The Annals of Statistics
2016, Vol. 44, No. 6, 2594–2623
DOI: 10.1214/15-AOS1424

ON MARGINAL SLICED INVERSE REGRESSION FOR ULTRAHIGH DIMENSIONAL MODEL-FREE FEATURE SELECTION

BY ZHOU YU[1], YUEXIAO DONG AND JUN SHAO[1,2]

East China Normal University, Temple University and University of Wisconsin-Madison

Model-free variable selection has been implemented under the sufficient dimension reduction framework since the seminal paper of Cook [*Ann. Statist.* **32** (2004) 1062–1092]. In this paper, we extend the marginal coordinate test for sliced inverse regression (SIR) in Cook (2004) and propose a novel marginal SIR utility for the purpose of ultrahigh dimensional feature selection. Two distinct procedures, Dantzig selector and sparse precision matrix estimation, are incorporated to get two versions of sample level marginal SIR utilities. Both procedures lead to model-free variable selection consistency with predictor dimensionality p diverging at an exponential rate of the sample size n. As a special case of marginal SIR, we ignore the correlation among the predictors and propose marginal independence SIR. Marginal independence SIR is closely related to many existing independence screening procedures in the literature, and achieves model-free screening consistency in the ultrahigh dimensional setting. The finite sample performances of the proposed procedures are studied through synthetic examples and an application to the small round blue cell tumors data.

1. Introduction. For regression problems between a response $Y \in \mathbb{R}$ and a p-dimensional predictor $\mathbf{x} = (X_1, X_2, \ldots, X_p)^{\mathrm{T}} \in \mathbb{R}^p$, variable or feature selection aims to identify the subset of important predictors and to enhance the model interpretability with parsimonious representation. Research on variable selection in linear models and generalized linear models has gained considerable momentum in the past two decades, which include key developments such as LASSO [Tibshirani (1996)], SCAD [Fan and Li (2001)], group LASSO [Yuan and Lin (2006)], adaptive LASSO [Zou (2006)], Dantzig selector [Candes and Tao (2007)], etc. In the presence of ultrahigh dimensional predictor space, where p diverges with an exponential rate of the sample size n, Fan and Lv (2008) first proposed the sure independence screening (SIS) procedure for feature screening in linear models.

Received May 2015; revised December 2015.

[1]Supported by the National Natural Science Foundation of China 11571111, the 111 project B14019, the program of Shanghai Subject Chief Scientist 14XD1401600 and Shanghai Rising-Star Program 16QA1401700.

[2]Supported in part by NSF Grant DMS-13-05474.

MSC2010 subject classifications. 62H20, 62H99, 62G99.

Key words and phrases. Marginal coordinate test, sliced inverse regression, sufficient dimension reduction, sure independence screening.

By ranking the marginal utility calculated as the Pearson correlation between the response and each individual predictor, the SIS is screening consistent in the sense that, with probability tending to one as $n \to \infty$, the top-ranked predictors retained by SIS include all the important predictors. Fan, Samworth and Wu (2009) and Fan and Song (2010) further extended SIS to generalized linear models by ranking the features according to the marginal likelihood-based utilities. Screening consistency, however, is weaker than selection consistency which further requires that, with probability tending to one, the set of selected predictors does not contain any irrelevant predictor.

Variable selection in the model-free setting aims to identify all the important predictors without knowledge about the link function between the response Y and the predictor $\mathbf{x}$. Many model-free variable selection methods in the literature are developed under the framework of sufficient dimension reduction [Cook (1998), Li (1991)], as sufficient dimension reduction searches linear combinations of $\mathbf{x}$ such that Y is independent of $\mathbf{x}$ given these linear combinations, without requiring estimation of the unknown link function between Y and $\mathbf{x}$. A series of sparse sufficient dimension reduction methods are motivated by combining penalized regression and sufficient dimension reduction, such as Ni, Cook and Tsai (2005), Li and Nachtsheim (2006), Li (2007), Li and Yin (2008), Zhou and He (2008), Bondell and Li (2009), Chen, Zou and Cook (2010) and Yu et al. (2013). By noticing that Pearson correlation measures dependence in linear models, model-free feature screening methods in the ultrahigh dimensional setting can be designed from more general dependence measures. For example, distance correlation [Székely, Rizzo and Bakirov (2007)], Kendall's tau and maximal correlation have been used for feature screening in Li, Zhong and Zhu (2012), Li et al. (2012) and Huang and Zhu (2014), respectively. Model-free feature screening in discriminant analysis has been studied in Mai and Zou (2013), Cui, Li and Zhong (2015) and Pan, Wang and Li (2015). More recently, Mai and Zou (2015) proposed the fused Kolmogorov filter approach, which connects feature screening for continuous response and discrete response.

Although there is a vast literature of applying sufficient dimension reduction for model-free variable selection, result on developing selection consistency for ultrahigh dimensional setting is scant. While model-free variable selection consistency has been established in the diverging p and $p < n$ setting [Jiang and Liu (2014), Wu and Li (2011)] and screening consistency has been proved in the ultrahigh dimensional setting [Li, Zhong and Zhu (2012), Mai and Zou (2015), Zhu et al. (2011); Yu, Dong and Zhu (2016)], there is no selection consistency result in the ultrahigh dimensional setting. One has to apply two methods in two stages to achieve selection consistency, as suggested by Jiang and Liu (2014). In the first stage, a consistent screening is applied in the ultrahigh dimensional setting to reduce the dimension to something less than n, and then a selection consistent method needs to be used in the second stage. Furthermore, due to the interplay between the regression coefficients and the correlation among the predictors, the independence

screening method may not consistently rank the utility of active predictors ahead of the utility of the inactive predictors even in linear models. Although independence screening is easy to implement, correlation among predictors may cause problems.

To fill the aforementioned gaps, we propose an approach called marginal sliced inverse regression (SIR) for model-free variable selection. Our first two procedures are based on the same population level marginal SIR utility, but use Dantzig selector and sparse precision matrix estimation as different sample level estimation schemes. Unlike the popular independence screening procedures in the literature, where the construction of the marginal utility assumes that the predictors are independent, our two marginal SIR procedures take into account the correlation among the predictors. By ranking and thresholding the corresponding sample level marginal SIR utilities, both procedures achieve model-free variable selection consistency in the ultrahigh dimensional setting. Marginal SIR with Dantzig selector exploits the intrinsic sparsity structure in the marginal utility, and requires the minimum assumptions to achieve the desirable selection consistency property. Marginal SIR with sparse precision matrix estimation uses a plug-in sample level utility, and incorporates the correlation among the predictors for variable selection by plugging in the sparse precision matrix estimator. As a special case of our marginal SIR, we also describe a marginal independence SIR, which is obtained by using a diagonal matrix as a working covariance matrix for $\mathbf{x}$. The population level marginal independence SIR utility can be viewed as a generalized Pearson correlation. As a result, our proposal of marginal independence SIR is closely related to the popular independence screening procedures based on Pearson correlation and distance correlation. Last but not least, to determine the threshold value for variable selection in application, we propose to minimize the classification error through cross-validation. The classification can be applied directly with the discrete response. For continuous response, SIR naturally leads to the discretized response through slicing, which can be used for our purpose of classification. We demonstrate through extensive numerical studies that the data-driven threshold works well across a wide range of models.

The rest of the paper is organized as follows. The principle of marginal SIR is discussed in Section 2, where we propose the population level marginal SIR utility. Two sample level utilities for marginal SIR are developed in Sections 3 and 4, respectively, where we use Dantzig selector and sparse precision matrix estimation to facilitate the sample utility estimation. Both procedures achieve model-free selection consistency with p diverging at the exponential rate of n. Section 5 studies marginal independence SIR for feature screening and its model-free screening consistency property. The connections and the differences between our proposals with some popular independence screening procedures are also discussed. Section 6 considers the threshold value for variable selection or feature screening. Finite sample performances of the proposed methods are studied in Section 7 and we conclude the paper with some discussions in Section 8. The proofs for the main theorems are relegated to the Appendix.

2. The principle of the marginal SIR. Throughout the paper, we denote $\|\mathbf{A}\|_\infty = \max_{1\le i\le p, 1\le j\le q} |a_{ij}|$ and $\|\mathbf{A}\|_1 = \max_{1\le j\le q} \sum_{i=1}^p |a_{ij}|$ for any matrix $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{p\times q}$. We assume that $\mathbf{\Sigma} = \mathrm{Var}(\mathbf{x})$ is finite. Without loss of generality, we assume that $E(X_k) = 0$ and $\mathrm{Var}(X_k) = 1$ for $k = 1, \ldots, p$. For predictor $\mathbf{x} = (X_1, X_2, \ldots, X_p)^{\mathrm{T}}$, denote $\mathcal{I} = \{1, 2, \ldots, p\}$ as the full index set. Let $\mathcal{A}$ be the active index set which corresponds to all relevant predictors for the response Y. Then $\mathcal{A}^c$, the complement of $\mathcal{A}$ in $\mathcal{I}$, is the index set that corresponds to all irrelevant predictors. Denote $\mathbf{x}_\mathcal{A} = \{X_k : k \in \mathcal{A}\}$ as the vector containing all the active predictors, and we have $Y \perp\!\!\!\perp \mathbf{x}|\mathbf{x}_\mathcal{A}$, that is, Y is independent of $\mathbf{x}$ given $\mathbf{x}_\mathcal{A}$. And the existence and the uniqueness of the active index set $\mathcal{A}$ is further guaranteed by Yin and Hilafu (2015).

We first briefly review the concept of sufficient dimension reduction and show its connection with model-free variable selection. The primary goal of sufficient dimension reduction is to make inference about the central space between Y and $\mathbf{x}$. The central space is a subspace of $\mathbb{R}^p$ denoted by $\mathcal{S}_{Y|\mathbf{x}}$. Let $\boldsymbol{\beta}_i \in \mathbb{R}^p$, $i = 1, \ldots, d$, be the basis of $\mathcal{S}_{Y|\mathbf{x}}$. Then the column space of $\boldsymbol{\beta}_i$'s satisfies $\mathrm{Span}(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d) = \mathcal{S}_{Y|\mathbf{x}}$. The central space is defined such that $Y \perp\!\!\!\perp \mathbf{x}|(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}, \ldots, \boldsymbol{\beta}_d^{\mathrm{T}}\mathbf{x})$, that is, Y is independent of $\mathbf{x}$ given $\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}, \ldots, \boldsymbol{\beta}_d^{\mathrm{T}}\mathbf{x}$ as d linear combinations of $\mathbf{x}$. Please refer to Cook (1998) for more discussions about the central space, and sufficient dimension reduction in general. For $j = 1, \ldots, p$, let $\beta_{i,j}$ be the jth element of $\boldsymbol{\beta}_i$. The two types of conditional independence $Y \perp\!\!\!\perp \mathbf{x}|\mathbf{x}_\mathcal{A}$ and $Y \perp\!\!\!\perp \mathbf{x}|(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}, \ldots, \boldsymbol{\beta}_d^{\mathrm{T}}\mathbf{x})$ imply that $\sum_{i=1}^d |\beta_{i,j}| > 0$ for $j \in \mathcal{A}$ and $\sum_{i=1}^d |\beta_{i,j}| = 0$ for $j \in \mathcal{A}^c$. In other words, if j belongs to the active set $\mathcal{A}$, then Y must depend on X_j through at least one of the d linear combinations. If j belongs to the inactive set $\mathcal{A}^c$, then none of the d linear combinations involve X_j.

The connection above implies that we may attempt to recover the active set $\mathcal{A}$ through estimation of the central space $\mathcal{S}_{Y|\mathbf{x}}$. SIR [Li (1991)] is one of the most popular methods in the sufficient dimension reduction literature. Suppose $\{J_1, \ldots, J_H\}$ is a measurable partition of the sample space of Y. The main idea of SIR is that the intraslice mean $E(\mathbf{x}|Y \in J_\ell)$ can be used to recover the central space through the relationship that $\mathbf{\Sigma}^{-1}E(\mathbf{x}|Y \in J_\ell) \in \mathcal{S}_{Y|\mathbf{x}}$, $\ell = 1, \ldots, H$. SIR relies on the assumption that $E(\mathbf{x}|\beta_1^{\mathrm{T}}\mathbf{x}, \ldots, \beta_d^{\mathrm{T}}\mathbf{x})$ is linear in $\beta_1^{\mathrm{T}}\mathbf{x}, \ldots, \beta_d^{\mathrm{T}}\mathbf{x}$, which is referred to as the linear conditional mean (LCM) assumption. The LCM assumption is satisfied when $\mathbf{x}$ has an elliptically contoured distribution. Diaconis and Freedman (1984) and Hall and Li (1993) showed that the LCM assumption holds to a reasonable approximation when p increases with d. Following Cook (2004), we further assume the coverage condition:

(C1) $\mathrm{Span}\{\mathbf{\Sigma}^{-1}E(\mathbf{x}|Y \in J_\ell), \ell = 1, \ldots, H\} = \mathcal{S}_{Y|\mathbf{x}}$.

The coverage condition (C1) holds for a wide range of models. We will assume (C1) to facilitate the theoretical derivations of our model-free variable selection procedures.

Let $p_\ell = E\{I(Y \in J_\ell)\}$ denote the probability of Y in the ℓth slice, where $I(\cdot)$ is the indicator function. Information across different slices of Y can be summarized as $\mathbf{\Lambda} = \sum_{\ell=1}^{H} p_\ell E(\mathbf{x}|Y \in J_\ell) E^{\mathrm{T}}(\mathbf{x}|Y \in J_\ell)$. Denote the kernel matrix for SIR as $\mathbf{M} = \mathbf{\Sigma}^{-1}\mathbf{\Lambda}\mathbf{\Sigma}^{-1}$. Assumption (C1) then implies that the column space of $\mathbf{M}$ satisfies $\mathrm{Span}(\mathbf{M}) = \mathfrak{S}_{Y|\mathbf{x}}$. Since $\mathbf{M}$ contains all the regression information between Y and $\mathbf{x}$, it is natural for us to consider the diagonal element of $\mathbf{M}$ as the marginal utility for the corresponding predictor. Specifically, let $\mathbf{e}_k$ be the standard unit vector in $\mathbb{R}^p$ with 1 being the kth element and 0 otherwise. We consider the following utility for X_k:

$$m_k = \mathbf{e}_k^{\mathrm{T}} \mathbf{\Sigma}^{-1} \mathbf{\Lambda} \mathbf{\Sigma}^{-1} \mathbf{e}_k. \tag{2.1}$$

We will refer to m_k as the population level marginal SIR utility. The key property of m_k is summarized in the next result.

Proposition 2.1. *Assume condition* (C1) *holds. Then* $m_k > 0$ *if* $k \in \mathcal{A}$ *and* $m_k = 0$ *if* $k \in \mathcal{A}^c$.

