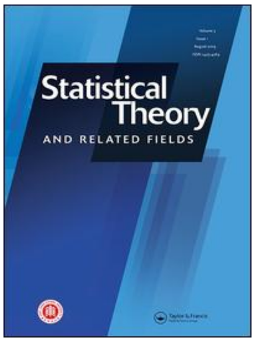

Taylor & Francis
Taylor & Francis Group

Statistical Theory and Related Fields

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tstf20

A selective overview of sparse sufficient dimension reduction

Lu Li , Xuerong Meggie Wen & Zhou Yu

To cite this article: Lu Li , Xuerong Meggie Wen & Zhou Yu (2020): A selective overview of sparse sufficient dimension reduction, Statistical Theory and Related Fields, DOI: 10.1080/24754269.2020.1829389

To link to this article: https://doi.org/10.1080/24754269.2020.1829389

Published online: 10 Nov 2020.

STATISTICAL THEORY AND RELATED FIELDS
https://doi.org/10.1080/24754269.2020.1829389

REVIEW

A selective overview of sparse sufficient dimension reduction

Lu Li[a], Xuerong Meggie Wen[b] and Zhou Yu[a]

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

ABSTRACT

High-dimensional data analysis has been a challenging issue in statistics. Sufficient dimension reduction aims to reduce the dimension of the predictors by replacing the original predictors with a minimal set of their linear combinations without loss of information. However, the estimated linear combinations generally consist of all of the variables, making it difficult to interpret. To circumvent this difficulty, sparse sufficient dimension reduction methods were proposed to conduct model-free variable selection or screening within the framework of sufficient dimension reduction. We review the current literature of sparse sufficient dimension reduction and do some further investigation in this paper.

ARTICLE HISTORY
Received 19 January 2020
Revised 16 September 2020
Accepted 23 September 2020

KEYWORDS
Minimax rate; sparse sufficient dimension reduction; variable selection; variable screening

1. Introduction

The rapid development of data collection technology in areas, such as biology, financial econometrics and signal processing, has posed a great challenge for traditional multivariate analysis. High-dimensional data analysis becomes ubiquitous and increasingly important. Dimension reduction, and in particular sufficient dimension reduction for regression, offers an appealing avenue to tackle high-dimensional problems. It is often desirable to reduce the dimensionality of the problem by replacing the original high-dimensional data with a low-dimensional space composed of a few linear combinations of predictors, which are usually much smaller than the original dimension. Although sufficient dimension reduction is an effective way to extract relevant information from high-dimensional data sets, while grasping the important features or patterns in the data, the linear combinations usually consist of all original predictors which makes the interpretation difficult. This limitation can be overcome via variable selection, where a subset of relevant predictor variables is selected. The removal of the excess variables not only can reduce the noise to the precise estimation, alleviate the collinearity issue, but also help reduce the computational cost caused by high-dimensional data.

As one of the most important dimension reduction approaches, many variable selection methods have been developed. Some most popular variable selection approaches are developed under the linear model or the generalised linear model paradigm, such as nonnegative garrotte (Breiman, 1995), the least absolute shrinkage and selection operator (Lasso, hereafter) (Tibshirani, 1996), the smoothly clipped absolute deviation (SCAD, hereafter) (Fan & Li, 2001), adaptive Lasso (Zou, 2006), group Lasso (Yuan & Lin, 2006), Dantzig selector (Candes & Tao, 2007) and the minimax concave plus penalty (MCP, hereafter) (Zhang, 2010).

These model-based variable selection methods assume the underlying true model is known up to a finite dimensional parameter or the imposed working model is usefully similar to the true model. However, the true model might be in a complex form and it is usually unknown. If the underlying modelling assumption is violated, these variable selection methods might fail. Hence, model-free variable selection method, which does not require the full knowledge of the underlying true model, is called for. It has been shown that the general framework of sufficient dimension reduction is useful for variable selection (Bondell & Li, 2009) since no pre-specified underlying models between the response and the predictors are required. So model-free variable selection can be achieved through the framework of SDR (Cook, 1998; Li, 1991, 2000).

