       | 46(7)  | 240.20      |

than IM-SIR and MX-Knockoff.

6 Discussion

Identifying the truly contributory variables is a very important task in statistical analysis. In this paper, we proposed a novel model-free variable selection procedure with a data-driven threshold in sufficient dimension reduction framework for a family of inverse regression methods via data splitting. The proposed MFSDA is computationally efficient in high-dimensional setting. Theoretical and numerical results show that the proposed MFSDA can asymptotically control the FDR at the target level.

It is interesting to adopt multiple data splitting in MFSDA to improve the stability and robustness but it maybe requires intensive computations. One can extend the idea of MFSDA for controlling other types of error rates such as per-family error rate or marginal false discovery rate. Equal size data splitting is used in our simulation and theory for simplicity. We have checked that larger sample size for $D_1$ will improve TPR to a certain extent. Benefiting from the use of the response transformation, the proposed method can work with many types of response, such as continuous, discrete, or categorical data. Our method may also be
able to dispose of the order determination problem in many situations which merits further
investigation, such as Luo and Li (2016) if the importance of $H$ slices has been determined
in advance with a diverging number of $H$.

**Appendix**

This Appendix contains two parts but we only present here the first part which gives
some key lemmas and the succinct proof of theorems. We only consider the proofs under
high-dimensional setting and the low-dimensional version can be easily verified. Additional
lemmas used in Appendix with their proofs and some additional simulation results, can be
found in the Supplemental Material.

For any vector $\mathbf{v} \in \mathbb{R}^H$, we have

$$(\widehat{\mathbf{B}}_{2j} - \mathbf{B}_j)^\top \mathbf{v} = e_j^\top \left( \frac{1}{n} \sum_{i=1}^n \mathbf{X}_{2S_i} \mathbf{X}_{2S_i}^\top \right)^{-1} \frac{1}{n} \sum_{i=1}^n \mathbf{X}_{2S_i} (\mathbf{f}_{2i} - \mathbf{B}_0^\top \mathbf{X}_{2S_i})^\top \mathbf{v}$$

$$= e_j^\top \left( \frac{1}{n} \sum_{i=1}^n \mathbf{X}_{2S_i} \mathbf{X}_{2S_i}^\top \right)^{-1} \frac{1}{n} \sum_{i=1}^n \mathbf{X}_{2S_i} \varepsilon_i^\top \mathbf{v}.$$ 

Here $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_H)^\top = \mathbf{f} - \mathbf{B}_0 S \mathbf{X}_S \in \mathbb{R}^H$ with $\mathbb{E}(\mathbf{X}_S \varepsilon^\top) = 0$. However, $\varepsilon$ is not inde-
pendent with $\mathbf{X}_S$ nor $\mathbb{E}(\varepsilon \mid \mathbf{X}_S) = 0$ due to the use of response transformation, which are
quite different from the usual setting. To establish the FDR control result of our proposed
procedure, we construct a modified statistic

$$\widetilde{W}_j = \frac{\widehat{\mathbf{B}}_{1j}^\top \widehat{\mathbf{B}}_{2j}}{s_{1j} s_{2j}}, \quad j = 1, \ldots, p,$$
where $\tilde{B}_{2j} = e_j^\top n^{-1} \Sigma_{S}^{-1} \sum_{i=1}^{n} X_{2S_i} e_{2i}$, $\tilde{s}_{2S_j}^2 = n^{-1} e_j^\top \Sigma_{S}^{-1} e_j$ for $j \in S$, and $\tilde{W}_j = 0$ otherwise.

Let $G(t) = q_{0n}^{-1} \sum_{j \in A^c} \Pr(\tilde{W}_j \geq t \mid D_1)$, $G_-(t) = q_{0n}^{-1} \sum_{j \in A^c} \Pr(\tilde{W}_j \leq -t \mid D_1)$ and $G^{-1}(y) = \inf\{t \geq 0 : G(t) \leq y\}$ for $0 \leq y \leq 1$.

Before we present the proofs of main theorems, we first state two key lemmas. The first lemma characterizes the closeness between $G(t)$ and $G_-(t)$, which plays an important role in the proof.

**Lemma A.1** (Symmetry property). *Suppose Assumptions 1–4 hold. For any $0 \leq t \leq G^{-1}(\alpha n/q_{0n})$*

$$\frac{G(t)}{G_-(t)} - 1 \to 0.$$  

The next lemma establishes the uniform convergence of $\sum_{j \in A^c} \mathbb{I}(\tilde{W}_j \geq t)/(q_{0n} G(t))$.

**Lemma A.2** (Uniform consistency). *Suppose Assumptions 1–4 and 7 hold. Then*

$$\sup_{0 \leq t \leq G^{-1}(\alpha n/q_{0n})} \left| \frac{\sum_{j \in A^c} \mathbb{I}(\tilde{W}_j \geq t)}{q_{0n} G(t)} - 1 \right| = o_p(1), \quad (A.1)$$

$$\sup_{0 \leq t \leq G^{-1}(\alpha n/q_{0n})} \left| \frac{\sum_{j \in A^c} \mathbb{I}(\tilde{W}_j \leq -t)}{q_{0n} G_-(t)} - 1 \right| = o_p(1). \quad (A.2)$$

We also need to show that the following two lemmas.

**Lemma A.3.** *Suppose Assumptions 1–5 hold. We have*

$$W_j - \tilde{W}_j = o_p(c_{np} a_{np} v_n q_n \sqrt{n \log q_n}).$$
Lemma A.4. Suppose Assumptions 1–5 hold with $c_n a_n v_n \bar{q}_n \sqrt{n (\log \bar{q}_n)^{3/2+\gamma}} \to 0$ for a small $\gamma > 0$. Then for any $M > 0$, we have

$$\sup_{M \leq t \leq (\alpha \eta \sqrt{n})/q_0} \left| \frac{\sum_{j \in A^c} \mathbb{I}(\tilde{W}_j \geq t)}{\sum_{j \in A^c} \mathbb{I}(W_j \geq t)} - 1 \right| = o_p(1),$$

and

$$\sup_{M \leq t \leq (\alpha \eta \sqrt{n})/q_0} \left| \frac{\sum_{j \in A^c} \mathbb{I}(\tilde{W}_j \leq -t)}{\sum_{j \in A^c} \mathbb{I}(W_j \leq -t)} - 1 \right| = o_p(1).$$

Proof of Theorem 4.1

We prove the finite-sample FDR control with $L_+$. The result for $L$ can be obtained similarly. The proof technique of this result has been extensively used in Barber et al. (2020) and Du et al. (2021). Fix $\epsilon > 0$ and for any threshold $t > 0$, define

$$R_\epsilon(t) = \frac{\sum_{j \in A^c} \mathbb{I}(W_j \geq t, \Delta_j \leq \epsilon)}{1 + \sum_{j \in A^c} \mathbb{I}(W_j \leq t)}.$$

Define an event that $\mathcal{A} = \{\Delta \equiv \max_{j \in A^c} \Delta_j \leq \epsilon\}$. Furthermore, consider a thresholding rule $L = T(W)$ that maps statistics $W$ to a threshold $L \geq 0$. Define

$$L_j = T(W_1, \ldots, W_{j-1}, |W_j|, W_{j+1}, \ldots, W_p) \geq 0, \ j = 1, \ldots, p.$$