PROOF. Let $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d$ be the basis of $\mathfrak{S}_{Y|\mathbf{x}}$. Under condition (C1), we have $\mathrm{Span}(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d) = \mathrm{Span}(\mathbf{M})$. Also note that $\mathbf{M}$ is positive definite. Thus, $\mathbf{M}$ can be written as $\mathbf{M} = \sum_{i=1}^{d} \delta_i \boldsymbol{\beta}_i \boldsymbol{\beta}_i^{\mathrm{T}}$ for some positive constants δ_i. Let $\beta_{i,j}$ be the jth element of $\boldsymbol{\beta}_i$, $j = 1, \ldots, p$. It follows from (2.1) that $m_k = \sum_{i=1}^{d} \delta_i \beta_{i,k}^2$. If $k \in \mathcal{A}$, then $\beta_{i,k} \neq 0$ for at least one of $i = 1, \ldots, d$, and $m_k > 0$ as a result. Similarly, if $k \in \mathcal{A}^c$, then $\beta_{i,k} = 0$ for all $i = 1, \ldots, d$, and thus $m_k = 0$. □

Proposition 2.1 implies that we may use the sample estimator of m_k to separate the active predictors in $\mathcal{A}$ from the irrelevant predictors in $\mathcal{A}^c$.

For subscript $k \in \mathcal{I}$, denote $\mathbf{x}_{-k} \in \mathbb{R}^{p-1}$ as $(X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_p)^{\mathrm{T}}$. From the definition of the active set $\mathcal{A}$, it is easy to see that $Y \perp\!\!\!\perp \mathbf{x}|\mathbf{x}_{-k}$ if and only if $k \in \mathcal{A}^c$. Proposition 2.1 thus implies that $m_k = 0$ if and only if $Y \perp\!\!\!\perp \mathbf{x}|\mathbf{x}_{-k}$. Thus, we can view m_k as the conditional dependence measure between Y and $\mathbf{x}$ given $\mathbf{x}_{-k}$. Unlike existing marginal screening methods that measure the dependence between Y and X_k without considering $\mathbf{x}_{-k}$, our marginal utility m_k naturally considers the potential correlation among the predictors. Furthermore, consider hypotheses

$$H_0 : Y \perp\!\!\!\perp \mathbf{x}|\mathbf{x}_{-k} \text{ versus } H_a : Y \text{ is not independent of } \mathbf{x} \text{ given } \mathbf{x}_{-k}. \tag{2.2}$$

Then the sample estimator of m_k in (2.1) can be used to construct a test statistic for hypotheses (2.2), which is known as the marginal coordinate test [Cook (2004)] in the sufficient dimension reduction literature.

3. Marginal SIR with the Dantzig selector. To apply Dantzig selector for the estimation of the marginal SIR utility m_k in (2.1), we introduce some notation first. Recall that $p_\ell = E\{I(Y \in J_\ell)\}$, $\ell = 1, \ldots, H$. Let $\mathbf{u}_\ell = E\{\mathbf{x}I(Y \in J_\ell)\}$. Then $\mathbf{\Lambda} = \sum_{\ell=1}^H p_\ell E(\mathbf{x}|Y \in J_\ell)E^{\mathrm{T}}(\mathbf{x}|Y \in J_\ell)$ can be written as $\mathbf{\Lambda} = \sum_{\ell=1}^H \mathbf{u}_\ell \mathbf{u}_\ell^{\mathrm{T}}/p_\ell$. Plugging $\mathbf{\Lambda}$ into (2.1), we obtain

$$m_k = \mathbf{e}_k^{\mathrm{T}}\left(\sum_{\ell=1}^H \boldsymbol{\alpha}_\ell \boldsymbol{\alpha}_\ell^{\mathrm{T}} \Big/ p_\ell\right)\mathbf{e}_k, \qquad \boldsymbol{\alpha}_\ell = \mathbf{\Sigma}^{-1}\mathbf{u}_\ell. \tag{3.1}$$

Under assumption (C1), $\boldsymbol{\alpha}_\ell = p_\ell \mathbf{\Sigma}^{-1} E(\mathbf{x}|Y \in J_\ell) \in \mathfrak{S}_{Y|\mathbf{x}}$. Let $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d$ be the basis of $\mathfrak{S}_{Y|\mathbf{x}}$ and denote $\beta_{i,j}$ as the jth element of $\boldsymbol{\beta}_i$, $j = 1, \ldots, p$. Then $\boldsymbol{\alpha}_\ell \in \mathfrak{S}_{Y|\mathbf{x}}$ can be written as a linear combination of the $\boldsymbol{\beta}_i$'s. Since $\sum_{i=1}^d |\beta_{i,j}| = 0$ for $j \in \mathcal{A}^c$, it follows that the jth element of $\boldsymbol{\alpha}_\ell$ must be zero for $j \in \mathcal{A}^c$. Due to the condition at the outset that $Y \perp\!\!\!\perp \mathbf{x}|\mathbf{x}_\mathcal{A}$, we conclude that $\boldsymbol{\alpha}_\ell$ has at most a nonzero elements, where a denotes the cardinality of $\mathcal{A}$. In other words, $\boldsymbol{\alpha}_\ell$ is intrinsically sparse under the coverage condition (C1). From expression (3.1), we can thus estimate m_k through finding the sparse estimators of $\boldsymbol{\alpha}_\ell = \mathbf{\Sigma}^{-1}\mathbf{u}_\ell$ for $\ell = 1, \ldots, H$.

Given an i.i.d. sample $\{(Y^{(i)}, \mathbf{x}^{(i)}) : i = 1, \ldots, n\}$, we consider the following optimization problem for the estimation of $\boldsymbol{\alpha}_\ell$:

$$\min \|\boldsymbol{\vartheta}\|_1 \text{ such that } \|\hat{\mathbf{\Sigma}}\boldsymbol{\vartheta} - \hat{\mathbf{u}}_\ell\|_\infty \le \varpi_n \text{ and } \boldsymbol{\vartheta} \in \mathbb{R}^p. \tag{3.2}$$

Here, $\hat{\mathbf{\Sigma}} = \sum_{i=1}^n \mathbf{x}^{(i)}(\mathbf{x}^{(i)})^{\mathrm{T}}/n$, $\hat{\mathbf{u}}_\ell = \sum_{i=1}^n \mathbf{x}^{(i)} I(Y^{(i)} \in J_\ell)/n$, and ϖ_n will be specified later. The solution of the constrained optimization problem (3.2) is denoted as $\hat{\boldsymbol{\alpha}}_\ell$. Note that (3.2) has exactly the same form as the original Dantzig selector in Candes and Tao (2007). The only difference is that the original response Y is now replaced by $I(Y \in J_\ell)$. Solving $\boldsymbol{\alpha}_\ell$ from (3.2), we avoid the direct estimation of $\mathbf{\Sigma}^{-1}$ when $n < p$. Define $\mathcal{Y}_\ell \in \mathbb{R}^n$ with the ith element as $I(Y^{(i)} \in J_\ell)$, and define $\mathcal{X} \in \mathbb{R}^{n \times p}$ with the element in the ith row and jth column as $X_j^{(i)}$, the jth element of $\mathbf{x}^{(i)}$. Then we have $\hat{\mathbf{\Sigma}} = \mathcal{X}^{\mathrm{T}}\mathcal{X}/n$, $\hat{\mathbf{u}}_\ell = \mathcal{X}^{\mathrm{T}}\mathcal{Y}_\ell$, and the constraint in (3.2) becomes $\|\mathcal{X}^{\mathrm{T}}\mathcal{X}\boldsymbol{\vartheta} - \mathcal{X}^{\mathrm{T}}\mathcal{Y}_\ell\|_\infty \le n\varpi_n$. The primal–dual interior point algorithm in Candes and Tao (2007) can be used to solve this constrained optimization problem.

After plugging into (3.1) the sample estimators $\hat{p}_\ell = \sum_{i=1}^n I(Y^{(i)} \in J_\ell)/n$ and $\hat{\boldsymbol{\alpha}}_\ell$, the marginal utility m_k is now estimated by

$$\hat{m}_k = \sum_{\ell=1}^H e_k^{\mathrm{T}} \hat{\boldsymbol{\alpha}}_\ell \hat{\boldsymbol{\alpha}}_\ell^{\mathrm{T}} e_k / \hat{p}_\ell. \tag{3.3}$$

For a given threshold γ_n, the active set $\mathcal{A}$ is estimated by including the predictors such that $\hat{m}_k$ exceeds γ_n, or $\hat{\mathcal{A}} = \{k \in \mathcal{I} : \hat{m}_k \ge \gamma_n\}$. The following conditions are needed for the theoretical development about the selection procedure based on $\hat{m}_k$:

(C2) There exist $0 < \varsigma < 1/4$ and $0 < b < \infty$ such that $E\{\exp(tX_k^2)\} \le b$ for all $|t| \le \varsigma$, $k = 1, \ldots, p$. In addition, there exist positive constants $\lambda_{\min}$ and $\lambda_{\max}$ such that $0 < \lambda_{\min} \le \lambda_{\min}(\boldsymbol{\Sigma}) \le \lambda_{\max}(\boldsymbol{\Sigma}) \le \lambda_{\max} < \infty$, where $\lambda_{\min}(\boldsymbol{\Sigma})$ and $\lambda_{\max}(\boldsymbol{\Sigma})$ are the smallest and largest eigenvalues of $\boldsymbol{\Sigma}$, respectively.

(C3) There exists $0 < f < \infty$ such that $\|\boldsymbol{\Sigma}^{-1}\|_1 \le f$.

(C4) There exists $0 < \xi < 1 - 2\phi$ such that $f^2 a^2 \log p = O(n^{\xi})$, where a is the cardinality of $\mathcal{A}$ and ϕ is specified in condition (C5).

(C5) There exist $0 < c < \infty$ and $\phi \le 1/2$ such that $\min_{k \in \mathcal{A}} m_k > 2cn^{-\phi}$.

Condition (C2) requires that $\mathbf{x}$ satisfy the subexponential tail probability uniformly in p. This is a commonly assumed condition in the high dimensional inference literature. See, for example, Wang (2009), Cai and Liu (2011), Cai, Liu and Luo (2011) and Li, Zhong and Zhu (2012). Condition (C3) is commonly used for sparse precision matrix estimation; see Cai, Liu and Luo (2011). Condition (C4) allows the predictor dimension p to diverge at the exponential rate of n. And condition (C3) and condition (C4) together indicate that we also allow the ℓ_1 norm of the precision matrix also diverge to infinity in a certain manner as n and p go to infinity. Condition (C5) is naturally motivated from Proposition 2.1, and requires that the marginal utility m_k for $k \in \mathcal{A}$ cannot be too weak. Conditions similar to (C5) have been used in Fan and Lv (2008), Li, Zhong and Zhu (2012) and Mai and Zou (2015).

Theorem 3.1. *Let* $\varpi_n = \pi_0 a(\log p/n)^{1/2}$, *where* π_0 *is defined in* (A.28) *in the Appendix.*

(a) *Under conditions* (C1), (C2), (C3) *and* (C4),

$$\Pr\Big\{\max_{1 \le k \le p} |\hat{m}_k - m_k| \ge \pi_1 f a(\log p/n)^{1/2}\Big\} \le 4p^{-\tau-2} + 8p^{-\tau-1} + 24p^{-\tau},$$

where $\tau > 0$ *and* π_1 *is defined in* (A.38) *in the Appendix.*

(b) *If, in addition, condition* (C5) *also holds, then with* $\gamma_n = cn^{-\phi}$,

$$\Pr(\mathcal{A} = \hat{\mathcal{A}}) \ge 1 - \big(8p^{-\tau-2} + 16p^{-\tau-1} + 48p^{-\tau}\big).$$

Theorem 3.1 confirms that the marginal SIR with the Dantzig selector achieves the variable selection consistency in the ultrahigh dimensional setting. By exploiting the specific form of m_k in (3.1), our proposal naturally connects the marginal coordinate test in the sufficient dimension reduction literature and Dantzig selector for the purpose of model-free variable selection in the ultrahigh dimensional setting. This result broadens the scope of model-free variable selection via sufficient dimension reduction methods, as the selection consistency of existing methods is established when p is fixed [Bondell and Li (2009); Chen, Zou and Cook (2010)], $p = o(n^{1/4})$ [Wu and Li (2011)] and $p = o(n^{1/2})$ [Jiang and Liu (2014)].

Moreover, we can follow Cai, Liu and Luo (2011) to further relax the exponential-type tails condition (C2) imposed on the predictors $\mathbf{x}$ as the

polynomial-type tails condition. When the predictors have polynomial-type tails, the predictor dimension p can still be much larger than the sample size n. However, the predictor dimension p can only diverge at the polynomial rate of n rather than the exponential rate of n.

4. Marginal SIR with sparse precision matrix estimation. As an alternative to the Dantzig selector approach described in the previous section, the marginal utility m_k can also be estimated if an estimator of $\mathbf{\Sigma}^{-1}$ is available. By combining a regularized estimator of $\mathbf{\Sigma}^{-1}$ and LASSO, Li and Shao (2015) demonstrate that variable selection consistency can be achieved in a linear regression model with ultrahigh dimension. We will extend their result to the model-free setting.