Let $\mathbf{X} = (X_1, \ldots, X_p)^\top$ be the predictor and Y be the scalar response. The goal of variable selection is to seek the smallest subset of the predictors $\mathbf{X}_\mathcal{A}$, with partition $\mathbf{X} = (\mathbf{X}_\mathcal{A}^\top, \mathbf{X}_{\mathcal{A}^c}^\top)^\top$, such that

$$Y \perp\!\!\!\perp \mathbf{X}_{\mathcal{A}^c} \mid \mathbf{X}_\mathcal{A}. \tag{1}$$

Here $\mathcal{A}$ denotes a subset of indices of $\{1, \ldots, p\}$ corresponding to the relevant predictor set $\mathbf{X}_\mathcal{A}$, and $\mathcal{A}^c$ is the complement of $\mathcal{A}$, i.e., $\mathbf{X}_\mathcal{A} = \{x_i : i \in \mathcal{A}\}$ and $\mathbf{X}_{\mathcal{A}^c} = \{x_i : i \in \mathcal{A}^c\}$. Condition (1) implies that $\mathbf{X}_\mathcal{A}$ contains all the active predictors in terms of predicting Y. The existence and uniqueness of $\mathcal{A}$ were discussed in details in Yin and Hilafu (2015). Ideally, we want to find the smallest index set $\mathcal{A}$ satisfying (1), in which case no inactive predictors are included in $\mathbf{X}_{\mathcal{A}^c}$.

CONTACT Zhou Yu ✉ zyu@stat.ecnu.edu.cn

Model-free variable selection is closely related to sufficient dimension reduction, which aims to find $\boldsymbol{\beta} \in \mathbb{R}^{p\times d}$ with $d \le p$, such that

$$Y \perp\!\!\!\perp \mathbf{X} \mid \boldsymbol{\beta}^\top \mathbf{X}, \tag{2}$$

that is, Y is independent of $\mathbf{X}$ conditioning on $\boldsymbol{\beta}^\top \mathbf{X}$. The column space of such $\boldsymbol{\beta}$, $\mathbf{Span}(\boldsymbol{\beta})$, is called a dimension reduction space. Under mild assumptions, such as given in Cook (1996) and Yin et al. (2008), the intersection of all such spaces is itself a dimension reduction space. In this case, we call the intersection the *central subspace* for the regression of Y on $\mathbf{X}$, and denote it by $\mathcal{S}_{Y|\mathbf{X}}$. And its dimension, $d = \dim(\mathcal{S}_{Y\,|\,\mathbf{X}})$, is usually much smaller than p, the dimension of the original predictor. Following the partition of $\mathbf{X}$, we can partition $\boldsymbol{\beta}$ accordingly as

$$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_{\mathcal{A}} \\ \boldsymbol{\beta}_{\mathcal{A}^c} \end{pmatrix}, \quad \boldsymbol{\beta}_{\mathcal{A}} \in \mathbb{R}^{|\mathcal{A}|\times d}, \quad \boldsymbol{\beta}_{\mathcal{A}^c} \in \mathbb{R}^{(p-|\mathcal{A}|)\times d},$$

where $|\mathcal{A}|$ is the cardinality of $\mathcal{A}$. Hence, (1) is equivalent to $\boldsymbol{\beta}_{\mathcal{A}^c} = 0$.

Many methods have been proposed for estimating the basis of $\mathcal{S}_{Y\,|\,\mathbf{X}}$ in the literature, including sliced inverse regression (SIR, hereafter) (Li, 1991), sliced average variance estimation (SAVE, hereafter) (Cook & Weisberg, 1991), principal Hessian directions (PHD, hereafter) (Li, 1992), minimum average variance estimation (MAVE, hereafter) (Xia et al., 2002), directional regression (DR, hereafter) (Li & Wang, 2007), principal fitted component (PFC, hereafter) (Cook & Forzani, 2008), semiparametric approach (Ma & Zhu, 2012), etc. Several methods have been also suggested for simultaneously selecting the contributing predictors. These include shrinkage SIR (Ni et al., 2005), sparse SIR (Li, 2007; Li & Nachtsheim, 2006), sparse SAVE and sparse PHD (Li, 2007), constrained canonical correlation (Zhou & He, 2008), the general shrinkage strategy for inverse regression estimation (Bondell & Li, 2009), the regularised SIR estimator with SCAD penalty (Wu & Li, 2011) and coordinate independent sparse estimation (CISE, hereafter) (Chen et al., 2010), conditional covariance minimisation (Chen et al., 2017), etc.