Then for the proposed MFSDA with threshold $L_+$, we can write

$$\frac{\sum_{j \in A^c} \mathbb{I}(W_j \geq L_+, \Delta_j \leq \epsilon)}{1 \vee \sum_{j \in A^c} \mathbb{I}(W_j \geq L_+)} = \frac{1 + \sum_{j \in A^c} \mathbb{I}(W_j \leq -L_+)}{1 \vee \sum_{j \in A^c} \mathbb{I}(W_j \geq L_+)} \times \frac{\sum_{j \in A^c} \mathbb{I}(W_j \geq L_+, \Delta_j \leq \epsilon)}{1 + \sum_{j \in A^c} \mathbb{I}(W_j \leq -L_+)} \leq \alpha \times R_\epsilon(L_+).$$
Next we derive an upper bound for $\mathbb{E} \{ R_\epsilon (L_+) \}$. Note that

$$
\mathbb{E} \{ R_\epsilon (L_+) \} = \mathbb{E} \left\{ \frac{\sum_{j \in A^c} \mathbb{I} (W_j \geq L_j, \Delta_j \leq \epsilon)}{1 + \sum_{j \in A^c, k \neq j} \mathbb{I} (W_k \leq -L_j)} \right\}
$$

$$
= \sum_{j \in A^c} \mathbb{E} \left\{ \frac{\mathbb{I} (W_j \geq L_j, \Delta_j \leq \epsilon)}{1 + \sum_{k \in A^c, k \neq j} \mathbb{I} (W_k \leq -L_j)} \right\}
$$

$$
= \sum_{j \in A^c} \mathbb{E} \left\{ \frac{\mathbb{I} (W_j \geq L_j, \Delta_j \leq \epsilon)}{1 + \sum_{k \in A^c, k \neq j} \mathbb{I} (W_k \leq -L_j)} \right\} \mathbb{P} (W_j > 0 \mid |W_j|, W_{-j}) \mathbb{I} (|W_j| \geq L_j, \Delta_j \leq \epsilon)
$$

where the last step holds since, after conditioning on $(|W_j|, W_{-j})$, the only unknown quantity is the sign of $W_j$. Recall the definition of $\Delta_j = |\mathbb{P} (W_j > 0 \mid |W_j|, W_{-j}) - 1/2|$, we have

$$
\mathbb{P} (W_j > 0 \mid |W_j|, W_{-j}) \leq 1/2 + \Delta_j. \quad \text{Thus}
$$

$$
\mathbb{E} \{ R_\epsilon (L_+) \} \leq \sum_{j \in A^c} \mathbb{E} \left\{ \frac{(1/2 + \Delta_j) \mathbb{I} (|W_j| \geq L_j, \Delta_j \leq \epsilon)}{1 + \sum_{k \in A^c, k \neq j} \mathbb{I} (W_k \leq -L_j)} \right\}
$$

$$
= \left( \frac{1}{2} + \Delta_j \right) \sum_{j \in A^c} \mathbb{E} \left\{ \frac{\mathbb{I} (W_j \geq L_j, \Delta_j \leq \epsilon)}{1 + \sum_{k \in A^c, k \neq j} \mathbb{I} (W_k \leq -L_j)} \right\} + \sum_{j \in A^c} \mathbb{E} \left\{ \frac{\mathbb{I} (W_j \leq -L_j, \Delta_j \leq \epsilon)}{1 + \sum_{k \in A^c, k \neq j} \mathbb{I} (W_k \leq -L_j)} \right\}
$$

$$
\leq \left( \frac{1}{2} + \Delta_j \right) \sum_{j \in A^c} \mathbb{E} \left\{ \frac{\mathbb{I} (W_j \geq L_j, \Delta_j \leq \epsilon)}{1 + \sum_{k \in A^c, k \neq j} \mathbb{I} (W_k \leq -L_j)} \right\} + \sum_{j \in A^c} \mathbb{E} \left\{ \frac{\mathbb{I} (W_j \leq -L_j)}{1 + \sum_{k \in A^c, k \neq j} \mathbb{I} (W_k \leq -L_j)} \right\}
$$

$$
= \left( \frac{1}{2} + \Delta_j \right) \mathbb{E} \{ R_\epsilon (L_+) \} + \sum_{j \in A^c} \mathbb{E} \left\{ \frac{\mathbb{I} (W_j \leq -L_j)}{1 + \sum_{k \in A^c, k \neq j} \mathbb{I} (W_k \leq -L_j)} \right\}
$$

(A.3)

The last expression summation in (A.3) can be simplified. If for all null $j$, $W_j > -L_j$, then
the summation is zero. Otherwise,

$$\sum_{j \in A^c} \mathbb{E} \left\{ \frac{1}{1 + \sum_{k \in A^c, k \neq j} \mathbb{I}(W_k \leq -L_j)} \mathbb{I}(W_j \leq -L_j) \right\} = \sum_{j \in A^c} \mathbb{E} \left\{ \frac{1}{1 + \sum_{k \in A^c, k \neq j} \mathbb{I}(W_k \leq -L_k)} \mathbb{I}(W_j \leq -L_j) \right\} = 1,$$

where the first equality holds by the fact: for any $j, k$, if $W_j \leq \min(L_j, L_k)$ and $W_k \leq -\min(L_j, L_k)$, then $L_j + L_k = L_k$; see Barber et al. Barber et al. (2020) for details.

Accordingly, we have $\mathbb{E} \{ R_\epsilon(L_+) \} \leq (1/2 + \epsilon) [\mathbb{E} \{ R_\epsilon(L_+) \} + 1]$. As a result

$$\mathbb{E} \{ R_\epsilon(L_+) \} \leq \frac{1/2 + \epsilon}{1/2 - \epsilon} \leq 1 + 4\epsilon.$$

Consequently, we can naturally get the conclusions in Theorem 4.1.

\[ \square \]

**Proof of Theorem 4.2**

By the definition of FDP, our test is equivalent to select the $j$th variable if $W_j \geq L$, where

$$L = \inf \left\{ t \geq 0 : \sum_j \mathbb{I}(W_j \leq -t) \leq \alpha \left( 1 + \sum_j \mathbb{I}(W_j \geq t) \right) \right\}.$$

We need to establish an asymptotic bound for this $L$ so that Lemmas A.1–A.4 can be applied. Let $t^* = G_-(\alpha \eta_n/q_0n)$. By Lemmas A.2 and A.4, it follows that

$$\frac{\alpha \eta_n}{q_0n} = G_-(t^*) = \frac{1}{q_0n} \sum_{j \in A^c} \mathbb{I}(\tilde{W}_j < -t^*) \{1 + o(1)\} = \frac{1}{q_0n} \sum_{j \in A^c} \mathbb{I}(W_j < -t^*) \{1 + o(1)\},$$

with $c_{np}a_{np}n \sqrt{n} \log(q_0n)^{3/2+\gamma} \to 0$. On the other hand, for any $j \in \mathcal{C}_B$, where $\mathcal{C}_B$ is defined
in Assumption 6 and \( \eta_n = |C_B| \), we can show that \( \Pr(W_j < t^*, j \in C_B) \rightarrow 0 \). In fact, it is straightforward to see that

\[
\Pr(W_j < t^*, \text{ for some } j \in C_B) \\
\leq \eta_n \Pr \left( \hat{B}_{1j}^\top \hat{B}_{2j} / (s_1 S_j s_2 S_j) - B_j^\top B_j / (s_1 S_j s_2 S_j) < t^* - B_j^\top B_j / (s_1 S_j s_2 S_j) \right) \\
\leq \eta_n \Pr \left( |B_j^\top (\hat{B}_{1j} - B_j + \hat{B}_{2j} - B_j)| + |(\hat{B}_{1j} - B_j)^\top (\hat{B}_{2j} - B_j)| > B_j^\top B_j - t^* s_1 S_j s_2 S_j \right).
\]

Denote \( d_j = \|B_j\|_2^2 - t^* s_1 S_j s_2 S_j \). Under Assumption 6, it follows that \( d_j = \|B_j\|_2^2 \{1 + o(1)\} \).