Given an i.i.d. sample $\{(Y^{(i)}, \mathbf{x}^{(i)}) : i = 1, \ldots, n\}$, our idea is to plug in the estimators of $\mathbf{\Sigma}^{-1}$ and $\mathbf{\Lambda}$ to get the sample version of $m_k = \mathbf{e}_k^{\mathrm{T}} \mathbf{\Sigma}^{-1} \mathbf{\Lambda} \mathbf{\Sigma}^{-1} \mathbf{e}_k$. Estimating the precision matrix $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$ in the high-dimensional setting is very challenging. Under some sparsity assumptions, there are several proposals for estimating the sparse precision matrix. See, for example, Bickel and Levina (2008), Friedman, Hastie and Tibshirani (2008), and Fan, Feng and Wu (2009). We adopt the constrained ℓ_1 minimization approach in Cai, Liu and Luo (2011), which has been shown to enjoy desirable theoretical properties. Denote $\mathbf{\Omega} = (\omega_{ij})_{1 \le i,j \le p}$. Let $\hat{\mathbf{\Omega}}_1 = (\hat{\omega}_{ij}^1)_{1 \le i,j \le p}$ be the solution of the following optimization problem:

$$\min \|\mathbf{\Omega}\|_1, \text{ such that } \|\hat{\mathbf{\Sigma}} \mathbf{\Omega} - \mathbf{I}_p\|_\infty \le \varrho_n \text{ and } \mathbf{\Omega} \in \mathbb{R}^{p \times p}.$$

Here, $\hat{\mathbf{\Sigma}} = \sum_{i=1}^n \mathbf{x}^{(i)} (\mathbf{x}^{(i)})^{\mathrm{T}} / n$, and ϱ_n will be specified later. As $\hat{\mathbf{\Omega}}_1$ is generally not symmetric, we get the final estimator $\hat{\mathbf{\Omega}} = (\hat{\omega}_{ij})_{1 \le i,j \le p}$ by symmetrizing $\hat{\mathbf{\Omega}}_1$ as follows:

$$\hat{\omega}_{ij} = \hat{\omega}_{ji} = \hat{\omega}_{ij}^1 I\big(\hat{\omega}_{ij}^1 \le \hat{\omega}_{ji}^1\big) + \hat{\omega}_{ji}^1 I\big(\hat{\omega}_{ji}^1 \le \hat{\omega}_{ij}^1\big).$$

On the other hand, the sample version of $\mathbf{\Lambda} = \sum_{\ell=1}^H \mathbf{u}_\ell \mathbf{u}_\ell^{\mathrm{T}} / p_\ell$ is the classical moment estimator. Recall that $p_\ell = E\{I(Y \in J_\ell)\}$ and $\mathbf{u}_\ell = E\{\mathbf{x} I(Y \in J_\ell)\}$, $\ell = 1, \ldots, H$. The estimator of $\mathbf{\Lambda}$ thus becomes $\hat{\mathbf{\Lambda}} = \sum_{\ell=1}^H \hat{\mathbf{u}}_\ell \hat{\mathbf{u}}_\ell^{\mathrm{T}} / \hat{p}_\ell$, where $\hat{p}_\ell = \sum_{i=1}^n I(Y^{(i)} \in J_\ell)/n$ and $\hat{\mathbf{u}}_\ell = \sum_{i=1}^n \mathbf{x}^{(i)} I(Y^{(i)} \in J_\ell)/n$. After plugging in $\hat{\mathbf{\Omega}}$ and $\hat{\mathbf{\Lambda}}$, we get the sample level marginal utility as

$$\hat{m}_k^* = \mathbf{e}_k^{\mathrm{T}} \hat{\mathbf{\Omega}} \hat{\mathbf{\Lambda}} \hat{\mathbf{\Omega}} \mathbf{e}_k. \tag{4.1}$$

We emphasize that $\hat{m}_k^*$ in (4.1) and $\hat{m}_k$ in (3.3) are different estimators of the same population utility m_k in (2.1). Recall that $\mathcal{I} = \{1, 2, \ldots, p\}$ denotes the full index set. For a given threshold γ_n^*, the active set $\mathcal{A}$ is simply estimated by including the predictors such that $\hat{m}_k^*$ exceeds γ_n^*, or $\hat{\mathcal{A}}^* = \{k \in \mathcal{I} : \hat{m}_k^* \ge \gamma_n^*\}$.

The following conditions are needed before we state the properties of the selection procedure based on $\hat{m}$:

(C3*) There exist $0 < f < \infty$ and $0 < s < \infty$ such that $\|\mathbf{\Omega}\|_1 \le f$ and $\max_{1\le i\le p}\sum_{j=1}^{p}|\omega_{ij}|^q \le s$ for $0 \le q < 1$.

(C4*) There exists $0 < \xi < 1 - 2\phi$ such that $p > n$ and $f^4 a^2 \log p = O(n^\xi)$, where a is the cardinality of $\mathcal{A}$ and ϕ is specified in condition (C5).

Condition (C3*) is commonly made for sparse precision matrix estimation. See, for example, Cai, Liu and Luo (2011). Compared with (C3), (C3*) requires the additional condition that $\mathbf{\Sigma}^{-1}$ is s-sparse in the sense that there are at most s nonzero elements in each row of $\mathbf{\Sigma}^{-1}$. Condition (C4*) also allows the predictor dimension p to diverge at the exponential rate of n, although the rate is slower than that in (C4) if f diverges to infinity.

Theorem 4.1. *Let* $\varrho_n = 2\varsigma^{-2}(2+\tau+\varsigma^{-1}e^2b^2)(\log p/n)^{1/2}$ *for some* $\tau > 0$.

(a) *Under conditions* (C1), (C2), (C3*) *and* (C4*),

$$\Pr\Big\{\max_{1\le k\le p}|\hat{m}_k^* - m_k| \ge \pi_2 f^2 a(\log p/n)^{1/2}\Big\} \le 24p^{-\tau-1} + 8p^{-\tau},$$

where π_2 *is defined in* (A.25) *in the Appendix.*

(b) *If, in addition, condition* (C5) *holds, then with* $\gamma_n^* = cn^{-\phi}$,

$$\Pr(\mathcal{A} = \hat{\mathcal{A}}^*) \ge 1 - \big(48p^{-\tau-1} + 16p^{-\tau}\big).$$

Theorem 4.1 ensures that the active set $\hat{\mathcal{A}}^*$ recovered by $\hat{m}_k^*$ in (4.1) is consistent for the true active set $\mathcal{A}$. Under different sets of conditions, Theorems 3.1 and 4.1 imply that we can estimate the marginal SIR utility either through $\hat{m}_k$ in (3.3) or $\hat{m}_k^*$ in (4.1) and achieve model-free variable selection consistency. By making use of the additional information that $\mathbf{\Sigma}^{-1}$ is sparse, we may have some gain in using $\hat{m}_k^*$ at least in terms of finite sample performance. We will evaluate the finite sample performances of both methods in Section 7.

In some applications $\mathbf{\Sigma}^{-1}$ is not sparse but $\mathbf{\Sigma}$ is. Using a sparse consistent estimator $\hat{\mathbf{\Sigma}}$ such as that in Bickel and Levina (2008), we can obtain a similar $\hat{m}_k^*$ with $\hat{\mathbf{\Omega}}$ in (4.1) replaced by $\hat{\mathbf{\Sigma}}^{-1}$. Selection consistency of this $\hat{m}_k^*$ can be similarly established.

5. Marginal independence SIR and its connections with others. In the seminal sure independence screening paper by Fan and Lv (2008), it was demonstrated that one can construct a marginal utility for feature screening without using the correlation among the predictors. Note that the diagonal elements of $\mathbf{\Sigma}$ are all ones by the assumption that $\mathrm{Var}(X_k) = 1$, $k = 1, \ldots, p$. If we ignore the correlation among the predictors and replace $\mathbf{\Sigma}$ with $\mathbf{I}_p$ in the marginal SIR utility $m_k = \mathbf{e}_k^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{\Lambda}\mathbf{\Sigma}^{-1}\mathbf{e}_k$, we obtain a new marginal utility

$$m_k^I = \sum_{\ell=1}^{H} p_\ell E^2(X_k|Y \in J_\ell), \tag{5.1}$$

which is the kth diagonal element of $\mathbf{\Lambda} = \sum_{\ell=1}^{H} p_\ell E(\mathbf{x}|Y \in J_\ell) E^{\mathrm{T}}(\mathbf{x}|Y \in J_\ell)$. We refer to m_k^I as the marginal independence SIR utility. Note that this utility ignores the correlation so that it can only lead to screening consistency like the SIS in Fan and Lv (2008). On the other hand, it is much simpler to estimate m_k^I compared with the estimation of m_k or m_k^*.

Recall that the SIR kernel matrix is $\mathbf{M} = \mathbf{\Sigma}^{-1}\mathbf{\Lambda}\mathbf{\Sigma}^{-1}$. Consider eigenvalue decomposition $\mathbf{M}\boldsymbol{\nu}_i = \lambda_i \boldsymbol{\nu}_i$, $i = 1, \ldots, p$. Here, $\boldsymbol{\nu}_i \in \mathbb{R}^p$ is the eigenvector corresponding to ith eigenvalue λ_i, and the jth component of $\boldsymbol{\nu}_i$ is denoted as $\nu_{i,j}$, $j = 1, \ldots, p$. Parallel to Proposition 2.1, we reveal the connection between m_k^I and the active set $\mathcal{A}$ in the following result.

Proposition 5.1. *Assume condition* (C1) *holds*, $\mathrm{Cov}(X_i, X_j)$ *has the same sign for* $1 \le i \ne j \le p$, *and* $\mathrm{Var}(X_i) = 1$ *for* $i = 1, \ldots, p$. *Suppose there exists* $\hbar \in \{1, \ldots, d\}$ *such that* $\nu_{\hbar,j}$ *has the same sign for all* $j \in \mathcal{A}$, *then* $m_k^I > 0$.

PROOF. Under the coverage assumption (C1), we have $\mathrm{Span}(\mathbf{M}) = \mathfrak{S}_{Y|\mathbf{x}}$. Since the basis of $\mathfrak{S}_{Y|\mathbf{x}}$ is d-dimensional, we know the rank of $\mathbf{M}$ is d. Thus the eigenvalue decomposition of $\mathbf{M}$ is $\mathbf{M} = \sum_{i=1}^{d} \lambda_i \boldsymbol{\nu}_i \boldsymbol{\nu}_i^{\mathrm{T}}$. From the definition of $\mathbf{M} = \mathbf{\Sigma}^{-1}\mathbf{\Lambda}\mathbf{\Sigma}^{-1}$, we have

$$\mathbf{\Lambda} = \mathbf{\Sigma}\mathbf{M}\mathbf{\Sigma} = \sum_{i=1}^{d} \lambda_i (\mathbf{\Sigma}\boldsymbol{\nu}_i)(\mathbf{\Sigma}\boldsymbol{\nu}_i)^{\mathrm{T}} = \sum_{i=1}^{d} \lambda_i \boldsymbol{\zeta}_i \boldsymbol{\zeta}_i^{\mathrm{T}},$$

where $\boldsymbol{\zeta}_i = \mathbf{\Sigma}\boldsymbol{\nu}_i$ for $i = 1, \ldots, d$. Recall that m_k^I is the kth diagonal element of $\mathbf{\Lambda}$. Thus, we have $m_k^I = \sum_{i=1}^{d} \lambda_i \zeta_{i,k}^2$, where $\zeta_{i,k}$ is the kth component of $\boldsymbol{\zeta}_i$. Note that $\mathbf{M}$ is positive definite and $\lambda_i > 0$ for $i = 1, \ldots, d$. It follows that $m_k^I > 0$ as long as $\zeta_{i,k} \ne 0$ for at least one of $i = 1, \ldots, d$.

It remains to show $\zeta_{\hbar,k} \ne 0$. Note that $\boldsymbol{\nu}_\hbar \in \mathrm{Span}(\mathbf{M}) = \mathfrak{S}_{Y|\mathbf{x}}$. By the definition of the active set $\mathcal{A}$ and the central space $\mathfrak{S}_{Y|\mathbf{x}}$, we have $\nu_{\hbar,j} = 0$ for $j \in \mathcal{A}^c$. Thus, the kth component of $\boldsymbol{\zeta}_\hbar = \mathbf{\Sigma}\boldsymbol{\nu}_\hbar$ becomes

$$\zeta_{\hbar,k} = \sum_{j=1}^{p} \mathrm{Cov}(X_k, X_j)\nu_{\hbar,j} = \sum_{j \in \mathcal{A}} \mathrm{Cov}(X_k, X_j)\nu_{\hbar,j}.$$

Since $\mathrm{Cov}(X_k, X_j)\nu_{\hbar,j}$ has the same sign for all $j \in \mathcal{A}$, we have $\zeta_{\hbar,k} \ne 0$ and $m_k^I > 0$ as a result. □

Consider an i.i.d. sample $\{(Y^{(i)}, \mathbf{x}^{(i)}) : i = 1, \ldots, n\}$, and denote the kth element of $\mathbf{x}^{(i)}$ as $X_k^{(i)}$. The estimator of $\mathbf{\Lambda}$ is $\hat{\mathbf{\Lambda}} = \sum_{\ell=1}^{H} \hat{\mathbf{u}}_\ell \hat{\mathbf{u}}_\ell^{\mathrm{T}} / \hat{p}_\ell$, where $\hat{p}_\ell = \sum_{i=1}^{n} I(Y^{(i)} \in J_\ell)/n$ and $\hat{\mathbf{u}}_\ell = \sum_{i=1}^{n} \mathbf{x}^{(i)} I(Y^{(i)} \in J_\ell)/n$. The sample estimator of m_k^I is the kth diagonal element of $\hat{\mathbf{\Lambda}}$, and can be calculated as

$$\hat{m}_k^I = \sum_{\ell=1}^{H} \hat{u}_{\ell,k}^2 \Big/ \hat{p}_\ell, \tag{5.2}$$

where $\hat{u}_{\ell,k} = \sum_{i=1}^{n} X_k^{(i)} I(Y^{(i)} \in J_\ell)/n$ is the kth element of $\hat{\mathbf{u}}_\ell$. For a given threshold γ_n^I, the active set $\mathcal{A}$ is estimated by including the predictors such that $\hat{m}_k^I$ exceeds γ_n^I, or $\hat{\mathcal{A}}^I = \{k \in \mathcal{I} : \hat{m}_k^I \geq \gamma_n^I\}$.

To study the theoretical property of the marginal independent SIR based independence screening method, we assume the following conditions:

(C6) There exist $0 < \varsigma < 1/4$ and $0 < b < \infty$ such that $E\{\exp(tX_k^2)\} \leq b$ for all $|t| \leq \varsigma$, $k = 1, \ldots, p$.

(C7) There exists $0 < \xi < 1 - 2\phi$ such that $\log p = O(n^\xi)$, where ϕ is specified in condition (C8).

(C8) There exist $0 < c < \infty$ and $\phi \leq 1/2$ such that $\min_{k \in \mathcal{A}} m_k^I > 2cn^{-\phi}$.

Condition (C6) is the first part of (C2). The assumptions about $\mathbf{\Sigma}$ in the second part of (C2) is no longer needed. Condition (C7) is parallel to (C4) and (C4*), and similar conditions have been used in Fan and Lv (2008) and Li et al. (2012). Condition (C8) is parallel to (C5), and we require that the marginal utility m_k^I for $k \in \mathcal{A}$ cannot be too small.

Theorem 5.1. (a) *Under conditions* (C6) *and* (C7),

$$\Pr\Big\{ \max_{1 \leq k \leq p} |\hat{m}_k^I - m_k^I| \geq \pi_3 (\log p/n)^{1/2} \Big\} \leq 8p^{-\tau-1},$$

where $\tau > 0$ *and* π_3 *is defined in* (A.9) *in the Appendix*.

(b) *If, in addition, condition* (C8) *holds, then with* $\gamma_n^I = cn^{-\phi}$,

$$\Pr(\mathcal{A} \subseteq \hat{\mathcal{A}}^I) \geq 1 - 8ap^{-\tau-2}.$$

Theorem 5.1 indicates that we can achieve screening consistency with m_k^I. Under the assumption that X_k and Y are uniformly bounded, Li, Zhong and Zhu (2012) established a similar sure screening property based on the marginal distance correlation utility. Our condition (C6) is weaker, as it only assumes the exponential tail of predictor $\mathbf{x}$.