Although these aforementioned methods can select the significant predictors without assuming an underlying parametric model, they are not designed for $p \gg n$ problems, in which the number of predictor variables is larger than the number of observations. The so-called large p small n problems are increasingly common with rapid technological advances in data collection and have attracted a lot of research interests. We hereby give a very brief review of model-free variable selections via sufficient dimension reduction approach under the $p \gg n$ setting. Li and Yin (2008) proposed sparse ridge SIR, which combined SIR with both ℓ_1- and ℓ_2-regularisation to achieve dimension reduction and variable selection simultaneously, even when $p > n$. Yu et al. (2013) suggested combining SIR with the Dantzig selector (Candes & Tao, 2007) to recover the central subspace in the general semiparametric models. A non-asymptotic error bound for the resulting estimator is derived and the error bound is of order $O_p((\log p/n)^{1/2})$, which appears to be optimal. Moreover, they proposed another regularised version of SIR with the adaptive Dantzig selector. The resulting estimators defined from variable selection are asymptotically normal even when the predictor dimension diverges to infinity. It is worth mentioning that the $|\mathcal{A}|$ is fixed in Yu et al. (2013). Yu, Dong, Zhu (2016) proposed trace pursuit for model-free variable selection under the sufficient dimension reduction paradigm. Two distinct algorithms are proposed: stepwise trace pursuit (STP, hereafter) and forward trace pursuit (FTP, hereafter). Stepwise trace pursuit achieved selection consistency with fixed p and is applicable in the setting with $p > n$. Furthermore, forward trace pursuit can serve as an initial screening step to speed up the computation in the case of ultrahigh dimensionality. Li and Dong (2020) extended trace pursuit method to matrix-valued predictors based on Yu, Dong, Zhu (2016). To test the importance of rows, columns and submatrices of the predictor matrix in terms of predicting the response, three types of hypotheses are formulated under a unified framework. The asymptotic properties of the test statistics under the null hypothesis are established and a permutation testing algorithm is also introduced to approximate the distribution of the test statistics. Tan et al. (2018) developed a convex formulation for fitting sparse SIR in high dimensions. They solved the resulting convex optimisation problem via the linearised alternating direction methods of multiple algorithms and established an upper bound on the subspace distance between the estimated and the true subspaces. Unlike Yu et al. (2013), Lin et al. (2019) allowed $|\mathcal{A}|$ goes to infinity. By constructing artificial response variables made up from top eigenvectors of the estimated conditional covariance matrix, Lin et al. (2019) introduced a simple Lasso regression method to obtain an estimator of the sufficient dimension reduction space. The resulting algorithm, Lasso-SIR, is shown to be consistent and achieves the optimal convergence rate under certain sparsity conditions when p is of order $o(n^2c^2)$, where c is the generalised signal-to-noise ratio, which is only the first step of Tan et al. (2020). Moreover, Tan et al. (2020) discovered the possible trade-off between statistical guarantee and computational performance for sparse SIR and proposed an adaptive estimation scheme for sparse SIR which is computationally tractable and rate optimal under the condition that $\log p = o(n)$, which is weaker than Lin et al. (2019).

There is considerable literature on applying sufficient dimension reduction for model-free selection, but the study of developing screening consistency for the ultra-high dimensional setting is still lacking. To fulfil the aforementioned gaps, Zhu et al. (2011) proposed a variable screening procedure under a unified model framework, which contains a wide variety of commonly used parametric and semiparametric models. The new method does not require imposing a specific model structure on regression functions and thus is particularly appealing to ultrahigh-dimensional regressions. They also showed the proposed method achieves screening consistency even with the number of predictors growing at an exponential rate of the sample size. Yu, Dong, Shao (2016) proposed an approach called marginal SIR for model-free variable selection. Furthermore, marginal SIR with Dantzig selector exploits the sparsity structure in the marginal utility and achieves the desirable selection consistency property. Lin et al. (2017) first introduced a large class of models depending on the smallest non-zero eigenvalue of the kernel matrix of SIR, then the minimax rate for estimating the central space is derived, which is the first paper studied the minimax estimation of sparse SIR. However, they only considered the projection loss (Li & Wang, 2007). More importantly, their theoretical study is based on the assumption that the covariance matrix is diagonal. As far as we know, most of mentioned work mainly focus on SIR with consistency on variable selection. Qian et al. (2019) provided simultaneous analysis for PFC and SAVE. Furthermore, their approach allows many quantities such as the structural dimension, the number of important predictors and the number of slices to diverge with n. To deliver the most essential messages, in the following section, we focus our discussion on the papers mentioned above.