We then get

\[
\Pr \left( |B_j^\top (\hat{B}_{1j} - B_j + \hat{B}_{2j} - B_j)| + |(\hat{B}_{1j} - B_j)^\top (\hat{B}_{2j} - B_j)| > d_j \right) \\
\leq \Pr \left( |B_j^\top (\hat{B}_{1j} - B_j + \hat{B}_{2j} - B_j)| > d_j / 2 \right) + \Pr \left( |(\hat{B}_{1j} - B_j)^\top (\hat{B}_{2j} - B_j)| > d_j / 2 \right) \\
= : \Lambda_1 + \Lambda_2.
\]

It follows that

\[
\Lambda_1 \leq \Pr \left( \|\hat{B}_{1j} - B_j\|_2 > d_j / (4\|B_j\|_2) \right) + \Pr \left( \|\hat{B}_{2j} - B_j\|_2 > d_j / (4\|B_j\|_2) \right) \rightarrow 0,
\]

\[
\Lambda_2 \leq \Pr \left( \|\hat{B}_{1j} - B_j\|_2 > c_{np} \right) + \Pr \left( \|\hat{B}_{2j} - B_j\|_2 > C \sqrt{q_n \log q_n / n} \right) \rightarrow 0,
\]

the above result follows from Assumption 5 and the Lemma S.4 stated in Supplementary Material.
Thus, we get \( \Pr(\sum_j I(W_j > t^*) \geq \eta_n) \to 1 \). Furthermore, we can conclude that

\[
\sum_j I(W_j < -t^*) \lesssim \alpha \eta_n \leq \alpha \sum_j I(W_j > t^*).
\]

Then, we get an upper bound for \( L \), that is \( L \lesssim t^* \). This implies that the proposed method can detect at least \( \sum_j I(W_j > t^*) \) signals.

Therefore, by Lemmas A.1, A.2 and A.4, we get

\[
\frac{\sum_{j \in A^c} I(W_j \geq L)}{\sum_{j \in A^c} I(W_j \leq -L)} - 1 \xrightarrow{p} 0. \quad \text{(A.4)}
\]

Write

\[
\text{FDP} = \frac{\sum_{j \in A^c} I(W_j \geq L)}{1 \vee \sum_j I(W_j \geq L)} = \frac{\sum_j I(W_j \leq -L)}{1 \vee \sum_j I(W_j \geq L)} \times \frac{\sum_{j \in A^c} I(W_j \geq L)}{\sum_j I(W_j \leq -L)} \leq \alpha \times R(L).
\]

Note that \( R(L) \leq \sum_{j \in A^c} I(W_j \geq L)/\sum_{j \in A^c} I(W_j \leq -L) \), and thus \( \limsup_{n \to \infty} \text{FDP} \leq \alpha \) in probability by (A.4). Thus the first assertion in Theorem 4.2 is proved.

Further, for any \( \epsilon > 0 \), we have

\[
\text{FDR} \leq (1 + \epsilon) \alpha R(L) + \Pr(\text{FDP} \geq (1 + \epsilon) \alpha R(L)),
\]

from which the second part of this theorem is proved.
References

Barber, R. F. and Candès, E. J. (2015), “Controlling the false discovery rate via knockoffs,” *The Annals of Statistics*, 43, 2055–2085.

— (2019), “A knockoff filter for high-dimensional selective inference,” *The Annals of Statistics*, 47, 2504–2537.

Barber, R. F., Candès, E. J., and Samworth, R. J. (2020), “Robust inference with knockoffs,” *The Annals of Statistics*, 48, 1409–1431.

Benjamini, Y. and Hochberg, Y. (1995), “Controlling the false discovery rate: a practical and powerful approach to multiple testing,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 57, 289–300.

Bondell, H. D. and Li, L. (2009), “Shrinkage inverse regression estimation for model-free variable selection,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71, 287–299.

Cai, T. T. and Liu, W. (2016), “Large-scale multiple testing of correlations,” *Journal of the American Statistical Association*, 111, 229–240.

Candès, E., Fan, Y., Janson, L., and Lv, J. (2018), “Panning for gold: ‘model-X’ knockoffs for high dimensional controlled variable selection,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80, 551–577.

Chen, X., Zou, C., and Cook, R. D. (2010), “Coordinate-independent sparse sufficient dimension reduction and variable selection,” *The Annals of Statistics*, 38, 3696–3723.
Cook, R. D. (2004), “Testing predictor contributions in sufficient dimension reduction,” *The Annals of Statistics*, 32, 1062–1092.

Cook, R. D. and Ni, L. (2006), “Using intraslice covariances for improved estimation of the central subspace in regression,” *Biometrika*, 93, 65–74.

Cook, R. D. and Weisberg, S. (1991), “Sliced inverse regression for dimension reduction: Comment,” *Journal of the American Statistical Association*, 86, 328–332.

Dong, Y. (2021), “A brief review of linear sufficient dimension reduction through optimization,” *Journal of Statistical Planning and Inference*, 211, 154–161.

Du, L., Guo, X., Sun, W., and Zou, C. (2021), “False discovery rate control under general dependence by symmetrized data aggregation,” *Journal of the American Statistical Association*, 1–15.

Fan, J., Han, X., and Gu, W. (2012), “Estimating false discovery proportion under arbitrary covariance dependence,” *Journal of the American Statistical Association*, 107, 1019–1035.

Fan, J., Li, R., Zhang, C.-H., and Zou, H. (2020), *Statistical foundations of data science*, CRC press.

Fan, J. and Lv, J. (2010), “A selective overview of variable selection in high dimensional feature space,” *Statistica Sinica*, 20, 101–148.

Fan, J. and Peng, H. (2004), “Nonconcave penalized likelihood with a diverging number of parameters,” *The Annals of Statistics*, 32, 928–961.
Guo, X., Ren, H., Zou, C., and Li, R. (2022), “Threshold selection in feature screening for error rate control,” *Journal of the American Statistical Association*, 1–13.

Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C., et al. (2001), “Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks,” *Nature Medicine*, 7, 673–679.

Li, B. and Wang, S. (2007), “On directional regression for dimension reduction,” *Journal of the American Statistical Association*, 102, 997–1008.

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

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

Li, L., Wen, X. M., and Yu, Z. (2020), “A selective overview of sparse sufficient dimension reduction,” *Statistical Theory and Related Fields*, 4, 121–133.

Luo, W. and Li, B. (2016), “Combining eigenvalues and variation of eigenvectors for order determination,” *Biometrika*, 103, 875–887.

Meinshausen, N., Meier, L., and Bühlmann, P. (2009), “P-values for high-dimensional regression,” *Journal of the American Statistical Association*, 104, 1671–1681.

Shao, Y., Cook, R. D., and Weisberg, S. (2007), “Marginal tests with sliced average variance estimation,” *Biometrika*, 94, 285–296.
Svetulevičienė, V. (1982), “Multidimensional local limit theorems for probabilities of moderate deviations,” *Lithuanian Mathematical Journal*, 22, 416–420.

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

Wasserman, L. and Roeder, K. (2009), “High dimensional variable selection,” *The Annals of Statistics*, 37, 2178–2201.

Wu, Y. and Li, L. (2011), “Asymptotic properties of sufficient dimension reduction with a diverging number of predictors,” *Statistica Sinica*, 2011, 707–703.