In the rest of this section, we reveal the connections of marginal independence SIR with some existing popular screening procedures. Recall that existing independence screening methods [Fan and Lv (2008); Huang and Zhu (2014), Li, Zhong and Zhu (2012)] are developed based on dependence measures like Pearson correlation, distance correlation, maximal correlation, etc. The next result states that m_k^I can be also viewed as a dependence measure between $\tilde{Y}$ and X_k, where $\tilde{Y} = \sum_{\ell=1}^{H} \ell I(Y \in J_\ell)$ is the discretized version of Y.

Proposition 5.2. *Let* $T(\tilde{Y})$ *be a transformation of* $\tilde{Y}$ *and denote the Pearson correlation between* $T(\tilde{Y})$ *and* X_k *as* $\mathrm{Corr}\{T(\tilde{Y}), X_k\}$. *Then* m_k^I *in* (5.1) *satisfies* $m_k^I = \max_T \mathrm{Corr}^2\{T(\tilde{Y}), X_k\}$, *where the maximization is over all mapping* $T : \mathbb{R} \mapsto \mathbb{R}$, *and the transformation to get the maximum is* $T(\tilde{Y}) = E(X_k|\tilde{Y})$ *or* $T(\tilde{Y}) = -E(X_k|\tilde{Y})$.

PROOF. Denote $L_k(T) = \text{Corr}^2\{T(\tilde{Y}), X_k\}$. Recall that $E(X_k) = 0$ and $\text{Var}(X_k) = 1$. Thus, $E(X_k^2) = 1$ and we have

$$L_k(T) = \frac{E^2\{T(\tilde{Y})X_k\}}{E\{T^2(\tilde{Y})\}E(X_k^2)} = \frac{E^2\{T(\tilde{Y})X_k\}}{E\{T^2(\tilde{Y})\}}.$$

Plugging in $E\{T(\tilde{Y})X_k\} = E\{T(\tilde{Y})E(X_k|\tilde{Y})\}$, we get

$$L_k(T) = \frac{E^2\{T(\tilde{Y})E(X_k|\tilde{Y})\}}{E\{T^2(\tilde{Y})\}E\{E^2(X_k|\tilde{Y})\}} E\big\{E^2(X_k|\tilde{Y})\big\}.$$

Because $E\{E^2(X_k|\tilde{Y})\} = \text{Var}\{E(X_k|\tilde{Y})\} = m_k^I$, we have

$$L_k(T) = \text{Corr}^2\big\{T(\tilde{Y}), E(X_k|\tilde{Y})\big\} m_k^I \le m_k^I,$$

where the maximum is achieved when $\text{Corr}^2\{T(\tilde{Y}), E(X_k|\tilde{Y})\} = 1$. Thus, the corresponding transformation is $T(\tilde{Y}) = \pm E(X_k|\tilde{Y})$. □

Denote the absolute Pearson correlation between X_k and Y as $m_k^P = |\text{Corr}(X_k, Y)|$. And denote the squared distance correlation [Székely, Rizzo and Bakirov (2007)] between X_k and Y as $m_k^D = \text{dCorr}^2(X_k, Y)$. Feature screening procedures based on m_k^P and m_k^D have been studied in Fan and Lv (2008) and Li, Zhong and Zhu (2012), respectively. We reveal the connections between m_k^I, m_k^P and m_k^D in the next result.

Proposition 5.3. *Suppose condition* (C1) *holds,* $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d$ *is the basis of* $\mathfrak{S}_{Y|\mathbf{x}}$, $\text{Var}(X_k) = \text{Var}(Y) = 1$ *and* $E(X_k) = 0$.

(a) *If* $\mathbf{x}$ *satisfies the LCM condition that* $E(\mathbf{x}|\beta_1^{\mathrm{T}}\mathbf{x}, \ldots, \beta_d^{\mathrm{T}}\mathbf{x})$ *is linear in* $\beta_1^{\mathrm{T}}\mathbf{x}, \ldots, \beta_d^{\mathrm{T}}\mathbf{x}$ *and* $m_k^P > 0$, *then* $m_k^I > 0$.

(b) *If* $(\mathbf{x}, Y)$ *is jointly normally distributed and* $m_k^D > 0$, *then* $m_k^I > 0$.

PROOF. For part (a), because $m_k^I \ge 0$, all we need is to derive a contradiction assuming that $m_k^I = 0$ and $m_k^P > 0$. From Proposition 2.1, we know $m_k^I = 0$ implies $k \in \mathcal{A}^c$ under condition (C1). As a result, $\beta_{i,k} = 0$ for $i = 1, \ldots, d$, where $\beta_{i,k}$ be the kth component of $\boldsymbol{\beta}_i$. Under the LCM assumption, it can be shown that $\boldsymbol{\Sigma}^{-1}E(\mathbf{x}Y) \in \mathfrak{S}_{Y|\mathbf{x}}$. This implies the existence of constants ρ_i, $i = 1, \ldots, d$, such that $E(\mathbf{x}Y) = \boldsymbol{\Sigma}(\sum_{i=1}^d \rho_i \boldsymbol{\beta}_i)$. The kth element of $E(\mathbf{x}Y)$ is thus $E(X_kY) = \boldsymbol{\Sigma}(\sum_{i=1}^d \rho_i \beta_{i,k}) = 0$. Under the assumption that $\text{Var}(X_k) = \text{Var}(Y) = 1$ and $E(X_k) = 0$, we have $m_k^P = |E(X_kY)| = 0$, which contradicts $m_k^P > 0$. The proof of part (a) is completed.

For part (b), since (X_k, Y) is bivariate normal, Theorem 7 in Székely, Rizzo and Bakirov (2007) implies that $m_k^D > 0$ if and only if $m_k^P > 0$. For normally

distributed $\mathbf{x}$, the LCM assumption is satisfied. From the conclusion of part (a), it follows that $m_k^D > 0$ guarantees $m_k^I > 0$. $\square$

Mai and Zou (2015) recently proposed the fused Kolmogorov filter for high dimensional feature screening, which uses the following marginal utility:

$$m_k^F = \max_{1\le \ell\neq \hbar\le H} \sup_{x\in\mathbb{R}} \big|F(X_k \le x|Y\in J_\hbar) - F(X_k\le x|Y\in J_\ell)\big|, \tag{5.3}$$

where $F(X_k|Y)$ is the conditional distribution function of X_k given Y. By noting that $\sum_{\hbar=1}^H p_\hbar = 1$ and $\sum_{\hbar=1}^H p_\hbar E(X_k|Y\in J_\hbar) = 0$, the marginal independence SIR utility in (5.1) can be rewritten as

$$m_k^I = \sum_{\ell=1}^H p_\ell \Bigg[\sum_{\hbar=1}^H p_\hbar \big\{E(X_k|Y\in J_\hbar) - E(X_k|Y\in J_\ell)\big\}\Bigg]^2.$$

There is clear resemblance between m_k^F and m_k^I. While the former measures the maximum difference between the conditional distribution functions $F(X_k|Y)$ across different slices, the latter measures the cumulative difference between the conditional mean $E(X_k|Y)$ across different slices.

The marginal independent SIR ignores the off-diagonal information contained in $\mathbf{\Sigma}$. As a result, the key condition (C8) for marginal independent SIR, which implies that $m_k^I > 0$ for $k\in\mathcal{A}$, might be violated. This is a limitation shared by all the marginal utilities that do not use the full information in $\mathbf{\Sigma}$. The following example illustrates the possible effect of the correlation among predictors on such screening methods.

EXAMPLE 1. Suppose $\mathbf{x} = (X_1, X_2, \ldots, X_p)^{\mathrm{T}} \sim N(0, \mathbf{\Sigma})$. Let $\mathrm{Var}(X_i) = 1$, $\mathrm{Cov}(X_i, X_j) = 0.6$ for $|i-j| = 1$, and $\mathrm{Cov}(X_i, X_j) = 0$ for $|i-j| > 1$, $1 \le i, j \le p$. Let $Y = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} + \varepsilon$, where $\varepsilon \sim N(0,1)$ is independent of $\mathbf{x}$, and $\boldsymbol{\beta} = (1.2, -2, 0, \ldots, 0)^{\mathrm{T}}$. The active set for the linear regression model is $\mathcal{A} = \{1, 2\}$. Consider five utilities for X_1: the marginal absolute Pearson correlation $m_1^P = |\mathrm{Corr}(X_1, Y)|$ from Fan and Lv (2008), the marginal squared distance correlation utility $m_1^D = \mathrm{dCorr}^2(X_1, Y)$ from Li, Zhong and Zhu (2012), the marginal fused Kolmogorov filter utility m_1^F as defined in (5.3), the marginal independence SIR utility m_1^I as defined in (5.1), and the marginal SIR utility m_1 as defined in (2.1). Because $\mathrm{Cov}(X_1, Y) = 0$, we have $m_1^P = 0$. Because (X_1, Y) is bivariate normal, the zero Pearson correlation guarantees that X_1 and Y are independent. It follows that their squared distance correlation $m_1^D = 0$. Because $F(X_1|Y\in J_\hbar) = F(X_1)$ for any $\hbar = 1, \ldots, H$, or the conditional distribution function $F(X_1|Y)$ is the same as the unconditional distribution function $F(X_1)$, we have $F(X_1|Y\in J_\hbar) - F(X_1|Y\in J_\ell) = 0$ and $m_1^F = 0$ as a result. It is easy to see that $m_1^I = 0$ as well. The zero marginal utilities $m_1^P = m_1^D = m_1^F = m_1^I = 0$ imply

that all four independence screening methods will fail to recover the active predictor X_1. The screening consistency result of Theorem 5.1 is no longer applicable here, as condition (C8) is not satisfied. On the other hand, Proposition 2.1 guarantees that $m_1 > 0$, and the procedures based on the marginal SIR estimators in Sections 3 and 4 can still recover X_1.

6. A data-driven threshold. Theorems 3.1, 4.1 and 5.1 show that properly chosen threshold values γ_n, γ_n^* and γ_n^I can lead to desirable theoretical properties. The theoretically optimal threshold values depend on the unknown signal strength ϕ in conditions (C5) and (C8), and will be difficult to determine in practice. We discuss how to choose data-driven threshold values in this section. Since the goal of independence screening is to screen out most of the irrelevant predictors and to retain all the active predictors, it is common practice to be more conservative and retain a larger predictor set. For marginal independence SIR in Section 5, we follow the convention in the independence screening literature, and retain the predictors corresponding to the top $[n/\log n]$ ranked $\hat{m}_k^I$ values, where $[n/\log n]$ is the integer part of $n/\log n$.

For marginal SIR utilities $\hat{m}_k$ in Section 3 and $\hat{m}_k^*$ in Section 4, the final active sets are estimated by $\hat{\mathcal{A}} = \{k \in \mathcal{I} : \hat{m}_k \geq \gamma_n\}$ and $\hat{\mathcal{A}}^* = \{k \in \mathcal{I} : \hat{m}_k^* \geq \gamma_n^*\}$. The goal of marginal SIR is variable selection, where we want to retain all the active predictors and exclude all the inactive predictors simultaneously, and more precise threshold values are needed as a result. We focus on the Dantzig selector based utility $\hat{m}_k$, and the threshold for the sparse precision matrix based utility $\hat{m}_k^*$ can be decided in a similar fashion. Recall that $\tilde{Y} = \sum_{\ell=1}^H \ell I(Y \in J_\ell)$ is the discretized version of Y. Let $\{\gamma_n^g = c_\gamma^g (\log p/n)^{1/2}, g = 1, \ldots, G\}$ be a set of threshold values to choose from. For example, we can take $c_\gamma^g = 0.1g$ for $g = 1, \ldots, 20$ in practice. As SIR with discretized response $\tilde{Y}$ is closely related to linear discriminant analysis [Cook and Yin (2001)], we recommend to perform cross-validation with linear discriminant analysis to choose the optimal threshold value. For $t = 1, \ldots, T$, let $D_1^{(t)}$ and $D_2^{(t)}$ be the tth partition of the sample $\{(\tilde{Y}^{(i)}, \mathbf{x}^{(i)}) : i = 1, \ldots, n\}$. Here $\tilde{Y}^{(i)}$ is the discretized version of $Y^{(i)}$, and takes on integer values from 1 to H. First use $D_1^{(t)}$ as the training data and $D_2^{(t)}$ as the testing data. For fixed threshold value γ_n^g, $g = 1, \ldots, G$, the working active set estimated from $D_1^{(t)}$ is denoted by $\hat{\mathcal{A}}_1^{g,(t)} = \{k \in \mathcal{I} : \hat{m}_k \geq \gamma_n^g\}$. Perform multiclass linear discriminant analysis based on predictors from the working active set $\hat{\mathcal{A}}_1^{g,(t)}$ in the training data $D_1^{(t)}$, and we get a classification rule for $\tilde{Y}^{(i)}$. Apply this classification rule to the testing data $D_2^{(t)}$, and denote the testing classification error as $\Xi_2^{g,(t)}$. Now switch the roles of $D_1^{(t)}$ and $D_2^{(t)}$, where $D_2^{(t)}$ becomes the training data and $D_1^{(t)}$ is the testing data. Repeat the above procedure and we get the testing classification error $\Xi_1^{g,(t)}$. The cumulative classification error over T partitions is

$\Xi^g = \sum_{t=1}^{T}(\Xi_1^{g,(t)} + \Xi_2^{g,(t)})$. The optimal threshold value γ_n^g is then determined by minimizing Ξ^g over $g = 1, \ldots, G$.

7. Numerical studies. In this section, the finite sample performances of the proposed methods are evaluated in simulation studies as well as a real data example.

7.1. *Synthetic examples.* We generate Y from the following four models:

$$\begin{aligned}
\text{I}: &\quad Y = (1.2X_1 - 2X_2 - 2X_{p-1} + 1.2X_p)^3 + \varepsilon,\\
\text{II}: &\quad Y = 1.2X_1 - 1.2X_2 + 0.5X_p + \exp(2X_{p-1} - 1.5X_p)\varepsilon,\\
\text{III}: &\quad Y = \exp(X_1 + X_2 - 2X_{p-1})\,\operatorname{sgn}(X_p) + \varepsilon,\\
\text{IV}: &\quad Y = X_1^3 - 0.5X_2^3 + 3\sin(X_{p-1}) - 3\sin(0.8X_p) + \varepsilon,
\end{aligned}$$

where $\mathbf{x} = (X_1, \ldots, X_p)^{\mathrm{T}}$ and $\varepsilon \sim N(0, 1)$ is independent of $\mathbf{x}$. We consider four ways to generate $\mathbf{x}$. In the first three cases, $\mathbf{x}$ is multivariate normal with mean $\mathbf{0}$ and covariance $\mathbf{\Sigma}$. Denote σ_{ij} as the element in the ith row and jth column of $\mathbf{\Sigma}$. In case (1), $\mathbf{\Sigma} = \mathbf{I}_p$. In case (2), $\sigma_{ij} = 0.6^{|i-j|}$ for $1 \le i, j \le p$. For this case, $\mathbf{\Sigma}$ has a banded structure, and the values of the entries of $\mathbf{\Sigma}$ decay as they move away from the diagonal. In case (3), $\sigma_{ij} = 0.6$ for $1 \le i \ne j \le p$ and $\sigma_{ii} = 1$ for $i = 1, \ldots, p$. This case serves as a dense $\mathbf{\Sigma}$ example. We consider discrete $\mathbf{x}$ in case (4). Let $W \sim \text{Geometric}(0.8)$ with mean $E(W) = 1.25$. Then $X_1, \ldots, X_p$ are i.i.d. with $W - E(W)$.