2. Review of sufficient dimension reduction

Sufficient dimension reduction aims to find the column space of $\boldsymbol{\beta}$ with the smallest dimension d. In other words, sufficient dimension reduction is proposed as a problem of estimating a space, instead of the classic statistical problem of estimating parameters. As mentioned in the introduction, there are many approaches in the literature of sufficient dimension reduction for estimating the column space $\boldsymbol{\beta}$: sliced inverse regression (SIR; Li, 1991), sliced average variance estimation (SAVE; Cook & Weisberg, 1991), minimum average variance estimation (MAVE; Xia et al., 2002), the kth moment estimation (Yin & Cook, 2002, 2003), inverse regression (Cook & Ni, 2005), directional regression (DR; Li & Wang, 2007), sliced regression (SR; Wang & Xia, 2008), likelihood acquired directions (LAD; Cook & Forzani, 2009), semiparametric approaches (Ma & Zhu, 2012, 2013a, 2013b, 2014), etc. We mainly review three inverse regression-based methods (SIR; SAVE and DR) for estimating $\mathcal{S}_{Y\mid\mathbf{X}}$ for our subsequent investigation.

Inverse regression methods constitute the oldest class of dimension reduction methods and are still under active development currently. The main idea of the inverse regression is to reverse the relation between the response and the predictors (Li, 1991). Instead of considering distributions or expectations of functions of Y conditional on $\mathbf{X}$, which suffers the curse of dimensionality when $\mathbf{X}$ is high dimensional, these inverse regression-based methods consider expectations of functions of $\mathbf{X}$ conditional on Y, which is suddenly a low dimensional problem because Y is univariate. The inverse regression-based methods often are based on some additional assumptions on the predictors to link the low dimensional problem and the original high dimensional problem. These additional assumptions are given as follows.

(W1) linearity condition $E(\mathbf{X}\mid\boldsymbol{\beta}^\top\mathbf{X}) = \boldsymbol{\Sigma}\boldsymbol{\beta}(\boldsymbol{\beta}^\top\boldsymbol{\Sigma}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^\top\mathbf{X}$.

(W2) constant variance condition $\mathrm{cov}(\mathbf{X}\mid\boldsymbol{\beta}^\top\mathbf{X}) = \boldsymbol{\Sigma} - \boldsymbol{\Sigma}\boldsymbol{\beta}(\boldsymbol{\beta}^\top\boldsymbol{\Sigma}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^\top\boldsymbol{\Sigma}\boldsymbol{\beta}(\boldsymbol{\beta}^\top\boldsymbol{\Sigma}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^\top\boldsymbol{\Sigma}$,

where $\boldsymbol{\Sigma} = \mathrm{cov}(\mathbf{X})$. As is known to all, SIR only requires the condition (W1) holds. However, SAVE and DR need both conditions.

When the linearity condition and the constant variance condition are satisfied, the inverse regression methods formulate the problem of estimating $\mathcal{S}_{Y\mid\mathbf{X}}$ into an eigen-decomposition problem. Let $\mathbf{M}$ be the kernel matrix of a specific inverse regression based dimension reduction method. For the sufficient dimension reduction methods that aim to estimate $\mathcal{S}_{Y\mid\mathbf{X}}$, the kernel matrices corresponding to the three most well-known inverse regression methods are summarised as below:

$$\begin{aligned}
&\text{SIR: } \mathbf{M}_{\mathrm{SIR}} = \mathrm{var}\{E(\mathbf{X}\mid Y)\};\\
&\text{SAVE: } \mathbf{M}_{\mathrm{SAVE}} = E\{\boldsymbol{\Sigma} - \mathrm{var}(\mathbf{X}\mid Y)\}^2;\\
&\text{DR: } \mathbf{M}_{\mathrm{DR}}\\
&\quad = 2E^2\{E(\mathbf{X}\mid Y)E(\mathbf{X}^\top\mid Y)\}\\
&\qquad + 2E\{E(\mathbf{X}^\top\mid Y)E(\mathbf{X}\mid Y)\}E\{E(\mathbf{X}\mid Y)E(\mathbf{X}^\top\mid Y)\}\\
&\qquad + 2E\{E^2(\mathbf{X}\mathbf{X}^\top)\} - 2\boldsymbol{\Sigma}.
\end{aligned}$$