Xia, Y., Tong, H., Li, W., and Zhu, L.-X. (2002), “An adaptive estimation of dimension reduction space,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64, 363–410.

Yin, X. and Cook, R. D. (2002), “Dimension reduction for the conditional kth moment in regression,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64, 159–175.

Yu, Z. and Dong, Y. (2016), “Model-free coordinate test and variable selection via directional regression,” *Statistica Sinica*, 26, 1159–1174.

Yu, Z., Dong, Y., and Shao, J. (2016a), “On marginal sliced inverse regression for ultrahigh dimensional model-free feature selection,” *The Annals of Statistics*, 44, 2594–2623.

Yu, Z., Dong, Y., and Zhu, L.-X. (2016b), “Trace pursuit: A general framework for model-free variable selection,” *Journal of the American Statistical Association*, 111, 813–821.
Zhu, L. (2020), “Review of sparse sufficient dimension reduction: comment,” *Statistical Theory and Related Fields*, 4, 134–134.

Zou, C., Ren, H., Guo, X., and Li, R. (2020), “A new procedure for controlling false discovery rate in large-scale t-tests,” *arXiv preprint arXiv:2002.12548*. 
Supplementary Material for “Model-free variable selection in sufficient dimension reduction via FDR control”

This Supplementary Material contains the proofs of some technical lemmas and additional simulation results.

S1. Additional lemmas

The first lemma is the standard Bernstein’s inequality.

Lemma S.1 (Bernstein’s inequality). Let $X_1, \ldots, X_n$ be independent centered random variables a.s. bounded by $A < \infty$ in absolute value. Let $\sigma^2 = n^{-1} \sum_{i=1}^{n} \mathbb{E}(X_i^2)$. Then for all $x > 0$,

$$\Pr\left(\sum_{i=1}^{n} X_i \geq x\right) \leq \exp\left(-\frac{x^2}{2n\sigma^2 + 2Ax/3}\right).$$

The second one is a moderate deviation result for the mean of random vector; See Theorem 1 in Svetulevičienė (1982).

Lemma S.2 (Moderate deviation for the independent sum). Suppose that $X_1, \ldots, X_n \in \mathbb{R}^s$ are independent identically distributed random vectors with mean zero and identity covariance matrix. Let $\mathbb{E}(\|X\|_2^q) < \infty$ for some $q > 2 + c^2$ with $c^2 > 0$. Then in the domain $\|x\|_2 \leq c\sqrt{\log n}$, one has uniformly

$$\frac{p_n(x)}{(2\pi)^{-s/2} \exp(-\|x\|_2^2/2)} \to 1,$$
as } n \to \infty \text{. Here } p_n(x) \text{ is the density function of } \sum_{i=1}^n X_i/\sqrt{n}.

S2. Useful lemmas

For simplicity of notation, the constant } c \text{ and } C \text{ may be slightly abused hereafter. The following lemma establishes uniform bounds for } \tilde{B}_{2j}, j = 1, \ldots, p.

Lemma S.3. Suppose Assumptions 1–4 hold. Then, for a large } C > 0 \text{ holds uniformly in } S. \text{ Here denote } \sigma^2 = H \max_{h=1}^H e_j^\top \Sigma_S^{-1} \mathbb{E}[X_s X_s^\top \varepsilon_{ih}^2] \Sigma_S^{-1} e_j / (e_j^\top \Sigma_S^{-1} e_j).

Proof. Note that

\[
\Pr \left( \tilde{s}_{2Sj}^{-1} \| \tilde{B}_{2j} \|_2 > x \right) = \Pr \left( \sum_{h=1}^H \tilde{B}_{2j,h}^2 > x^2 \tilde{s}_{2Sj}^2 \right)
\leq \Pr \left( \tilde{B}_{2j,h}^2 > x^2 \tilde{s}_{2Sj}^2 / H, \exists h = 1, \ldots, H \right)
\leq H \max_h \Pr \left( |\tilde{B}_{2j,h}| > x \tilde{s}_{2j} / \sqrt{H} \right).
\]

Recall that } \tilde{B}_{2j,h} = e_j^\top \Sigma_S^{-1} n^{-1} \sum_{i=1}^n X_{2Si} \varepsilon_{ih}. \text{ Let } B = \{ \max_{1 \leq h \leq H} \max_{1 \leq i \leq n} \| \Sigma_S^{-1} X_{2Si} \varepsilon_{ih} \|_\infty \leq m_n \}, \text{ where } m_n = (n \tilde{q}_n)^{1/\infty + \gamma} K_n \text{ for some small } \gamma > 0.

In what follows, we first work on the case of the occurrence of } B. \text{ Denote } \sigma_h^2 = e_j^\top \Sigma_S^{-1} \mathbb{E}[X_s X_s^\top \varepsilon_{ih}^2] \Sigma_S^{-1} e_j
and $\sigma^2 = \sigma_h^2 H/(e_j^\top \Sigma_S^{-1} e_j)$. The Bernstein’s inequality in Lemma S.1 yields that

$$\Pr \left( \| \hat{B}_{2j,h} \| > x \tilde{s}_{2Sj}/\sqrt{H} \text{ for some } j \mid D_1 \right) \leq q_n \max_{j \in S} \Pr \left( \left| \sum_{i=1}^n e_j^\top \Sigma_S^{-1} X_{2Si} \varepsilon_{ih} \right| > n x \tilde{s}_{2Sj}/\sqrt{H} \mid D_1 \right)$$

$$\leq 2\tilde{q}_n \max_{j \in S} \exp \left\{ -\frac{n^2 x^2 \tilde{s}_{2Sj}^2 / H}{2n \sigma_h^2 + 2n x \tilde{s}_{2Sj} / \sqrt{H} \cdot m_n / 3} \right\}$$

$$\leq 2\tilde{q}_n \max_{j \in S} \exp \left\{ -\frac{x^2 e_j^\top \Sigma_S^{-1} e_j / H}{2\sigma_h^2 + 2x \sqrt{e_j^\top \Sigma_S^{-1} e_j / nH} \cdot m_n / 3} \right\}$$

$$= 2\tilde{q}_n \max_{j \in S} \exp \left\{ -\frac{x^2}{2\sigma^2 + 2m_n x / (3\sqrt{n}) \sqrt{H / e_j^\top \Sigma_S^{-1} e_j}} \right\}$$

$$= o(1/\tilde{q}_n),$$

holds uniformly in $S$. Here we use the condition $m_n \sqrt{\log \tilde{q}_n / n} = o(1)$ which is implied by Assumption 3 and $x = \sqrt{C \log \tilde{q}_n}$.

Next, we turn to consider the case on $B^c$. By Assumption 3 and Markov’s inequality

$$\Pr(B^c) \leq nH \max_{i,h} \Pr(\| \Sigma_S^{-1} X_{2Si} \varepsilon_{ih} \|_\infty^\infty > m_n^\infty) = o(1/\tilde{q}_n).$$

The lemma is proved. \qed

The next lemma establishes uniform bounds for $\hat{B}_{2j}$, $j = 1, \ldots, p$.