The performances of five model-free feature screening and selection methods are compared across different model configurations. Our proposals are the marginal independence SIR (I-MSIR), the marginal SIR with sparse precision matrix estimation (SP-MSIR) and the marginal SIR with Dantzig selector (DS-MSIR). We also include two benchmark methods in the model-free feature screening literature: the distance correlation based sure independence screening (DC-SIS) in Li, Zhong and Zhu (2012), and the fused Kolmogorov filter (FKF) in Mai and Zou (2015). For DC-SIS, FKF and I-MSIR, we retain the predictors with the top $[n/\log n]$ ranked marginal utilities, which are the sample estimators of m_k^D, m_k^F and m_k^I, respectively. For SP-MSIR and DS-MSIR, we use the data-driven method in Section 6 to determine the threshold values γ_n and γ_n^*.

We consider sample size $n = 300$ and $p = 1000$. The active predictors in all four models are X_1, X_2, X_{999} and X_{1000}. Based on 100 repetitions, we report the frequencies of each active predictor being selected as f_1, f_2, f_{999} and f_{1000}. A frequency close to one is desirable, as it means that the corresponding active predictor is selected with high frequency. According to the marginal utilities, the average ranks of each active predictor are reported as r_1, r_2, r_{999} and r_{1000}. A small average rank is desirable, as this means that the corresponding active predictor is highly ranked among all predictors. The average selected model size and the

TABLE 1

Results for normal $\mathbf{x}$ *in case* (1). *Based on* 100 *repetitions*, *the frequencies of active predictors being selected*, *the average ranks of active predictors*, *the average selected model size*, *and the average oracle model size are reported*

Model	Method	f_1	f_2	f_{999}	f_{1000}	r_1	r_2	r_{999}	r_{1000}	SMS	OMS
I	DC-SIS	1.00	1.00	1.00	1.00	3.55	1.48	1.54	3.50	52.0	4.07
	FKF	0.99	1.00	1.00	1.00	4.51	1.48	1.52	3.70	52.0	5.19
	I-MSIR	1.00	1.00	1.00	1.00	3.71	1.51	1.49	3.52	52.0	4.23
	SP-MSIR	0.99	1.00	1.00	1.00	3.62	1.50	1.50	3.57	4.31	4.19
	DS-MSIR	1.00	1.00	1.00	1.00	3.51	1.47	1.53	3.57	4.51	4.06
II	DC-SIS	0.81	0.79	1.00	1.00	48.27	46.94	1.09	2.62	52.0	70.98
	FKF	0.99	1.00	1.00	1.00	2.00	1.83	2.66	3.73	52.0	4.20
	I-MSIR	1.00	1.00	1.00	1.00	2.54	2.38	1.74	3.44	52.0	4.08
	SP-MSIR	1.00	1.00	1.00	0.99	2.62	2.40	1.68	3.55	7.92	4.23
	DS-MSIR	1.00	1.00	1.00	1.00	2.81	2.42	1.91	4.26	9.75	5.27
III	DC-SIS	0.79	0.69	1.00	0.93	39.88	62.52	1.01	22.56	52.0	93.02
	FKF	0.96	0.98	1.00	1.00	9.73	7.66	2.00	1.00	52.0	13.50
	I-MSIR	1.00	1.00	1.00	1.00	4.35	4.69	1.05	1.95	52.0	5.99
	SP-MSIR	0.98	0.98	1.00	1.00	4.11	4.60	1.06	1.94	6.32	5.64
	DS-MSIR	0.96	0.93	1.00	1.00	4.58	5.66	1.19	1.81	8.97	7.04
IV	DC-SIS	1.00	1.00	1.00	1.00	1.89	4.19	1.59	2.53	52.0	4.14
	FKF	1.00	0.96	1.00	1.00	2.88	11.98	1.26	1.90	52.0	12.02
	I-MSIR	1.00	1.00	1.00	1.00	1.76	4.28	1.78	2.49	52.0	4.31
	SP-MSIR	1.00	0.98	1.00	1.00	1.74	4.65	1.88	2.40	10.99	4.67
	DS-MSIR	1.00	0.96	1.00	1.00	1.64	6.67	2.01	2.39	8.69	6.71

average oracle model size are reported as SMS and OMS. Here, the selected model size is determined by either $[n/\log n]$ or the data-driven threshold, and the oracle model size is the smallest model size to include all the active predictors. The OMS will be large if one of r_1, r_2, r_{999} and r_{1000} is large. The OMS is always larger than or equal to four, and a value close to four is desirable.

For $\mathbf{x} \sim N(\mathbf{0}, \mathbf{I}_p)$ in case (1), we summarize the results in Table 1. DC-SIS works well for Model I and Model IV, as all four active predictors are selected with frequency one. DC-SIS does not work well for Model II and Model III, where X_1 and X_2 can be missed with large frequency. We confirm from r_1 and r_2 that the average DC-SIS ranks for X_1 and X_2 in Models II and III are very large, and the OMS values for these two models based on DC-SIS are large as well. The other four methods generally work well, with I-MSIR being the most consistent and selecting all active predictors with frequency one across all the models. Our data-driven threshold values for SP-MSIR and DS-MSIR work well, and the average selected model size is generally close to the average oracle model size.

We summarize in Table 2 the results of $\mathbf{x} \sim N(\mathbf{0}, \mathbf{\Sigma})$, where $\sigma_{ij} = 0.6^{|i-j|}$ for $1 \le i, j \le p$. The independence assumption among the predictor no longer holds.

TABLE 2

Results for normal $\mathbf{x}$ *in case* (2). *Based on* 100 *repetitions, the frequencies of active predictors being selected, the average ranks of active predictors, the average selected model size, and the average oracle model size are reported*

Model	Method	f_1	f_2	f_{999}	f_{1000}	r_1	r_2	r_{999}	r_{1000}	SMS	OMS
I	DC-SIS	0.04	1.00	1.00	0.07	488.58	1.40	1.60	520.48	52.0	672.57
	FKF	0.01	1.00	1.00	0.07	549.00	1.46	1.55	510.25	52.0	690.63
	I-MSIR	0.01	1.00	1.00	0.02	531.54	1.39	1.61	526.72	52.0	681.94
	SP-MSIR	0.85	1.00	1.00	0.82	19.36	1.58	1.42	28.46	17.78	38.85
	DS-MSIR	1.00	1.00	1.00	1.00	3.66	1.48	1.52	3.42	4.63	4.08
II	DC-SIS	0.79	0.85	1.00	0.90	48.76	26.66	1.04	16.31	52.0	74.58
	FKF	0.84	0.93	1.00	0.95	37.82	13.76	1.36	14.09	52.0	55.84
	I-MSIR	0.85	0.92	1.00	0.93	34.48	13.98	1.13	14.03	52.0	54.06
	SP-MSIR	1.00	1.00	1.00	0.99	2.67	2.50	1.37	3.98	15.03	4.52
	DS-MSIR	0.98	0.99	1.00	0.96	4.52	3.45	1.99	7.24	4.63	16.81
III	DC-SIS	0.97	0.95	1.00	0.99	8.77	13.72	1.00	3.49	52.0	16.72
	FKF	1.00	1.00	1.00	1.00	4.29	4.09	2.00	1.00	52.0	4.86
	I-MSIR	1.00	1.00	1.00	1.00	3.87	3.85	1.30	1.70	52.0	4.47
	SP-MSIR	1.00	0.99	1.00	1.00	3.57	3.54	1.20	1.80	5.29	4.11
	DS-MSIR	0.99	0.97	1.00	1.00	3.64	4.52	1.33	1.68	6.17	5.16
IV	DC-SIS	1.00	0.94	1.00	1.00	1.07	20.02	2.06	4.14	52.0	20.28
	FKF	1.00	0.17	1.00	0.99	1.91	244.74	1.38	5.46	52.0	245.20
	I-MSIR	1.00	0.10	0.99	0.91	1.07	453.82	3.29	18.29	52.0	455.06
	SP-MSIR	1.00	0.73	1.00	1.00	1.36	34.57	2.09	2.55	16.58	34.57
	DS-MSIR	1.00	0.91	1.00	1.00	1.49	14.52	2.17	2.37	17.95	14.55

As discussed in Example 1 of Section 6, the correlation among the predictors can negatively affect the performances of independence screening methods such as DC-SIS, FKF and I-MSIR. All three independence screening methods cannot select X_1 and X_{1000} in Model I, and have a large frequency to miss X_1 in Model II. FKF and I-MSIR will miss X_2 in Model IV as well. By taking into account the correlation among the predictors, SP-MSIR and DS-MSIR enjoy much better overall performances. Due to the exponentially decaying correlations in the off-diagonal elements, the sparse precision matrix assumption holds in this case, and the decent performance of SP-MSIR is as expected. DS-MSIR has the best overall performances, and the data-driven threshold values work well as before.

For $\mathbf{x} \sim N(\mathbf{0}, \mathbf{\Sigma})$ in case (3), we have $\sigma_{ij} = 0.6$ for $1 \le i \ne j \le p$ and $\sigma_{ii} = 1$ for $i = 1, \ldots, p$. From the results in Table 3, we clearly see that the independence screening methods DC-SIS, FKF and I-MSIR do not work well. This is similar to our findings in Table 2. Moreover, due to the constant correlation among the predictors, the precision matrix is no longer sparse, with the off-diagonal elements of the precision matrix all equal to -0.0025. Thus, we expect SP-MSIR to fail. This is confirmed by the low SP-MSIR frequencies, such as $f_1 = 0.10$ and $f_{1000} =$

TABLE 3

Results for normal $\mathbf{x}$ *in case* (3). *Based on* 100 *repetitions*, *the frequencies of active predictors being selected*, *the average ranks of active predictors*, *the average selected model size and the average oracle model size are reported*

Model	Method	f_1	f_2	f_{999}	f_{1000}	r_1	r_2	r_{999}	r_{1000}	SMS	OMS
I	DC-SIS	0.00	1.00	1.00	0.00	999.38	1.50	1.50	999.44	52.0	1000.0
	FKF	0.00	1.00	1.00	0.00	998.51	1.49	1.51	996.88	52.0	999.96
	I-MSIR	0.00	1.00	1.00	0.00	999.16	1.48	1.52	999.41	52.0	999.99
	SP-MSIR	0.10	1.00	1.00	0.13	190.55	1.54	1.46	207.46	7.45	293.12
	DS-MSIR	0.98	1.00	1.00	0.99	4.17	1.45	1.55	3.92	5.33	4.99
II	DC-SIS	1.00	0.04	1.00	0.57	2.37	703.59	1.25	181.32	52.0	750.07
	FKF	1.00	0.05	0.97	0.53	1.82	751.61	5.73	140.24	52.0	776.07
	I-MSIR	0.00	0.04	1.00	0.58	2.37	774.94	1.95	116.65	52.0	804.12
	SP-MSIR	1.00	0.93	1.00	0.96	1.79	4.55	1.70	4.00	7.15	5.90
	DS-MSIR	0.96	0.94	0.94	0.94	7.26	3.36	3.06	11.90	8.36	19.02
III	DC-SIS	0.03	0.01	1.00	1.00	711.55	677.93	1.20	1.80	52.0	808.44
	FKF	0.10	0.06	1.00	1.00	504.78	457.78	2.00	1.00	52.0	644.95
	I-MSIR	0.04	0.07	1.00	0.00	589.14	527.76	1.53	1.43	52.0	706.09
	SP-MSIR	0.67	0.68	1.00	1.00	19.97	24.97	1.00	2.00	5.78	32.78
	DS-MSIR	0.95	0.95	1.00	1.00	5.35	4.86	1.00	2.05	5.94	7.09
IV	DC-SIS	1.00	0.00	1.00	0.13	1.81	837.09	1.19	651.26	52.0	853.08
	FKF	1.00	0.00	1.00	0.22	2.35	838.80	1.00	512.52	52.0	857.91
	I-MSIR	1.00	0.00	1.00	0.13	1.65	885.48	1.35	672.18	52.0	891.67
	SP-MSIR	1.00	0.43	1.00	0.99	1.72	39.78	1.34	3.16	6.05	39.79
	DS-MSIR	1.00	0.93	1.00	1.00	1.95	10.01	1.77	2.40	8.67	10.12

0.13 in Model I, $f_1 = 0.67$ and $f_2 = 0.68$ in Model III and $f_2 = 0.43$ in Model IV. As it does not rely on the sparse precision matrix estimation, DS-MSIR still works well in this setting.

We summarize in Table 4 the results of discrete $\mathbf{x}$. The results here are similar to the normal independent predictor case presented in Table 1. We see that all the methods generally work well, with the exception of DC-SIS in Model II. We conclude that our proposed methods I-MSIR, SP-MSIR and DS-MSIR still work well with nonnormal predictors. In addition, we observe from Table 1 and Table 4 that the average oracle model size of SP-MSIR is generally smaller than that of DS-MSIR. This observation suggests that when the precision matrix is very sparse, we can potentially benefit from the sparse precision matrix estimation.

The computational costs for the proposed variable selection or screening methods are investigated next. We report in Table 5 the average computation time of each method for Model I based on 100 replications, where n is fixed to be 300 and p is set as 1000, 2000 and 3000. All the computations are done on a Lenovo laptop with 2.4 GHz CPU and 8 GB memory.