Assuming $d = \mathcal{S}_{Y|\mathbf{X}}$ is known, the procedure for a generalised eigenvalue-decomposition of the kernel matrix $\mathbf{M}$, that is

$$\begin{aligned}
\mathbf{M}\boldsymbol{\beta}_i = \lambda_i\boldsymbol{\Sigma}\boldsymbol{\beta}_i, \quad &\text{with } \boldsymbol{\beta}_i^\top\boldsymbol{\Sigma}\boldsymbol{\beta}_j = 1 \quad \text{if } i = j,\\
&\boldsymbol{\beta}_i^\top\boldsymbol{\Sigma}\boldsymbol{\beta}_j = 0 \quad \text{else } i \neq j,
\end{aligned}$$

where $i = 1,\ldots,p$, and $\lambda_1 \geq \cdots \geq \lambda_d > 0 = \lambda_{d+1} = \cdots = \lambda_p$ are the eigenvalues. Then the eigenvectors corresponding to the nonzero eigenvalues $\boldsymbol{\beta} = (\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_d)$ form a basis of $\mathcal{S}_{Y\mid\mathbf{X}}$. Thus the sufficient dimension reduction directions $\boldsymbol{\beta}$ can also be identified through the following optimisation problem (Tan et al., 2020):

$$\widehat{\boldsymbol{\beta}} = \arg\max_{\mathbf{B}\in\mathbb{R}^{p\times d}} \mathrm{Tr}(\mathbf{B}^\top\mathbf{M}\mathbf{B}) \text{ s.t. } \mathbf{B}^\top\boldsymbol{\Sigma}\mathbf{B} = \mathbf{I}_d. \quad (3)$$

3. The current literature of variable selection via sufficient dimension reduction

3.1. Oracle property under the setting $p < n$

In the general framework of condition (1), the shrinkage SIR method is developed in Ni et al. (2005) by applying the Lasso approach to SIR. When a subset of predictors are irrelevant, then the corresponding row estimates of $\boldsymbol{\beta}$ is equal to 0, and consequently to achieve variable selection. Let $\alpha = (\alpha_1, \ldots, \alpha_p)^\top$, with $\alpha_i \in \mathbb{R}$, $i = 1, \ldots, p$, be the shrinkage vector. Then based on the expression (3), the estimation of the shrinkage vector can be rewritten to minimise the following function over α (Ni et al., 2005):

$$\begin{aligned}\widehat{\alpha} = \arg\min_{\alpha} n(&\mathrm{Vec}(\widehat{\boldsymbol{\beta}}) \\ &- \mathrm{Vec}\{\mathrm{diag}(\alpha)\widehat{\boldsymbol{\beta}}\})^\top\widehat{\mathbf{M}}(\mathrm{Vec}(\widehat{\boldsymbol{\beta}}) - \mathrm{Vec}\{\mathrm{diag}(\alpha)\widehat{\boldsymbol{\beta}}\}), \\ &\text{subject to } \sum_{i=1}^{p} |\alpha_i| \le t, \quad t \ge 0, \qquad (4)\end{aligned}$$

where $\widehat{\mathbf{M}}$ is the estimator of kernel matrix $\mathbf{M}$. To investigate the asymptotic behaviour, we consider the Lagrangian formulation of the constrained optimisation problem. Specially, the optimisation problem in expression (4) can be reformulated as

$$\widehat{\alpha} = \arg\min_{\alpha}(||\mathbf{U} - \mathbf{W}\alpha||^2 + \tau_n \sum_{i=1}^{p} |\alpha_i|),$$

for some non-negative penalty constant τ_n. In which,

$$\mathbf{U} = n^{1/2}\widehat{\mathbf{M}}^{1/2}\mathrm{Vec}(\widehat{\boldsymbol{\beta}}), \quad \mathbf{W} = n^{1/2}\widehat{\mathbf{M}}^{1/2}\widehat{\boldsymbol{\beta}}.$$