**Lemma S.4.** Suppose Assumptions 1–4 hold. Then, for a large $C > 0$

$$\Pr \left( s_{2Sj}^{-1} \| \hat{B}_{2j} - B_j \|_2 > \sigma \sqrt{C \tilde{q}_n \log \tilde{q}_n} \mid D_1 \right) = o(1/\tilde{q}_n),$$

uniformly holds for $j \in S$. 3
Proof. By the similar proof of Lemma S.3 and Assumption 4, we have

\[
\Pr \left( s_{2S_j}^{-1} \left\| \hat{B}_{2j} - B_j \right\|_2 > x \right) \leq H \max_h \Pr \left( \left\| \hat{B}_{2j,h} - B_{j,h} \right\| > x s_{2S_j} / \sqrt{H} \right)
\leq H \max_h \Pr \left( \left\| \hat{B}_{2j,h} - B_{j,h} \right\| > x / (\kappa \sqrt{n H}) \right).
\]

Conditionally on \(D_1\), we know that \(\hat{B}_{2j,h} - B_{j,h} = e_j^T \left( n^{-1} \sum_{i=1}^n X_{2Si} X_{2Si}^\top \right)^{-1} n^{-1} \sum_{i=1}^n X_{2Si} \varepsilon_{ih}\).

Using Assumption 4

\[
\max_{j \in S} \left| \hat{B}_{2j,h} - B_{j,h} \right| \leq \kappa^{-1} \left| \frac{1}{n} \sum_{i=1}^n e_j^T X_{2Si} \varepsilon_{ih} \right| \leq \kappa^{-1} \sqrt{q_n} \max_{j \in S} |\zeta_j|,
\]

where \(\zeta = n^{-1} \sum_{i=1}^n e_j^T X_{2Si} \varepsilon_{ih} = (\zeta_1, \ldots, \zeta_p)^\top\).

Let \(z = c \sqrt{\log q_n / n}\) with \(c\) being very large and \(L\) is a positive constant. By Lemma S.1 again, we have

\[
\Pr \left( \max_{j \in S} |\zeta_j| > z \right) \leq \bar{q}_n \max_{j \in S} \Pr \left( \left| \sum_{i=1}^n e_j^T X_{2Si} \varepsilon_{ih} \right| > nz \right)
\leq 2\bar{q}_n \max_{j \in S} \exp \left\{ - \frac{n^2 z^2}{2n E[ (e_j^T X_{2Si})^2 \varepsilon_{ih}^2 ]} + 2 (\max_{1 \leq h \leq H, 1 \leq i \leq n} |e_j^T X_{2Si} \varepsilon_{ih}|) \times nz / 3 \right\}
\to 0,
\]

where we use the condition \(\max_{1 \leq h \leq H, 1 \leq i \leq n} E \left\| X_{2Si} \varepsilon_{ih} \right\|_\infty^\omega \leq L K_n^\omega\) and \((n\bar{q}_n)^{1/\omega+\gamma} \times \sqrt{\log q_n / n} \to 0\) implied by Assumption 3.
Thus, for some large $C > 0$, we have

\[
\Pr \left( s_{2S_j}^{-1} \| \tilde{B}_{2j} - B_j \|_2 > x \right) \leq H \max_h \Pr \left( \| \tilde{B}_{2j,h} - B_{j,h} \|_2 > x / (\tilde{\kappa} \sqrt{nH}) \right)
\]

\[
\leq H \max_h \Pr \left( \max_{j \in S} \left| \frac{1}{n} \sum_{i=1}^{n} e_j^\top X_{2S_i} \varepsilon_{ih} \right| > x / (\tilde{\kappa} \sqrt{nq_n H}) \right)
\]

\[
= H \max_h \Pr \left( \max_{j \in S} |\zeta_j| > z \right) \to 0,
\]

hold if $x = C \sqrt{q_n \log q_n}$.

\[\square\]

S3. Proofs of lemmas in Appendix

Proof of Lemma A.1.

Define $b_n = \sigma \sqrt{C \log q_n}$ where $C > 0$. Hereafter for simplicity, we denote $T_{1j} = \tilde{B}_{1j} / s_{1S_j}$ and $T_{2j} = \tilde{B}_{2j} / \tilde{s}_{2S_j}$. Thus $\tilde{W}_j = T_{1j}^\top T_{2j}$. We observe that

\[
\frac{G(t)}{G_-(t)} - 1 = \frac{\sum_{j \in A^c} \left\{ \Pr(\tilde{W}_j \geq t \mid D_1) - \Pr(\tilde{W}_j \leq -t \mid D_1) \right\}}{\sum_{j \in A^c} \Pr(\tilde{W}_j \leq -t \mid D_1)}
\]

\[
+ \frac{\sum_{j \in A^c} \left\{ \Pr(T_{1j}^\top T_{2j} \geq t, \|T_{2j}\|_2 \leq b_n \mid D_1) - \Pr(T_{1j}^\top T_{2j} \leq -t, \|T_{2j}\|_2 \leq b_n \mid D_1) \right\}}{q_n G_-(t)}
\]

\[
+ \frac{\sum_{j \in A^c} \left\{ \Pr(T_{1j}^\top T_{2j} \geq t, \|T_{2j}\|_2 > b_n \mid D_1) - \Pr(T_{1j}^\top T_{2j} \leq -t, \|T_{2j}\|_2 > b_n \mid D_1) \right\}}{q_n G_-(t)}
\]

\[:= \Lambda_1 + \Lambda_2.\]

Note that $G_-(t)$ is a decreasing function by definition. Firstly, for the term $\Lambda_2$, by Lemma
S.3, we have
\[
\sum_{j \in A^c} \Pr(T_{1j}^T T_{2j} \geq t, \|T_{2j}\|_2 > b_n \mid D_1) \leq \sum_{j \in A^c} \Pr(\|T_{2j}\|_2 > b_n \mid D_1) \leq \bar{q}_n \times o(1/\bar{q}_n) \eta_n, 
\]
which implies \(\Lambda_2 = o(1)\).

Recall that
\[
T_{2j}^\top = e_j^\top \Sigma_S^{-1} \sum_{i=1}^n X_{2Si} e_{2i}^\top / \sqrt{ne_j \Sigma_S^{-1} e_j}.
\]

By Lemma S.2 and Assumption 3, we obtain that
\[
\Pr(T_{1j}^T T_{2j} \geq t, \|T_{2j}\|_2 \leq b_n \mid D_1) \rightarrow 1,
\]
where \(U \sim \mathcal{N}(0, \bar{\Sigma})\) which is independent of \(T_{1j}\). Here \(\bar{\sigma}_{th} = e_j^\top \Sigma_S^{-1} \mathbb{E}[\epsilon_i X_S^\top X_S \epsilon_h^\top] \Sigma_S^{-1} e_j / e_j^\top \Sigma_S^{-1} e_j\) and \(\bar{\Sigma} = (\bar{\sigma}_{th})_{i,h=1}^H\).

Similarly we get
\[
\Pr(T_{1j}^T T_{2j} \leq -t, \|T_{2j}\|_2 \leq b_n \mid D_1) \rightarrow 1.
\]

Note that \(\Pr(T_{1j}^T U \leq -t, \|U\|_2 \leq b_n \mid D_1) = \Pr(T_{1j}^T U \geq t, \|U\|_2 \leq b_n \mid D_1)\), from which we get \(\Lambda_1 = o(1)\) and accordingly we can claim the assertion.

Proof of Lemma A.2.