TABLE 4

Results for discrete $\mathbf{x}$ *in case* (4). *Based on* 100 *repetitions, the frequencies of active predictors being selected, the average ranks of active predictors, the average selected model size and the average oracle model size are reported*

Model	Method	f_1	f_2	f_{999}	f_{1000}	r_1	r_2	r_{999}	r_{1000}	SMS	OMS
I	DC-SIS	1.00	1.00	1.00	1.00	1.71	3.17	3.43	1.69	52.0	4.00
	FKF	1.00	1.00	1.00	1.00	3.19	1.76	1.94	3.11	52.0	4.00
	I-MSIR	1.00	1.00	1.00	1.00	3.40	1.63	1.68	3.29	52.0	4.00
	SP-MSIR	1.00	1.00	1.00	1.00	3.27	1.72	1.73	3.28	5.16	4.00
	DS-MSIR	1.00	1.00	1.00	1.00	3.22	1.93	1.76	3.09	6.10	4.00
II	DC-SIS	0.82	0.78	1.00	0.39	37.12	47.20	1.09	110.45	52.0	116.42
	FKF	1.00	1.00	1.00	1.00	1.39	1.68	3.30	3.83	52.0	4.20
	I-MSIR	1.00	1.00	1.00	1.00	1.41	1.69	2.95	4.84	52.0	4.89
	SP-MSIR	1.00	1.00	1.00	0.90	1.45	1.65	3.01	5.80	5.79	5.90
	DS-MSIR	1.00	1.00	1.00	0.95	1.41	1.71	3.23	15.27	16.24	15.60
III	DC-SIS	1.00	1.00	0.99	0.94	1.67	1.65	3.96	19.71	52.0	19.71
	FKF	1.00	1.00	1.00	0.78	2.58	2.42	1.00	45.36	52.0	45.36
	I-MSIR	1.00	1.00	1.00	0.97	2.48	2.54	1.00	7.67	52.0	7.69
	SP-MSIR	1.00	1.00	1.00	0.84	2.55	2.43	1.03	6.67	5.58	6.68
	DS-MSIR	1.00	1.00	1.00	0.93	2.45	2.61	1.02	15.53	12.38	15.61
IV	DC-SIS	1.00	1.00	1.00	1.00	2.57	4.00	1.45	1.99	52.0	4.01
	FKF	1.00	1.00	1.00	1.00	3.06	3.47	1.99	1.48	52.0	4.00
	I-MSIR	1.00	1.00	1.00	1.00	2.94	3.85	1.80	1.41	52.0	4.00
	SP-MSIR	1.00	1.00	1.00	1.00	2.98	3.76	1.74	1.52	5.34	4.00
	DS-MSIR	1.00	1.00	1.00	0.95	2.98	3.85	1.73	1.49	5.94	4.05

Among the three independence screening methods DC-SIS, FKF and I-MSIR, we see from Table 5 that our proposed I-MSIR is the fastest in all settings. Between SP-MSIR and DS-MSIR, SP-MSIR is faster when $p = 1000$, but becomes prohibitively slow when $p = 3000$. The computation time of DS-MSIR is very

TABLE 5

Average running time (*in seconds*) *of each method for Model I based on* 100 *replications*

	$p = 1000$				$p = 2000$				$p = 3000$			
Method	Case (1)	Case (2)	Case (3)	Case (4)	Case (1)	Case (2)	Case (3)	Case (4)	Case (1)	Case (2)	Case (3)	Case (4)
DC-SIS	35.2	34.9	34.9	35.0	69.6	69.9	69.9	69.6	105	104	105	104
FKF	11.9	12.1	12.0	11.9	24.5	25.0	24.3	24.1	35.5	35.6	35.6	35.9
I-MSIR	0.521	0.509	0.510	0.514	1.01	1.01	1.00	1.03	1.51	1.50	1.51	1.49
SP-MSIR	19.5	22.1	45.2	18.9	135	156	502	136	488	531	2070	507
DS-MSIR	32.7	31.0	34.7	33.8	41.8	40.9	42.3	42.5	51.4	51.3	53.1	52.2

reasonable. DS-MSIR is not as fast as FKF or I-MSIR, but is faster than DC-SIS. This is very encouraging, as the independence screening methods aim to achieve screening consistency, while DS-MSIR is designed to achieve variable selection consistency.

For all methods, the computation time increases as p increases. For fixed p across the four cases of the distribution of $\mathbf{x}$, the computation time of each method is generally rather stable. The exception here is case (2) and case (3) for the SP-MSIR method. Recall that case (2) corresponds to normal $\mathbf{x}$ with AR type covariance structure, while case (3) corresponds to normal $\mathbf{x}$ with constant correlations among the predictors. As SP-MSIR relies on the sparsity assumption of the precision matrix, our observation here implies that the computation time of the SP-MSIR algorithm depends on the sparseness of the precision matrix. Generally, it takes longer for the SP-MSIR algorithm in the case when the precision matrix is less sparse. The performances of these variable selection methods with $p = 2000$ and 3000 are consistent with the simulation results with $p = 1000$, and thus are omitted here.

7.2. *An application to the small round blue cell tumors classification.* In this section, we apply our proposals to the children cancer data [Khan et al. (2001)] for classifying small round blue cell tumors (SRBCT). The SRBCT data consists of four types of tumor in childhood, including Ewing's sarcoma, rhabdomyosarcoma, neuroblastoma and Burkitt lymphoma. There are 83 tumor samples and the expression measurements on 2308 genes for each sample are provided. We randomly split the SRBCT data into the training set of 55 observations and the testing set of 28 observations. We first perform feature screening and selection based on the training set, build a classification rule with the linear discriminant analysis, and then apply this rule to the testing set. The same five methods are compared as in the previous section. Based on 100 repetitions, Table 6 reports the average training error, the average testing error and the average number of genes selected. We observe that I-MSIR has smaller classification errors than DC-SIS and FKF. Compared with the two existing methods DC-SIS and FKF, SP-MSIR has similar or better testing classification errors with fewer genes selected. DS-MSIR enjoys the best classification performances, and only selects an average of 9 genes out of the 2308 total genes.

TABLE 6
Classification results for the SRBCT data based on 100 *repetitions*

Method	**DC-SIS**	**FKF**	**I-MSIR**	**SP-MSIR**	**DS-MSIR**
Average training error (%)	2.16	5.76	1.33	7.05	0.76
Average testing error (%)	11.00	19.89	6.36	11.07	6.32
Average model size	13	13	13	6.60	9.06

8. Discussions. Marginal independence SIR and marginal SIR are proposed for model-free feature screening and variable selection in this paper. Marginal independence SIR is closely related to existing independence screening methods such as the SIS, distance correlation based SIS and fused Kolmogrov filter. While these independence screening methods share similar theoretical properties, our proposal of marginal independence SIR has better overall finite sample performances. Marginal SIR naturally connects the marginal coordinate tests in the sufficient dimension reduction literature and the ultrahigh dimensional feature screening literature, and opens new avenues for model-free feature selection. Both the sparse precision matrix based and the Dantzig selector based marginal SIR procedures achieve selection consistency in the ultrahigh dimensional setting. As our proposed methods are developed based on the marginal coordinate test wit SIR, they are expected to inherit the limitations of SIR. Although our discussions in this paper focus on SIR only, other popular sufficient dimension reduction methods in the literature like sliced average variance estimation (SAVE) [Cook and Weisberg (1991)] and directional regression (DR) [Li and Wang (2007)] can be considered as well to fix such limitations. Marginal coordinate tests for SAVE and DR have been studied in Shao, Cook and Weisberg (2007) and Yu and Dong (2016). How to extend these procedures in the ultrahigh dimensional setting is worth future investigation.

Another issue for real application is the choice of number of slices when the response is continuous. For the original SIR, Cook and Zhang (2014) proved that combining several slicing schemes works better than the usual practice relying on a single slicing scheme. The combining slicing scheme suggested in Cook and Zhang (2014) motivates us to consider the fused version of marginal SIR. Suppose $m_k(H)$ is the marginal SIR utility (Dantzig selector version or the sparse precision matrix version) with H slices. The fused marginal SIR utility can be defined as $\sum_{H=2}^{T} m_k(H)$, $k = 1, \ldots, p$, where different slicing scheme is combined to get the marginal utility for the kth variable. For the marginal independence SIR utility $m_k^I(H)$, the fused utility can be defined similarly as $\sum_{H=2}^{T} m_k^I(H)$. We can follow Cook and Zhang (2014) and Mai and Zou (2015) to choose $T = [\log n]$. However, the theoretical properties of these fused utilities based on SIR as well as the choice of the tuning parameter T warrant future research.

APPENDIX: PROOFS OF THEOREMS

We first prove Theorem 5.1, and then we provide the proofs of Theorems 4.1 and 3.1.

PROOF OF THEOREM 5.1. For part (a), first note that $|I(Y^{(i)} \in J_\ell) - p_\ell| \le 1$ and $\mathrm{Var}\{I(Y^{(i)} \in J_\ell) - p_\ell\} \le 1/4$. By the Bernstein's inequality [van der Vaart and

Wellner (1996)], we have

$$\Pr\left(\left|\sum_{i=1}^{n}\{I(Y^{(i)}\in J_\ell)-p_\ell\}\right|\ge(2+\tau)(n\log p)^{1/2}\right)$$
$$\le 2\exp\bigl(-(2+\tau)^2 n\log p/[2\{n/4+(2+\tau)(n\log p)^{1/2}/3\}]\bigr)\le 2p^{-\tau-2}.$$

It follows that

$$\Pr\{|\hat p_\ell-p_\ell|\ge(2+\tau)(\log p/n)^{1/2}\}\le 2p^{-\tau-2}. \tag{A.1}$$

Let $p_{\min}=\min\{p_1,\dots,p_H\}$. From $|\hat p_\ell^{-1}-p_\ell^{-1}|=|\hat p_\ell-p_\ell|/p_\ell\hat p_\ell$, we have

$$\begin{aligned}&\Pr\{|\hat p_\ell^{-1}-p_\ell^{-1}|\ge(4+2\tau)p_{\min}^{-2}(\log p/n)^{1/2}\}\\&\quad\le\Pr\{|\hat p_\ell-p_\ell|\ge(2+\tau)(\log p/n)^{1/2}\}+\Pr(p_\ell\hat p_\ell\le p_{\min}^2/2).\end{aligned} \tag{A.2}$$

We have $\log p/n\le p_{\min}^2/(4+2\tau)^2<1/4$ by condition (C7). It follows that

$$\Pr(p_\ell\hat p_\ell\le p_{\min}^2/2)\le 2p^{-\tau-2}. \tag{A.3}$$

(A.1), (A.2) and (A.3) together lead to

$$\Pr\{|\hat p_\ell^{-1}-p_\ell^{-1}|\ge(4+2\tau)p_{\min}^{-2}(\log p/n)^{1/2}\}\le 4p^{-\tau-2}. \tag{A.4}$$

Condition (C6) guarantees $\max_k|u_{\ell,k}|\le\max_k E|X_k|\le 2b^{1/2}$. Together with (A.4), we have

$$\Pr\{|(\hat p_\ell^{-1}-p_\ell^{-1})u_{\ell,k}^2|\ge(16+8\tau)bp_{\min}^{-2}(\log p/n)^{1/2}\}\le 4p^{-\tau-2}. \tag{A.5}$$

From condition (C6), we have

$$\begin{aligned}E\{\exp(tX_k^2I(Y\in J_\ell))\}\le E\{\exp(tX_k^2)\}\le b\qquad&\text{for }|t|\le\varsigma\quad\text{and}\\E\{\exp(t|X_kI(Y\in J_\ell)|)\}\le E\{\exp(|tX_k|)\}\le eb\qquad&\text{for }|t|\le\varsigma.\end{aligned}$$

Let $\pi_4=2+\tau+\varsigma^{-1}e^2b^2$. Following similar arguments in the proof of Theorems 1 and 4 in Cai, Liu and Luo (2011), we have

$$\Pr\{|\hat u_{\ell,k}-u_{\ell,k}|\ge\varsigma^{-1}\pi_4(\log p/n)^{1/2}\}\le 2p^{-\tau-2}. \tag{A.6}$$

Note that $|\hat u_{\ell,k}^2-u_{\ell,k}^2|\le|\hat u_{\ell,k}-u_{\ell,k}|(2|u_{\ell,k}|+|\hat u_{\ell,k}-u_{\ell,k}|)$. Let $\pi_5=p_{\min}^{-1}\varsigma^{-1}\pi_4(4b^{1/2}+\varsigma^{-1}\pi_4/2)$. Then

$$\begin{aligned}&\Pr\{|p_\ell^{-1}(\hat u_{\ell,k}^2-u_{\ell,k}^2)|\ge\pi_5(\log p/n)^{1/2}\}\\&\quad\le\Pr\{|\hat u_{\ell,k}-u_{\ell,k}|\ge\varsigma^{-1}\pi_4(\log p/n)^{1/2}\}\\&\qquad+\Pr\{(2|u_{\ell,k}|+|\hat u_{\ell,k}-u_{\ell,k}|)\ge 4b^{1/2}+\varsigma^{-1}\pi_4(\log p/n)^{1/2}\}.\end{aligned}$$

From (A.6), the first term of the right-hand side of the above inequality is bounded by $2p^{-\tau-2}$. Together with $\max_k |u_{\ell,k}| \le 2b^{1/2}$, the second term of the right-hand side of the above inequality is also bounded by $2p^{-\tau-2}$. Thus, we have

$$\Pr\{|p_\ell^{-1}(\hat{u}_{\ell,k}^2 - u_{\ell,k}^2)| \ge \pi_5(\log p/n)^{1/2}\} \le 4p^{-\tau-2}. \tag{A.7}$$

From the definition of m_k^I and $\hat{m}_k^I$, it can be shown that

$$|\hat{m}_k^I - m_k^I| \le \sum_{\ell=1}^{H}(|(\hat{p}_\ell^{-1} - p_\ell^{-1})u_{\ell,k}^2| + |p_\ell^{-1}(\hat{u}_{\ell,k}^2 - u_{\ell,k}^2)|). \tag{A.8}$$

Define positive constant π_3 as

$$\pi_3 = H\{(16+8\tau)bp_{\min}^{-2} + \pi_5\}. \tag{A.9}$$

Combining (A.5), (A.7) and (A.8), we can derive that

$$\Pr\{|\hat{m}_k^I - m_k^I| \ge \pi_3(\log p/n)^{1/2}\} \le 8p^{-\tau-2}. \tag{A.10}$$

The conclusion of part (a) is then completed by noting that

$$\begin{aligned}&\Pr\Big\{\max_{1\le k\le p}|\hat{m}_k^I - m_k^I| \ge \pi_3(\log p/n)^{1/2}\Big\}\\ &\quad\le p\max_{1\le k\le p}\Pr\{|\hat{m}_k^I - m_k^I| \ge \pi_3(\log p/n)^{1/2}\}.\end{aligned}$$

Now we turn to part (b). If $\mathcal{A} \not\subseteq \hat{\mathcal{A}}^I$, then there must exist some $k \in \mathcal{A}$ such that $\hat{m}_k^I < cn^{-\phi}$. It follows from condition (C8) that $|\hat{m}_k^I - m_k^I| > cn^{-\phi}$ for some $k \in \mathcal{A}$. Recall that a denotes the cardinality of $\mathcal{A}$. Thus,

$$\begin{aligned}\Pr(\mathcal{A} \subseteq \hat{\mathcal{A}}^I) &\ge 1 - \Pr\{|\hat{m}_k^I - m_k^I| > cn^{-\phi} \text{ for some } k \in \mathcal{A}\}\\ &\ge 1 - a\Pr\{|\hat{m}_k - m_k| \ge cn^{-\phi}\}\\ &\ge 1 - a\Pr\{|\hat{m}_k - m_k| \ge cn^{(\xi-1)/2}\},\end{aligned} \tag{A.11}$$

where the last inequality follows from condition (C7). (A.11) together with $\log p = O(n^\xi)$ and (A.10) lead to $\Pr(\mathcal{A} \subseteq \hat{\mathcal{A}}^I) \ge 1 - 8ap^{-\tau-2}$. □