Then the central dimension reduction subspace $\mathcal{S}_{Y\,|\,\mathbf{X}}$ is estimated by $\mathbf{Span}\{\mathrm{diag}(\widehat{\boldsymbol{\alpha}})\widehat{\boldsymbol{\beta}}\}$. Li (2007) extended shrinkage SIR method to SAVE and PHD methods, where the central dimension subspace is estimated the same as Ni et al. (2005), and $\widehat{\boldsymbol{\beta}}$ corresponds to the estimated central dimension reduction directions of SAVE and PHD methods, respectively. Bondell and Li (2009) proposed a general shrinkage estimation strategy for the entire inverse regression estimation family that is capable of simultaneous sufficient dimension reduction and variable selection. They considered the adaptive Lasso,

$$\widehat{\alpha} = \arg\min_{\alpha}\left(||\mathbf{U} - \mathbf{W}\alpha||^2 + \tau_n \sum_{i=1}^{p} w_i|\alpha_i|\right),$$

where $\mathbf{w} = (w_1, \ldots, w_p)^\top$ is a known weights vector. They also demonstrated that the proposed class of shrinkage estimators has the desirable oracle property of consistency in variable selection while retaining root n estimation consistency.

However, most existing sparse dimension reduction methods mentioned above are conducted stepwise, estimating a sparse solution for a basis matrix of the central subspace column by column. Instead, Chen et al. (2010) proposed a unified one-step approach to reduce the number of variables appearing in the estimate of $\mathcal{S}_{Y\,|\,\mathbf{X}}$. Their approach, which depends operationally on Grassmann manifold optimisation, can achieve dimension reduction and variable selection simultaneously. Additionally, their proposed method has the oracle property: under mild conditions, the proposed estimator would perform asymptotically as well as if the true irrelevant predictors were known. More importantly, Chen et al. (2010) is an extension to Bondell and Li (2009), which combined SIR, SAVE, DR with adaptive Lasso to variable selection. Zhou and He (2008) proposed a constrained canonical correlation procedure (C^3) based on imposing the L_1-norm constraint on the effective dimension reduction estimates in CANCOR, followed by a simple variable selection method. Using the B-spline basis functions generated for the response variable, the CANCOR method (Fung et al., 2002) is asymptotically equivalent to SIR. Suppose that the range of Y is a bounded interval $[a, b]$, given k_n interval knots in $[a, b]$ and the spline order m, we generate $m + k_n$ B-spline basis functions. Under the linearity condition, CANCOR estimates a set of effective dimension reduction directions by estimating the canonical variates between the B-spline basis functions and $\mathbf{X}$. Since the generated $m + k_n$ B-spline basis functions add to 1, we use in CANCOR the first $m + k_n - 1$ basis function of Y, $\pi(Y) = (\pi_1(Y), \ldots, \pi_{m+k_n-1}(Y))^\top$. Let $\mathcal{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^\top$ and $\Pi_{n\times(m+k_n-1)} = (\pi(Y_1), \ldots, \pi(Y_n))^\top$ be the data matrices containing the predictor values and the B-spline basis function values. Then the CANCOR method is to estimate the canonical correlations between the columns of $\mathcal{X}$ and the columns of Π. The dimensionality of the central dimension reduction subspace is selected by performing the following sequential tests on the number of the non-zero canonical correlations, $H_0 : \gamma_s > \gamma_{s+1} = 0$ versus $H_1 : \gamma_{s+1} > 0$ for $s = 0, 1, \ldots, p - 1$, where γ_s are the asymptotic canonical correlations between $\pi(Y)$ and $\mathbf{X}$ in decreasing order. The dimensionality estimate for d is the smallest s such that H_0 is not rejected. The CANCOR method actually solves an optimisation problem that sequentially finds the directions $\boldsymbol{\beta}$ with the maximum correlation between $\boldsymbol{\beta}^\top\mathbf{X}$ and some functions of Y. Their procedure is attractive because they demonstrated that it also has the oracle property.