We prove the first formula and the second one can be deduced similarly. By the proof of
Lemma A.1, it suffices to show that

\[ G(t) = \frac{1}{q_0 n} \sum_{j \in \mathcal{A}^c} \Pr(\tilde{W}_j \geq t, \|T_{2j}\|_2 \leq b_n \mid \mathcal{D}_1) \{1 + o(1)\} := \bar{G}(t) \{1 + o(1)\}. \]

Accordingly

\[ \frac{1}{q_0 n} \sum_{j \in \mathcal{A}^c} \mathbb{I}(\tilde{W}_j \geq t) = \frac{1}{q_0 n} \sum_{j \in \mathcal{A}^c} \mathbb{I}(\tilde{W}_j \geq t, \|T_{2j}\|_2 \leq b_n) \{1 + o_p(1)\}. \]

Thus, we need to show the following assertion

\[ \sup_{0 \leq t \leq \bar{G}^{-1}(\alpha \eta_n / q_0 n)} \left| \frac{\sum_{j \in \mathcal{A}^c} \mathbb{I}(\tilde{W}_j \geq t, \|T_{2j}\|_2 \leq b_n)}{q_0 n \bar{G}(t)} - 1 \right| = o_p(1). \]

Note that the \( \bar{G}(t) \) is a decreasing and continuous function. Let \( a_p = \alpha \eta_n, z_0 < z_1 < \cdots < z_{h_n} \leq 1 \) and \( t_i = \bar{G}^{-1}(z_i) \), where \( z_0 = a_p / q_0 n, z_i = a_p / q_0 n + b_p \exp(i \zeta) / q_0 n, h_n = \{\log((q_0 n - a_p)/b_p)\}^{1/\zeta} \) with \( b_p/a_p \to 0 \) and \( 0 < \zeta < 1 \). Note that \( \bar{G}(t_i)/\bar{G}(t_{i+1}) = 1 + o(1) \) uniformly in \( i \). It is therefore enough to derive the convergence rate of the following formula

\[ D_n = \sup_{0 \leq i \leq h_n} \left| \frac{\sum_{j \in \mathcal{A}^c} \left\{ \mathbb{I}(\tilde{W}_j \geq t, \|T_{2j}\|_2 \leq b_n) - \Pr(\tilde{W}_j \geq t, \|T_{2j}\|_2 \leq b_n \mid \mathcal{D}_1) \right\}}{q_0 n \bar{G}(t_i)} \right|. \]

Define \( \mathcal{B} = \{\|T_{2j}\|_2 \leq b_n, j \in \mathcal{A}^c\} \) and then we have

\[ D(t) = \mathbb{E} \left[ \left( \sum_{j \in \mathcal{A}^c} \left\{ \mathbb{I}(\tilde{W}_j > t, \|T_{2j}\|_2 \leq b_n) - \Pr(\tilde{W}_j > t, \|T_{2j}\|_2 \leq b_n \mid \mathcal{D}_1) \right\} \right)^2 \mid \mathcal{D}_1 \right] \]

\[ = \sum_{j \in \mathcal{A}^c} \sum_{l \in \mathcal{A}^c} \left\{ \Pr(\tilde{W}_j > t, \tilde{W}_l > t \mid \mathcal{D}_1, \mathcal{B}) - \Pr(\tilde{W}_j > t \mid \mathcal{D}_1, \mathcal{B}) \Pr(\tilde{W}_l > t \mid \mathcal{D}_1, \mathcal{B}) \right\} \{1 + o(1)\}. \]
Further define \( \mathcal{M}_j = \{ l \in \mathcal{A}^c : |\rho_{jl}| \geq C(\log n)^{-2-\nu} \} \). Here \( \rho_{jl} \) denotes the conditional correlation between \( \tilde{W}_j \) and \( \tilde{W}_l \) given \( \mathcal{D}_1 \). Recall that

\[
\tilde{W}_j = \frac{e_j^\top \Sigma_{S}^{-1} \sum_{i=1}^{n} X_{2S}^\top e_{2i} T_{1j}}{\sqrt{n e_j^\top \Sigma_{S}^{-1} e_j}}.
\]

It follows that

\[
\text{var}(\tilde{W}_j \mid \mathcal{D}_1) = \frac{e_j^\top \Sigma_{S}^{-1} \mathbb{E}[|X_S| X_S^\top (\varepsilon^\top T_{1j})^2] \Sigma_{S}^{-1} e_j}{e_j^\top \Sigma_{S}^{-1} e_j},
\]

\[
\text{cov}(\tilde{W}_j, \tilde{W}_l \mid \mathcal{D}_1) = \frac{e_j^\top \Sigma_{S}^{-1} \mathbb{E}[|X_S| X_S^\top |\varepsilon^\top T_{1j} T_{1l}^\top \varepsilon|] \Sigma_{S}^{-1} e_l}{\sqrt{e_j^\top \Sigma_{S}^{-1} e_j \cdot \sqrt{e_l^\top \Sigma_{S}^{-1} e_l}}}. 
\]

So we get

\[
\rho_{jl} = \frac{e_j^\top \Sigma_{S}^{-1} \mathbb{E}[|X_S| X_S^\top \varepsilon^\top T_{1j} T_{1l}^\top \varepsilon] \Sigma_{S}^{-1} e_l}{\sqrt{e_j^\top \Sigma_{S}^{-1} \mathbb{E}[|X_S| X_S^\top (\varepsilon^\top T_{1j})^2] \Sigma_{S}^{-1} e_j} \cdot \sqrt{e_l^\top \Sigma_{S}^{-1} \mathbb{E}[|X_S| X_S^\top (\varepsilon^\top T_{1l})^2] \Sigma_{S}^{-1} e_l}}.
\]

But now the conditional correlation between \( \tilde{W}_j \) and \( \tilde{W}_l \) given \( \mathcal{D}_1 \), \( \rho_{jl} \), still depends on the \( \mathcal{D}_1 \) and \( \varepsilon \) is not independent with \( X_S \). By Assumption 7

\[
D(t) \leq \sum_{j \in \mathcal{A}^c} \sum_{l \in \mathcal{M}_j^c} \Pr(\tilde{W}_j > t \mid \mathcal{D}_1) + \sum_{j \in \mathcal{A}^c} \sum_{l \in \mathcal{M}_j^c} \left\{ \Pr(\tilde{W}_j > t, \tilde{W}_l > t \mid \mathcal{D}_1) - \Pr(\tilde{W}_j > t \mid \mathcal{D}_1) \Pr(\tilde{W}_l > t \mid \mathcal{D}_1) \right\}
\]

\[
\leq r_{pG_0n} G(t) + \sum_{j \in \mathcal{A}^c} \sum_{l \in \mathcal{M}_j^c} \left\{ \Pr(\tilde{W}_j > t, \tilde{W}_l > t \mid \mathcal{D}_1, \mathcal{B}) - \Pr(\tilde{W}_j > t \mid \mathcal{D}_1, \mathcal{B}) \Pr(\tilde{W}_l > t \mid \mathcal{D}_1, \mathcal{B}) \right\}.
\]

While for each \( j \in \mathcal{A}^c \) and \( l \in \mathcal{M}_j^c \), conditional on \( \mathcal{D}_1 \), by Lemma 1 in Cai and Liu Cai
and Liu (2016) we have

\[
\begin{align*}
\left| \frac{\Pr(\tilde{W}_j > t, \tilde{W}_r > t \mid \mathcal{D}_1, \mathcal{B}) - \Pr(\tilde{W}_j > t \mid \mathcal{D}_1, \mathcal{B})}{\Pr(\tilde{W}_j > t \mid \mathcal{D}_1, \mathcal{B})} - \frac{\Pr(\tilde{W}_r > t \mid \mathcal{D}_1, \mathcal{B})}{\mathcal{G}(t)} \right| & \leq A_n,
\end{align*}
\]

uniformly holds, where \(A_n = (\log n)^{-1-\nu_1}\) for \(\nu_1 = \min(\nu, 1/2)\).