PROOF OF THEOREM 4.1. For part (a), it can be shown that

$$\hat{m}_k^* - m_k = T_{k,1} + 2T_{k,2} + 2T_{k,3} + T_{k,4} + T_{k,5}, \tag{A.12}$$

where

$$\begin{aligned}&T_{k,1} = \mathbf{e}_k^{\mathrm{T}}\boldsymbol{\Omega}(\hat{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda})\boldsymbol{\Omega}\mathbf{e}_k, \qquad T_{k,2} = \mathbf{e}_k^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{\Lambda}(\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega})\mathbf{e}_k,\\ &T_{k,3} = \mathbf{e}_k^{\mathrm{T}}(\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega})\boldsymbol{\Lambda}(\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega})\mathbf{e}_k, \qquad T_{k,4} = \mathbf{e}_k^{\mathrm{T}}(\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega})(\hat{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda})\boldsymbol{\Omega}\mathbf{e}_k \quad \text{and}\\ &T_{k,5} = \mathbf{e}_k^{\mathrm{T}}(\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega})(\hat{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda})(\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega})\mathbf{e}_k.\end{aligned}$$

By the definitions of $\hat{\mathbf{\Lambda}}$ and $\mathbf{\Lambda}$, we have

$$\|\hat{\mathbf{\Lambda}} - \mathbf{\Lambda}\|_\infty = \max_{1\le k,j\le p} \left| \sum_{\ell=1}^{H} \hat{u}_{\ell,k}\hat{u}_{\ell,j}/\hat{p}_\ell - u_{\ell,k}u_{\ell,j}/p_\ell \right|. \tag{A.13}$$

(A.13) together with (A.8), (A.9) and (A.10) in the proof of Theorem 5.1 lead to

$$\Pr\{\|\hat{\mathbf{\Lambda}} - \mathbf{\Lambda}\|_\infty \ge \pi_3 (\log p/n)^{1/2}\} \le 8p^{-\tau-1}. \tag{A.14}$$

Let $\pi_6 = 8\varsigma^{-2}(2+\tau+\varsigma^{-1}e^2b^2)^2$. From Theorem 1 and Theorem 4 in Cai, Liu and Luo (2011), we have

$$\Pr\{\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}\|_\infty \ge \pi_6 f^2 (\log p/n)^{1/2}\} \le 4p^{-\tau}. \tag{A.15}$$

Now we turn to $T_{k,i}$, $i = 1, \ldots, 5$. For $T_{k,1} = \mathbf{e}_k^{\mathrm{T}}\mathbf{\Omega}(\hat{\mathbf{\Lambda}} - \mathbf{\Lambda})\mathbf{\Omega}\mathbf{e}_k$, we have

$$\begin{aligned}&\Pr\Big\{\max_{1\le k\le p} |T_{k,1}| \ge \pi_3 f^2 (\log p/n)^{1/2}\Big\} \\ &\quad\le \Pr\{\|\hat{\mathbf{\Lambda}} - \mathbf{\Lambda}\|_\infty \|\mathbf{\Omega}\|_1^2 \ge \pi_3 f^2 (\log p/n)^{1/2}\}.\end{aligned}$$

From (A.14) and condition (C3), we have

$$\Pr\Big\{\max_{1\le k\le p} |T_{k,1}| \ge \pi_3 f^2 (\log p/n)^{1/2}\Big\} \le 8p^{-\tau-1}. \tag{A.16}$$

Let $\pi_7 = d\lambda_{\min}^{-1}\lambda_{\max}\pi_6$. For $T_{k,2} = \mathbf{e}_k^{\mathrm{T}}\mathbf{\Omega}\mathbf{\Lambda}(\hat{\mathbf{\Omega}} - \mathbf{\Omega})\mathbf{e}_k$, we have

$$\begin{aligned}&\Pr\Big\{\max_{1\le k\le p} |T_{k,2}| \ge \pi_7 f^2 a (\log p/n)^{1/2}\Big\} \\ &\quad\le \Pr\{\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}\|_\infty \|\mathbf{\Omega}\mathbf{\Lambda}\mathbf{\Omega}\mathbf{\Sigma}\|_1 \ge \pi_7 a f^2 (\log p/n)^{1/2}\}.\end{aligned} \tag{A.17}$$

Recall that $\mathbf{M} = \mathbf{\Omega}\mathbf{\Lambda}\mathbf{\Omega}$ has eigenvalue decomposition $\mathbf{M} = \sum_{i=1}^{d} \lambda_i \boldsymbol{\nu}_i \boldsymbol{\nu}_i^{\mathrm{T}}$. It follows that $\nu_{i,k}^2 \le 1$ and $\sum_{j=1}^{p} |\nu_{i,j}| \le a$, where $\nu_{i,j}$ is the jth element of $\boldsymbol{\nu}_i$, $j = 1, \ldots, p$. In addition, we see that $\lambda_i \le \lambda_{\min}^{-1}$ because M can be rewritten as $\mathbf{M} = \mathbf{\Sigma}^{-1/2}\operatorname{Cov}\{E(\mathbf{z}|\tilde{Y})\}\mathbf{\Sigma}^{-1/2}$, where $\mathbf{z} = \mathbf{\Sigma}^{-1/2}\{\mathbf{x} - E(\mathbf{x})\}$ is the standardized predictor. We also have $\boldsymbol{\nu}_i^{\mathrm{T}}\mathbf{\Sigma}^2\boldsymbol{\nu}_i \le \lambda_{\max}^2$ from condition (C2). Thus, we have

$$\|\mathbf{\Omega}\mathbf{\Lambda}\mathbf{\Omega}\mathbf{\Sigma}\|_1 \le \sum_{i=1}^{d} \lambda_i |\boldsymbol{\nu}_i|_1 |\mathbf{\Sigma}\boldsymbol{\nu}_i|_\infty \le \sum_{i=1}^{d} \lambda_i \lambda_{\max} a \le d\lambda_{\min}^{-1}\lambda_{\max} a. \tag{A.18}$$

From (A.15), (A.17) and (A.18), we have

$$\Pr\Big\{\max_{1\le k\le p} |T_{k,2}| \ge \pi_7 f^2 a (\log p/n)^{1/2}\Big\} \le 4p^{-\tau}. \tag{A.19}$$

We now deal with $T_{k,3} = \mathbf{e}_k^{\mathrm{T}}(\hat{\mathbf{\Omega}} - \mathbf{\Omega})\mathbf{\Lambda}(\hat{\mathbf{\Omega}} - \mathbf{\Omega})\mathbf{e}_k$. From Lemma 1 in Cai, Liu and Luo (2011), we know that $\|\hat{\mathbf{\Omega}}\|_1 \le \|\mathbf{\Omega}\|_1 \le f$ and $\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}\|_1 \le 2f$. Together

with (A.18), we have

$$\begin{aligned} T_{k,3} &\le \big\|(\hat{\mathbf{\Omega}}-\mathbf{\Omega})\mathbf{\Sigma}\big\|_\infty \|\mathbf{\Omega}\mathbf{\Lambda}\mathbf{\Omega}\mathbf{\Sigma}\|_1 \|\hat{\mathbf{\Omega}}-\mathbf{\Omega}\|_1 \\ &\le 2d\lambda_{\min}^{-1}\lambda_{\max} f a\big\{\big\|\hat{\mathbf{\Omega}}(\mathbf{\Sigma}-\hat{\mathbf{\Sigma}})\big\|_\infty + \|\hat{\mathbf{\Omega}}\hat{\mathbf{\Sigma}}-\mathbf{I}_p\|_\infty\big\}. \end{aligned}$$

Plug in $\|\mathbf{\Omega}\|_1 \le f$ and the constraint $\|\hat{\mathbf{\Omega}}\hat{\mathbf{\Sigma}}-\mathbf{I}_p\|_\infty \le \varrho_n$, we get

$$T_{k,3} \le 2d\lambda_{\min}^{-1}\lambda_{\max} f a\big(f\|\mathbf{\Sigma}-\hat{\mathbf{\Sigma}}\|_\infty + \varrho_n\big). \tag{A.20}$$

Let $\pi_8 = 8d\lambda_{\min}^{-1}\lambda_{\max}\varsigma^{-2}\pi_4^2$. From equation (28) in Cai, Liu and Luo (2011), we have

$$\Pr\big\{\|\mathbf{\Sigma}-\hat{\mathbf{\Sigma}}\|_\infty \le 2\varsigma^{-2}\pi_4^2(\log p/n)^{1/2}\big\} \le 4p^{-\tau}. \tag{A.21}$$

Plug (A.21) in to (A.20), and we have

$$\Pr\Big\{\max_{1\le k\le p}|T_{k,3}| \ge \pi_8 f^2 a(\log p/n)^{1/2}\Big\} \le 4p^{-\tau}. \tag{A.22}$$

For $T_{k,4} = \mathbf{e}_k^{\mathrm{T}}(\hat{\mathbf{\Omega}}-\mathbf{\Omega})(\hat{\mathbf{\Lambda}}-\mathbf{\Lambda})\mathbf{\Omega}\mathbf{e}_k$, we have $\max_{1\le k\le p}|T_{k,4}| \le \|\hat{\mathbf{\Omega}}-\mathbf{\Omega}\|_1\|(\hat{\mathbf{\Lambda}}-\mathbf{\Lambda})\|_\infty|\mathbf{\Omega}|_1$. From $\|\mathbf{\Omega}\|_1 \le f$ and $\|\hat{\mathbf{\Omega}}-\mathbf{\Omega}\|_1 \le 2f$, we have

$$\begin{aligned} &\Pr\Big\{\max_{1\le k\le p}|T_{k,4}| \ge 2\pi_3 f^2(\log p/n)^{1/2}\Big\} \\ &\quad\le \Pr\big\{\|\hat{\mathbf{\Omega}}-\mathbf{\Omega}\|_1\big\|(\hat{\mathbf{\Lambda}}-\mathbf{\Lambda})\big\|_\infty\|\mathbf{\Omega}\|_1 \ge 2\pi_3 f^2(\log p/n)^{1/2}\big\} \\ &\quad\le \Pr\big\{\big\|(\hat{\mathbf{\Lambda}}-\mathbf{\Lambda})\big\|_\infty \ge \pi_3(\log p/n)^{1/2}\big\}. \end{aligned}$$

Together with (A.14), we have

$$\Pr\Big\{\max_{1\le k\le p}|T_{k,4}| \ge 2\pi_3 f^2(\log p/n)^{1/2}\Big\} \le 8p^{-\tau-1}. \tag{A.23}$$

For $T_{k,5} = \mathbf{e}_k^{\mathrm{T}}(\hat{\mathbf{\Omega}}-\mathbf{\Omega})(\hat{\mathbf{\Lambda}}-\mathbf{\Lambda})(\hat{\mathbf{\Omega}}-\mathbf{\Omega})\mathbf{e}_k$, we have

$$\begin{aligned} &\Pr\Big\{\max_{1\le k\le p}|T_{k,5}| \ge 4\pi_3 f^2(\log p/n)^{1/2}\Big\} \\ &\quad\le \Pr\big\{\|\hat{\mathbf{\Omega}}-\mathbf{\Omega}\|_1\big\|(\hat{\mathbf{\Lambda}}-\mathbf{\Lambda})\big\|_\infty\|\hat{\mathbf{\Omega}}-\mathbf{\Omega}\|_1 \ge 4\pi_3 f^2(\log p/n)^{1/2}\big\} \\ &\quad\le \Pr\big\{\big\|(\hat{\mathbf{\Lambda}}-\mathbf{\Lambda})\big\|_\infty \ge \pi_3(\log p/n)^{1/2}\big\}. \end{aligned}$$

Together with (A.14), we have

$$\Pr\Big\{\max_{1\le k\le p}|T_{k,5}| \ge 4\pi_3 f^2(\log p/n)^{1/2}\Big\} \le 8p^{-\tau-1}. \tag{A.24}$$

Define positive constant π_2 as

$$\pi_2 = 7\pi_3 + 2\pi_7 + 2\pi_8. \tag{A.25}$$

Apply (A.16), (A.19), (A.22), (A.23) and (A.24) to (A.12). Evoke the definition of π_2 in (A.25) and we get the desired result in part (a).

Now we turn to part (b). From the definition $\hat{\mathcal{A}}^* = \{k \in \mathcal{I} : \hat{m}_k^* \geq \gamma_n^*\}$, we have $\mathcal{A} = \hat{\mathcal{A}}^*$ if and only if both $\max_{k\in\mathcal{A}^c} |\hat{m}_k^*| < \gamma_n$ and $\min_{k\in\mathcal{A}} |\hat{m}_k^*| \geq \gamma_n$. It follows that

$$\Pr(\mathcal{A} = \hat{\mathcal{A}}) \geq 1 - \Pr\Big(\max_{k\in\mathcal{A}^c} |\hat{m}_k^*| \geq \gamma_n\Big) - \Pr\Big(\min_{k\in\mathcal{A}} |\hat{m}_k^*| < \gamma_n\Big). \tag{A.26}$$

Because $m_k = 0$ for $k \in \mathcal{A}^c$ from Proposition 2.1, we have

$$\Pr\Big(\max_{k\in\mathcal{A}^c} |\hat{m}_k^*| \geq \gamma_n\Big) = \Pr\Big(\max_{k\in\mathcal{A}^c} |\hat{m}_k^* - m_k| \geq \gamma_n\Big) \leq \Pr\Big(\max_{1\leq k\leq p} |\hat{m}_k^* - m_k| \geq \gamma_n\Big).$$

From condition (C5), we have

$$\Pr\Big(\min_{k\in\mathcal{A}} |\hat{m}_k^*| < \gamma_n\Big) \leq \Pr\Big(\max_{k\in\mathcal{A}} |\hat{m}_k^* - m_k| \geq \gamma_n\Big) \leq \Pr\Big(\max_{1\leq k\leq p} |\hat{m}_k^* - m_k| \geq \gamma_n\Big).$$

Plug the two inequalities above into (A.26) and we get

$$\Pr(\mathcal{A} = \hat{\mathcal{A}}) \geq 1 - 2\Pr\Big(\max_{1\leq k\leq p} |\hat{m}_k^* - m_k| \geq \gamma_n\Big). \tag{A.27}$$

From condition (C4*) and the result of part (a), we have $\Pr(\max_{1\leq k\leq p} |\hat{m}_k^* - m_k| \geq \gamma_n) \leq \Pr(\max_{1\leq k\leq p} |\hat{m}_k^* - m_k| \geq \pi_2 f^2 a \log p/n^{1/2}) \leq 24p^{-\tau-1} + 8p^{-\tau}$. Plug it into (A.27) and we get the desired result in part (b). □

PROOF OF THEOREM 3.1. For part (a), define positive constant π_0 as follows:

$$\pi_0 = 2\varsigma^{-2}\pi_4^2 d^{1/2}\lambda_{\min}^{-1/2} + \varsigma^{-1}\pi_4. \tag{A.28}$$