Sparse sufficient dimension reduction methods mentioned above focus on the cases when p is fixed. For regressions with diverging p, estimation and variable selection methods are also developed in the framework of sufficient dimension reduction: Zhu et al. (2006) studied the asymptotic properties of SIR as p diverges, but their result is for SIR only, and variable selection is not studied at all. Zhu and Zhu (2009a) investigated weighted partial least squares with a diverging p, but again variable selection is not derived. Zhu and Zhu (2009b) investigated variable selection with a diverging number of predictors through inverse regression, but focused on single-index models only. By contrast, Wu and Li (2011) established asymptotic properties for a family of inverse regression estimators that includes SIR, studied simultaneous dimension reduction and variable selection with a particular emphasis on the latter and encompassed more general forms, while the number of predictors p is allowed to diverge as the sample size n approaches infinity. Wu and Li (2011) adopted the SCAD type penalty that was first introduced by Fan and Li (2001), and combined it with sufficient dimension reduction estimator, that is

$$\widehat{\alpha} = \arg\min_{\alpha} \left(||\mathbf{U} - \mathbf{W}\alpha||^2 + \sum_{i=1}^{p} p_{\tau_n}(|\alpha_i|) \right).$$

The penalty $p_{\tau_n}(\cdot)$ are not necessarily the same for all i. Wu and Li (2011) also showed that the penalised estimator selects all truly contributing predictors and excludes all irrelevant ones with probability approaching one.

Based on the work in kernel dimension reduction, Chen et al. (2017) proposed a method to perform feature selection via a constrained optimisation problem. The corresponding SDR method can refer to Fukumizu et al. (2009); Fukumizu Leng (2014). Many previous kernel approaches are filter methods based on the Hilbert–Schmidt Independence Criterion (HSIC, Gretton et al., 2005). Chen et al. (2017) proposed to use the trace of the conditional covariance operator as a criterion for feature selection. Let (H_1, k_1) denote an RKHS supported on $\mathbf{X} \subset \mathbb{R}^p$. Then the trace of the conditional covariance operator, trace$(\mathbf{\Sigma}_{YY\,|\,\mathbf{X}})$ can be interpreted as a dependence measure, as long as the H_1 is large enough. Then the problem of supervised feature selection reduces to minimising the trace of the conditional covariance operator over subsets of features with controlled cardinality:

$$\min_{T:|T|=d} Q(T) := \mathrm{Tr}(\mathbf{\Sigma}_{YY\,|\,\mathbf{X}_T}).$$

They also showed that empirical estimate of the criterion is consistent as the sample size increases. It is worth noting that kernel feature selection methods have the advantage of capturing nonlinear relationships between the features and the labels.

Theorem 3.1: *Assume* $n^{1/2}\{\mathrm{Vec}(\widehat{\boldsymbol{\beta}}) - \mathrm{Vec}(\boldsymbol{\beta})\} \to \mathcal{N}(0, \Gamma)$, *for some* $\Gamma > 0$, *and that* $\mathbf{M}_n^{1/2} = \mathbf{M}^{1/2} + o(\frac{1}{\sqrt{n}})$. *Suppose that* $\tau_n \to \infty$ *and* $\frac{\tau_n}{\sqrt{n}} \to 0$ *with* $p < n$, *then the shrinkage estimator* $\widehat{\boldsymbol{\beta}}$ *satisfies*

(a) *consistency in variable selection,* $\Pr(\widehat{\mathcal{A}} = \mathcal{A}) \to 1$, *and*
(b) *asymptotic normality,* $n^{1/2}\{\mathrm{Vec}(\widehat{\boldsymbol{\beta}}_{\mathcal{A}}) - \mathrm{Vec}(\boldsymbol{\beta}_{\mathcal{A}})\} \to \mathcal{N}(0, \Lambda)$, *for some* $\Lambda > 0$.