From the above results and Chebyshev’s inequality, for \(\xi > 0\) we have

\[
\Pr(D_n \geq \xi \mid \mathcal{D}_1) \leq \sum_{i=0}^{h_n} \Pr \left( \left. \left| \sum_{j \in \mathcal{A}^c} \mathbb{I}(W_j > t_i, \|T_{2j}\|_2 \leq b_n) - \Pr(W_j > t_i, \|T_{2j}\|_2 \leq b_n \mid \mathcal{D}_1) \right| \geq \xi \right| \mathcal{D}_1 \right)
\]

\[
\leq \frac{1}{\xi^2} \sum_{i=0}^{h_n} \frac{1}{q_{0n}\mathcal{G}^2(t_i)} D(t_i)
\]

\[
\leq \frac{1}{\xi^2} \sum_{i=0}^{h_n} \frac{1}{q_{0n}\mathcal{G}^2(t_i)} \left\{ r_p q_{0n} \mathcal{G}(t_i) + q_{0n}^2 \mathcal{G}^2(t_i) A_n \right\}
\]

\[
\leq \frac{1}{\xi^2} \left\{ \sum_{i=0}^{h_n} \frac{r_p}{q_{0n} \mathcal{G}(t_i)} + h_n A_n \right\}.
\]

Moreover, observe that

\[
\sum_{i=0}^{h_n} \frac{1}{q_{0n} \mathcal{G}(t_i)} = \frac{1}{a_p} + \sum_{i=1}^{h_n} \frac{1}{a_p + b_p e^{\zeta c}} \lesssim b_p^{-1}.
\]

Note that \(\zeta\) can be arbitrarily close to 1 such that \(h_n A_n \to 0\). Because \(b_p\) can be made arbitrarily large as long as \(b_p/a_p \to 0\), we have \(D_n = o_p(1)\) providing that \(r_p/b_p \to 0\).

\[\square\]

Proof of Lemma A.3.
By definition

\[ W_j - \tilde{W}_j = \frac{\hat{B}_{1j}^\top s_{1Sj}s_{2Sj}}{\hat{B}_{2j}^\top s_{1Sj}s_{2Sj}} - \frac{\hat{B}_{1j}^\top s_{1Sj}}{\tilde{s}_{2Sj}} \left( \hat{B}_{2j} - \tilde{B}_{2j} \right), \]

where \( \|\hat{B}_{1j}/s_{1Sj}\|_\infty = O_p(\sqrt{nc_{np}}) \) uniformly holds for \( j \in S \) by Assumption 5.

Recall that

\[ \frac{\hat{B}_{2j}^\top s_{2Sj}}{s_{2Sj}} - \frac{\tilde{B}_{2j}^\top s_{2Sj}}{s_{2Sj}} = \frac{e_j^\top A^{-1} \sum_{i=1}^n X_{2Si} \varepsilon_i^\top}{\sqrt{ne_j^\top Ae_j}} - \frac{e_j^\top \Sigma^{-1}_S \sum_{i=1}^n X_{2Si} \varepsilon_i^\top}{\sqrt{ne_j^\top \Sigma^{-1}_S e_j}} \]

\[ = \frac{1}{\sqrt{n}} e_j^\top (A^{-1} - \Sigma^{-1}_S) \sum_{i=1}^n X_{2Si} \varepsilon_i^\top \frac{1}{\sqrt{e_j^\top \Sigma^{-1}_S e_j}} \]

\[ + \left( \frac{1}{\sqrt{e_j^\top A^{-1} e_j}} - \frac{1}{\sqrt{e_j^\top \Sigma^{-1}_S e_j}} \right) \frac{1}{\sqrt{n}} e_j^\top \Sigma^{-1}_S \sum_{i=1}^n X_{2Si} \varepsilon_i^\top \]

\[ + \frac{1}{\sqrt{n}} e_j^\top (A^{-1} - \Sigma^{-1}_S) \sum_{i=1}^n X_{2Si} \varepsilon_i^\top \left( \frac{1}{\sqrt{e_j^\top A^{-1} e_j}} - \frac{1}{\sqrt{e_j^\top \Sigma^{-1}_S e_j}} \right) \]

\[ := \Lambda_1 + \Lambda_2 + \Lambda_3, \]

where \( A = n^{-1} \sum_{i=1}^n X_{2Si} X_{2Si}^\top \). Firstly note that

\[ \left\| A^{-1} - \Sigma^{-1}_S \right\|_\infty \leq \left\| \Sigma^{-1}_S \right\|_\infty \left\| A^{-1} \right\|_\infty \left\| \Sigma_S - A \right\|_\infty = O_p(nq_n a_{np}). \]

Further observe that

\[ \left\| \frac{1}{\sqrt{n}} e_j^\top (A^{-1} - \Sigma^{-1}_S) \sum_{i=1}^n X_{2Si} \varepsilon_i^\top \right\|_\infty \leq \left\| A^{-1} - \Sigma^{-1}_S \right\|_\infty \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n X_{2Si} \varepsilon_i^\top \right\|_\infty = O_p(nq_n a_{np}\sqrt{\log q_n}). \]
While by Assumption 4, there exists a constant $c$ such that

\[
\begin{vmatrix}
\sqrt{e_j^\top A^{-1} e_j} - 1 \\
\sqrt{e_j^\top \Sigma^{-1} e_j} - 1
\end{vmatrix} = \frac{|e_j^\top \Sigma^{-1} e_j - e_j^\top A^{-1} e_j|}{\sqrt{e_j^\top A^{-1} e_j + e_j^\top \Sigma^{-1} e_j}} \leq c \| \Sigma^{-1} - A^{-1} \|_\infty = O_p(\nu n\bar{q}_n a_{np}).
\]

Thus, $\Lambda_1 = O_p(cnpq_n a_{np}\sqrt{\log \bar{q}_n})$, $\Lambda_2 = O_p(cnpq_n a_{np}\sqrt{\log \bar{q}_n})$, and $\Lambda_3 = O_p(c_n^2 a_{np}^2 \nu n\sqrt{\log \bar{q}_n})$ is a small order than $\Lambda_1$ and $\Lambda_2$. By triangle inequality and Lemma S.4, $W_j - \tilde{W}_j = O_p(\nu n\bar{q}_n a_{np}\sqrt{\log \bar{q}_n}) = O_p(cnp a_{np} q_n \sqrt{\log \bar{q}_n}).$ 

**Proof of Lemma A.4.**

By Lemma A.3, with probability tending to one

\[
\begin{align*}
\left| \sum_{j \in A^c} \mathbb{I}(W_j \geq t) - \sum_{j \in A^c} \mathbb{I}(\tilde{W}_j \geq t) \right| &\leq \sum_{j \in A^c} \left\{ \mathbb{I}(\tilde{W}_j \geq t + l_n) - \mathbb{I}(\tilde{W}_j \geq t) \right\} \\
&\quad + \sum_{j \in A^c} \left\{ \mathbb{I}(\tilde{W}_j \geq t - l_n) - \mathbb{I}(\tilde{W}_j \geq t) \right\} \\
&:= \Lambda_1 + \Lambda_2,
\end{align*}
\]

where $l_n/(c_{np} a_{np} q_n \sqrt{n \log \bar{q}_n}) \to \infty$ as $(n, p) \to \infty$.