By the definition of $\boldsymbol{\alpha}_\ell$, we know that $\sum_{\ell=1}^H \boldsymbol{\alpha}_\ell\boldsymbol{\alpha}_\ell^{\mathrm{T}}/p_\ell = \mathbf{M} = \sum_{i=1}^d \lambda_i \boldsymbol{\nu}_i\boldsymbol{\nu}_i^{\mathrm{T}}$. It follows that $\|\boldsymbol{\alpha}_\ell\|_\infty \leq (\sum_{i=1}^d \lambda_i)^{1/2} \max_{1\leq i,j\leq p} |\nu_{i,j}| \leq d^{1/2}\lambda_{\min}^{-1/2}$. Because there are at most a nonzero elements in $\boldsymbol{\alpha}_\ell$, we further have

$$\|\boldsymbol{\alpha}_\ell\|_1 \leq d^{1/2}\lambda_{\min}^{-1/2} a. \tag{A.29}$$

Due to the constraint $\|\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{\alpha}}_\ell - \hat{\mathbf{u}}_\ell\|_\infty \leq \varpi_n$, we have

$$\Pr\big\{\big\|\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\alpha}}_\ell - \boldsymbol{\alpha}_\ell)\big\|_\infty \geq 2\varpi_n\big\} \leq \Pr\{\|\hat{\boldsymbol{\Sigma}}\boldsymbol{\alpha}_\ell - \hat{\mathbf{u}}_\ell\|_\infty \geq \varpi_n\}. \tag{A.30}$$

From (A.6), (A.21) and (A.29), we get

$$\begin{aligned}
&\Pr\{\|\hat{\boldsymbol{\Sigma}}\boldsymbol{\alpha}_\ell - \hat{\mathbf{u}}_\ell\| \geq \varpi_n\} \\
&\quad= \Pr\{\|\hat{\boldsymbol{\Sigma}}\boldsymbol{\alpha}_\ell - \hat{\mathbf{u}}_\ell\| \geq \pi_0 a(\log p/n)^{1/2}\} \\
&\quad\leq \Pr\{\big\|(\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\big\|_\infty \|\boldsymbol{\alpha}_\ell\|_1 \geq 2\varsigma^{-2}\pi_4^2 d^{1/2}\lambda_{\min}^{-1/2} a(\log p/n)^{1/2}\} \\
&\qquad+ \Pr\{\|\hat{\mathbf{u}}_\ell - \mathbf{u}_\ell\|_\infty \geq \varsigma^{-1}\pi_4(\log p/n)^{1/2}\} \leq 4p^{-\tau} + 2p^{-\tau-1}.
\end{aligned} \tag{A.31}$$

It follows from (A.30) and (A.31) that

$$\Pr\{\|\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\alpha}}_\ell - \boldsymbol{\alpha}_\ell)\|_\infty \geq 2\varpi_n\} \leq 4p^{-\tau} + 2p^{-\tau-1}. \tag{A.32}$$

Because $\Pr\{\|(\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})(\hat{\boldsymbol{\alpha}}_\ell - \boldsymbol{\alpha}_\ell)\|_\infty \geq 2\varpi_n\} \leq \Pr\{\|\hat{\boldsymbol{\alpha}}_\ell - \boldsymbol{\alpha}_\ell\|_1 \geq 2d^{1/2}\lambda_{\min}^{-1/2}a\} + \Pr\{\|(\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\|_\infty \geq 2\varsigma^{-2}\pi_2^2(\log p/n)^{1/2}\} \leq (4p^{-\tau} + 2p^{-\tau-1}) + 4p^{-\tau}$, together with (A.32), we have

$$\begin{aligned}
&\Pr\{\|\boldsymbol{\Sigma}(\hat{\boldsymbol{\alpha}}_\ell - \boldsymbol{\alpha}_\ell)\|_\infty \geq 4\varpi_n\} \\
&\leq \Pr\{\|\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\alpha}}_\ell - \boldsymbol{\alpha}_\ell)\|_\infty \geq 2\varpi_n\} \\
&\quad + \Pr\{\|(\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})(\hat{\boldsymbol{\alpha}}_\ell - \boldsymbol{\alpha}_\ell)\|_\infty \geq 2\varpi_n\} = 12p^{-\tau} + 4p^{-\tau-1}.
\end{aligned} \tag{A.33}$$

From (A.33) and $\|\hat{\boldsymbol{\alpha}}_\ell - \boldsymbol{\alpha}_\ell\|_\infty \leq \|\boldsymbol{\Sigma}^{-1}\|_1\|\boldsymbol{\Sigma}(\hat{\boldsymbol{\alpha}}_\ell - \boldsymbol{\alpha}_\ell)\|_\infty$, we have

$$\Pr\{|\hat{\boldsymbol{\alpha}}_\ell - \boldsymbol{\alpha}_\ell|_\infty \geq 4\pi_0 fa(\log p/n)^{1/2}\} \leq 4p^{-\tau-1} + 12p^{-\tau}. \tag{A.34}$$

By the triangular inequality, we have

$$|\hat{m}_k - m_k| \leq \sum_{\ell=1}^{H}(|(\hat{p}_\ell^{-1} - p_\ell^{-1})\boldsymbol{\alpha}_{\ell,k}^2| + |p_\ell^{-1}(\hat{\boldsymbol{\alpha}}_{\ell,k}^2 - \boldsymbol{\alpha}_{\ell,k}^2)|). \tag{A.35}$$

Let $\pi_9 = 8p_{\min}^{-1}\pi_0(d^{1/2}\lambda_{\min}^{-1/2} + \pi_0)$. Then we have

$$\begin{aligned}
&\Pr\{|p_\ell^{-1}(\hat{\alpha}_{\ell,k}^2 - \alpha_{\ell,k}^2)| \geq \pi_9 fa(\log p/n)^{1/2}\} \\
&\leq \Pr\{(2|\alpha_{\ell,k}| + |\hat{\alpha}_{\ell,k} - \alpha_{\ell,k}|) \geq 2d^{1/2}\lambda_{\min}^{-1/2} + 2\pi_0\} \\
&\quad + \Pr\{|\hat{\alpha}_{\ell,k} - \alpha_{\ell,k}| \geq 4\pi_0 fa(\log p/n)^{1/2}\}.
\end{aligned}$$

From condition (C4), we assume that $fa(\log p/n)^{1/2} \leq 1/2$. Together with (A.34), we have

$$\Pr\{|p_\ell^{-1}(\hat{\alpha}_{\ell,k}^2 - \alpha_{\ell,k}^2)| \geq \pi_9 fa(\log p/n)^{1/2}\} \leq 8p^{-\tau-1} + 24p^{-\tau}. \tag{A.36}$$

Let $\pi_{10} = (4 + 2\tau)p_{\min}^{-2}d\lambda_{\min}^{-1}$. Similar to (A.5), we have

$$\Pr\{|(\hat{p}_\ell^{-1} - p_\ell^{-1})\alpha_{\ell,k}^2| \geq \pi_{10}(\log p/n)^{1/2}\} \leq 4p^{-\tau-2}. \tag{A.37}$$

Define positive constant π_1 as

$$\pi_1 = \pi_9 faH + \pi_{10}H. \tag{A.38}$$

(A.35), (A.36), (A.37) and definition of π_1 in (A.38) together lead to the result of part (a).

The proof of part (b) is similar to part (b) of Theorem 4.1, and is thus omitted. □

Acknowledgments. The authors would like to thank the Editor, the Associate Editor and two anonymous referees for giving useful comments that led to a much-improved presentation of the paper.

REFERENCES

BICKEL, P. J. and LEVINA, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.* **36** 199–227. MR2387969

BONDELL, H. D. and LI, L. (2009). Shrinkage inverse regression estimation for model-free variable selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 287–299. MR2655534

CAI, T. and LIU, W. (2011). A direct estimation approach to sparse linear discriminant analysis. *J. Amer. Statist. Assoc.* **106** 1566–1577. MR2896857

CAI, T., LIU, W. and LUO, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.* **106** 594–607. MR2847973

CANDES, E. and TAO, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35** 2313–2351. MR2382644

CHEN, X., ZOU, C. and COOK, R. D. (2010). Coordinate-independent sparse sufficient dimension reduction and variable selection. *Ann. Statist.* **38** 3696–3723. MR2766865

COOK, R. D. (1998). *Regression Graphics*: *Ideas for Studying Regressions Through Graphics*. Wiley, New York. MR1645673

COOK, R. D. (2004). Testing predictor contributions in sufficient dimension reduction. *Ann. Statist.* **32** 1062–1092. MR2065198

COOK, R. D. and WEISBERG, S. (1991). Discussion of "Sliced inverse regression for dimension reduction," by K. Li. *J. Amer. Statist. Assoc.* **86** 328–332.

COOK, R. D. and YIN, X. (2001). Dimension reduction and visualization in discriminant analysis. *Aust. N. Z. J. Stat.* **43** 147–199. MR1839361

COOK, R. D. and ZHANG, X. (2014). Fused estimators of the central subspace in sufficient dimension reduction. *J. Amer. Statist. Assoc.* **109** 815–827. MR3223752

CUI, H., LI, R. and ZHONG, W. (2015). Model-free feature screening for ultrahigh dimensional discriminant analysis. *J. Amer. Statist. Assoc.* **110** 630–641. MR3367253

DIACONIS, P. and FREEDMAN, D. (1984). Asymptotics of graphical projection pursuit. *Ann. Statist.* **12** 793–815. MR0751274

FAN, J., FENG, Y. and WU, Y. (2009). Network exploration via the adaptive lasso and SCAD penalties. *Ann. Appl. Stat.* **3** 521–541. MR2750671

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 849–911. MR2530322

FAN, J., SAMWORTH, R. and WU, Y. (2009). Ultrahigh dimensional variable selection: Beyond the lienar model. *J. Mach. Learn. Res.* **10** 1829–1853.

FAN, J. and SONG, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *Ann. Statist.* **38** 3567–3604. MR2766861

FRIDEMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.

HALL, P. and LI, K. (1993). On almost linearity of low-dimensional projections from high-dimensional data. *Ann. Statist.* **21** 867–889. MR1232523

HUANG, Q. and ZHU, Y. (2014). Model-free sure screening via maximum correlation. Available at arXiv:1403.0048.

JIANG, B. and LIU, J. S. (2014). Variable selection for general index models via sliced inverse regression. *Ann. Statist.* **42** 1751–1786. MR3262467

KHAN, J., WEI, J. S., RINGNÉR, M., SAAL, L. H., LADANYI, M., WESTERMANN, F., BERTHOLD, F., SCHWAB, M., ANTONESCU, C. R., PETERSON, C. and MELTZER, P. S. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Med.* **7** 673–679.

LI, K. (1991). Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.* **86** 316–342. MR1137117

LI, L. (2007). Sparse sufficient dimension reduction. *Biometrika* **94** 603–613. MR2410011

LI, L. and NACHTSHEIM, C. J. (2006). Sparse sliced inverse regression. *Technometrics* **48** 503–510. MR2328619

LI, Q. and SHAO, J. (2015). Regularizing LASSO: A consistent variable selection method. *Statist. Sinica* **25** 975–992. MR3409733

LI, B. and WANG, S. (2007). On directional regression for dimension reduction. *J. Amer. Statist. Assoc.* **102** 997–1008. MR2354409

LI, L. and YIN, X. (2008). Sliced inverse regression with regularizations. *Biometrics* **64** 124–131, 323. MR2422826

LI, R., ZHONG, W. and ZHU, L. (2012). Feature screening via distance correlation learning. *J. Amer. Statist. Assoc.* **107** 1129–1139. MR3010900

LI, G., PENG, H., ZHANG, J. and ZHU, L. (2012). Robust rank correlation based screening. *Ann. Statist.* **40** 1846–1877. MR3015046

MAI, Q. and ZOU, H. (2013). The Kolmogorov filter for variable screening in high-dimensional binary classification. *Biometrika* **100** 229–234. MR3034336

MAI, Q. and ZOU, H. (2015). The fused Kolmogorov filter: A nonparametric model-free screening method. *Ann. Statist.* **43** 1471–1497. MR3357868

NI, L., COOK, D. and TSAI, C. (2005). A note on shrinkage sliced inverse regression. *Biometrika* **92** 242–247. MR2158624

PAN, R., WANG, H. and LI, R. (2015). Ultrahigh dimensional multi-class linear discriminant analysis by pairwise sure independence screening. *J. Amer. Statist. Assoc.* **110** 630–641.

SHAO, Y., COOK, R. D. and WEISBERG, S. (2007). Marginal tests with sliced average variance estimation. *Biometrika* **94** 285–296. MR2331487

SZÉKELY, G. J., RIZZO, M. L. and BAKIROV, N. K. (2007). Measuring and testing dependence by correlation of distances. *Ann. Statist.* **35** 2769–2794. MR2382665

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **58** 267–288. MR1379242

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York. MR1385671

WANG, H. (2009). Forward regression for ultra-high dimensional variable screening. *J. Amer. Statist. Assoc.* **104** 1512–1524. MR2750576

WU, Y. and LI, L. (2011). Asymptotic properties of sufficient dimension reduction with a diverging number of predictors. *Statist. Sinica* **21** 707–730. MR2829852

YIN, X. and HILAFU, H. (2015). Sequential sufficient dimension reduction for large p, small n problems. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 879–892. MR3382601

YU, Z. and DONG, Y. (2016). Model-free coordinate test and variable selection via directional regression. *Statist. Sinica* **26** 1159–1174.

YU, Z., DONG, Y. and ZHU, L.-X. (2016). Trace pursuit: A general framework for model-free variable selection. *J. Amer. Statist. Assoc.* **111** 813–821. MR3538707

YU, Z., ZHU, L., PENG, H. and ZHU, L. (2013). Dimension reduction and predictor selection in semiparametric models. *Biometrika* **100** 641–654. MR3094442

YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 49–67. MR2212574

ZHOU, J. and HE, X. (2008). Dimension reduction based on constrained canonical correlation and variable filtering. *Ann. Statist.* **36** 1649–1668. MR2435451

ZHU, L., LI, L., LI, R. and ZHU, L. (2011). Model-free feature screening for ultrahigh-dimensional data. *J. Amer. Statist. Assoc.* **106** 1464–1475. MR2896849

ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469

Z. YU
SCHOOL OF STATISTICS
EAST CHINA NORMAL UNIVERSITY
SHANGHAI 200241
CHINA
E-MAIL: zyu@stat.ecnu.edu.cn

Y. DONG
DEPARTMENT OF STATISTICS
TEMPLE UNIVERSITY
PHILADELPHIA, PENNSYLVANIA 19122
USA
E-MAIL: ydong@temple.edu

J. SHAO
DEPARTMENT OF STATISTICS
UNIVERSITY OF WISCONSIN-MADISON
MADISON, WISCONSIN 53705
USA
E-MAIL: shao@stat.wisc.edu