Define event $C_t = \{s_{18j}^{-1} \| \tilde{B}_{1j} \|_2 > t/(\sigma \sqrt{C \log \bar{q}_n}), s_{28j}^{-1} \| \tilde{B}_{2j} \|_2 > t/(\sqrt{\nu c_{np}}), j \in A^c \}$. Then

\[
\mathbb{E}(\Lambda_1) = \mathbb{E} \left\{ \sum_{j \in A^c} \mathbb{I}(t \leq \tilde{W}_j \leq t + l_n) \right\} \leq \sum_{j \in A^c} \Pr(t \leq \tilde{W}_j \leq t + l_n \mid C_t) + \sum_{j \in A^c} \Pr(t \leq \tilde{W}_j \leq t + l_n, \mathcal{C}_t^c),
\]

11
where we use Lemma S.3 and Assumption 5 to get \( \Pr(t \leq \tilde{W}_j \leq t + l_n, \mathcal{C}_t) = o(\bar{q}_n^{-1}) \). Recall that \( b_n = \sigma \sqrt{C \log \bar{q}_n} \) in Lemma A.1. Under the event \( \mathcal{C}_t \) and the condition \( l_n \to 0 \), we have

\[
\|T_{2j}\|_2 \leq \frac{t + l_n}{\|T_{1j}\|_2} \leq \frac{\sigma(t + l_n) \sqrt{C \log \bar{q}_n}}{t} = \sigma \sqrt{C \log \bar{q}_n} + \frac{\sigma l_n \sqrt{C \log \bar{q}_n}}{t} \leq \sigma \sqrt{C \log \bar{q}_n} = b_n. \tag{S.1}
\]

Let \( U_j = B_j/\sigma_j \) and \( Z \) from \( p \)-dimensional standard normal distribution. By Lemma A.1 and Equation (S.1), it suffices to show that

\[
\sum_{j \in A^c} \Pr \left( t \leq T_{1j}^\top (U_j + Z) \leq t + l_n | \mathcal{C}_t \right)
= \sum_{j \in A^c} \Pr \left( \frac{t - \|T_{1j}\|_2 \|U_j\|_2}{\|T_{1j}\|_2} \leq \|Z\|_2 \leq \frac{t + l_n - \|T_{1j}\|_2 \|U_j\|_2}{\|T_{1j}\|_2} | \mathcal{C}_t \right)
\leq \sum_{j \in A^c} \mathbb{E} \left\{ \phi \left( \frac{t}{\|T_{1j}\|_2} - \|U_j\|_2 \right) - \Phi \left( \frac{t}{\|T_{1j}\|_2} - \|U_j\|_2 \right) | \mathcal{C}_t \right\}
\leq l_n \sum_{j \in A^c} \mathbb{E} \left\{ \phi \left( \frac{t}{\|T_{1j}\|_2} - \|U_j\|_2 \right) | \mathcal{C}_t \right\}
\leq l_n M^{-1} \log \bar{q}_n \sum_{j \in A^c} \mathbb{E} \left\{ \Phi \left( \frac{t}{\|T_{1j}\|_2} - \|U_j\|_2 \right) | \mathcal{C}_t \right\},
\]

where \( \Phi(x) = 1 - \Phi(x) \), \( \phi(x) \) and \( \Phi(x) \) are the density function and cumulative distribution function of standard normal distribution, respectively. When the event \( \{t \leq \tilde{W}_j \leq t + l_n, \mathcal{C}_t\} \) occurs, the second to last inequality is due to

\[
\phi(x) < \frac{x^2 + 1}{x} \Phi(x), \text{ for all } x > 0.
\]
On the other hand, by definition,

$$\sum_{j \in A^c} \Pr(\tilde{W}_j > t) = \sum_{j \in A^c} \mathbb{E} \left\{ \tilde{\Phi} \left( \frac{t}{\|T_i\|_2} - \|U_j\|_2 \right) \mid C_t \right\} \{ 1 + o(1) \}.$$ 

Note that $h_n$ can be made arbitrarily small as $n \to \infty$ in Lemma A.2. By the similar proof in Lemma A.2, the result holds if $c_{np}a_{np}v_nq_n\sqrt{n \log q_n \log q_n h_n} \to 0$. The part of $\Lambda_2$ is proved similarly as the part of $\Lambda_1$. Accordingly, we can claim the assertion. 

S4. Additional simulations

Table S1: FDR, TPR, $P_a = \Pr(A \subseteq \hat{A}(L))($%) and computing time for several methods against different $X$ distributions under Scenarios 2a–2c when $(n, p, p_1, \rho) = (800, 1000, 10, 0.5)$. 

| Scenario | Method         | normal FDR | normal TPR | normal $P_a$ | time | mixed FDR  | mixed TPR | mixed $P_a$ | time |
|----------|----------------|-------------|-------------|--------------|------|------------|------------|--------------|------|
| 2a       | MFSDA          | 18.5        | 100.0       | 100.0        | 18.6 | 17.9       | 100.0      | 99.6         | 24.2 |
|          | MFSDA-DB       | 18.1        | 93.7        | 63.4         | 213.5 | 18.1       | 96.7       | 79.0         | 209.9 |
|          | MSIR-BH        | 2.7         | 37.0        | 0.0          | 14.1 | 4.4        | 38.7       | 0.0          | 20.2 |
|          | IM-SIR1        | 0.0         | 50.0        | 0.0          | 28.5 | 0.0        | 50.0       | 0.0          | 34.8 |
|          | IM-SIR2        | 83.1        | 100.0       | 100.0        | 28.5 | 83.1       | 100.0      | 100.0        | 34.6 |
|          | MX-Knockoff    | 71.8        | 6.0         | 0.0          | 51.6 | 27.4       | 11.9       | 0.0          | 56.7 |
| 2c       | MFSDA          | 17.7        | 99.5        | 94.6         | 18.4 | 17.9       | 99.5       | 95.2         | 27.7 |
|          | MFSDA-DB       | 17.3        | 91.6        | 41.6         | 215.7 | 17.9      | 93.6       | 56.2         | 234.8 |
|          | MSIR-BH        | 4.4         | 39.5        | 0.0          | 14.0 | 4.6        | 39.1       | 0.0          | 20.8 |
|          | IM-SIR1        | 0.0         | 50.0        | 0.0          | 28.6 | 0.0        | 50.0       | 0.0          | 35.5 |
|          | IM-SIR2        | 83.1        | 100.0       | 100.0        | 28.6 | 83.1       | 100.0      | 100.0        | 35.9 |
|          | MX-Knockoff    | 10.0        | 13.9        | 0.0          | 51.6 | 31.3       | 23.6       | 0.0          | 57.7 |
| 2c       | MFSDA          | 17.6        | 99.1        | 92.8         | 21.0 | 16.8       | 98.9       | 90.6         | 38.1 |
|          | MFSDA-DB       | 18.0        | 85.6        | 25.2         | 203.4 | 16.6      | 89.8       | 38.6         | 248.7 |
|          | MSIR-BH        | 5.2         | 38.4        | 0.0          | 16.1 | 4.7        | 37.8       | 0.0          | 30.0 |
|          | IM-SIR1        | 0.0         | 50.0        | 0.0          | 24.4 | 0.0        | 50.0       | 0.0          | 40.8 |
|          | IM-SIR2        | 83.1        | 100.0       | 100.0        | 24.2 | 83.1       | 100.0      | 100.0        | 41.3 |
|          | MX-Knockoff    | 10.3        | 11.1        | 0.0          | 54.9 | 31.5       | 23.6       | 0.0          | 60.5 |
Figure S1: FDR and TPR (%) curves against different covariate dimension $p$, different correlation $\rho$ and different signal number $p_1$ under Scenario 2c when $n = 500$ and $X$ normal distribution. Left panel: $(p_1, \rho) = (10, 0.5)$; Middle panel: $(p, p_1) = (1000, 10)$; Right panel: $(p, \rho) = (1000, 0.5)$. The gray solid line denotes the target FDR level.