Remark 3.1: Theorem 3.1, part (a), indicates that the sparse sufficient dimension reduction estimator can select contributing predictors consistently, i.e., for all $i \notin \mathcal{A}$ we have $\Pr(\widehat{\alpha}_i \neq 0) \to 0$, and for all $i \in \mathcal{A}$ we have $\Pr(\widehat{\alpha}_i \neq 0) \to 1$. Theorem 3.1, part (b), further shows that the estimator for $\boldsymbol{\beta}_{\mathcal{A}}$ that corresponds to the contributing predictors is root n consistent. The oracle property as shown in Theorem 3.1 is given in Bondell and Li (2009), Chen et al. (2010), Wu and Li (2011) and Zhou and He (2008). Most of the methods mentioned above cannot achieve the desired property with $p > n$, however, Wu and Li (2011) showed that their proposed method can obtain selection consistency when p diverge as the sample size n goes to infinity. Then we turn to investigate the oracle property with $p > n$.

3.2. Oracle property under the setting p ≫ n

Large-p-small-n problems appear frequently in fields such as biology, economics and finance. While those variable selection methods have been successfully applied in many high-dimensional analyses, modern applications in areas such as genomics and high-frequency finance further push the dimensionality of data to an even larger scale, where p may grow exponentially with n. Such ultrahigh-dimensional data present simultaneous challenges of computational expediency, statistical accuracy and algorithm stability. It is difficult to directly apply the aforementioned variable selection methods to those ultrahigh-dimensional statistical learning problems due to the computational complexity inherent in those methods. To reduce the predictor dimension in semiparametric regressions, Yu et al. (2013) proposed a ρ_1-minimisation of SIR with the Dantzig selector (Candes & Tao, 2007), which is defined as

$$\begin{aligned} &\min ||\eta_k||_{\ell_1}, \quad (l = 1, \ldots, k-1) \\ &\text{such that } ||\widehat{\mathbf{M}}\eta_k - \nu_k \widehat{\mathbf{\Sigma}}\eta_k|| \\ &\quad \leq \zeta_k, \quad |\eta_k^\top \widehat{\mathbf{\Sigma}}\eta_k - 1| \leq \zeta_k, \quad |\eta_k^\top \widehat{\mathbf{\Sigma}}\eta_l| \leq \zeta_k, \end{aligned} \tag{5}$$

where $k = 1, \ldots, d$, $\nu_k = \eta_k^\top \widehat{\mathbf{M}}\eta_k$, $|\eta|_{\ell_1} = \sum_{i=1}^{p} |\eta_i|$ and η_0 is a $p \times 1$ zero vector. Furthermore, they established a non-asymptotic error bound for the resulting estimator when $|\mathcal{A}|$ is fixed. Yu et al. (2013) also extended the regularisation concept to SIR with an adaptive Dantzig selector, which is defined by

$$\begin{aligned} &\min ||W_k \eta_k||_{\ell_1}, \\ &\quad \text{such that } ||W_k^{-1}(\widehat{\mathbf{M}}\widehat{\boldsymbol{\beta}}_k^0 - \widehat{\lambda}_k^0 \widehat{\mathbf{\Sigma}}\eta_k)|| \leq \zeta_k, \end{aligned} \tag{6}$$

where $W_k = \mathrm{diag}(w_{k1}, \ldots, w_{kp})$ is the a known weight matrix and w_{kj} is a specified positive value, which should vary inversely with the magnitude of $\widehat{\boldsymbol{\beta}}^0_{kj}$. Yu et al. (2013) proposed a two-step estimation procedure to select the contributing predictors. In the first step, they screened out informative predictors based on (5). This is called Dantzig selector based SIR. In the second step, they enhance the sparsity and the estimation efficiency with (6), based on the predictors selected in the first step, called iterative adaptive Dantzig selector based SIR. This ensures that all contributing predictors are selected with high probability and that the resulting estimator is asymptotically normal even when the predictor dimension diverges to infinity.

However, there is a gap between the optimisation problem and the theoretical results: there is no guarantee that the estimator obtained from solving the proposed biconvex optimisation problem is the global minimum. Most existing work in the high-dimensional sufficient dimension reduction literature involves nonconvex optimisation problems. Moreover, they seek to estimate a set of reduced predictors that are not identifiable by definition, rather than the central subspace. Yin and Hilafu (2015) proposed a sequential approach for estimating high-dimensional SIR. Both proposals are stepwise procedures that do not correspond to solving a convex optimisation problem. Moreover, as discussed in Yin and Hilafu (2015), theoretical properties for their proposed estimators are hard to establish due to the sequential procedure used to obtain the estimators. In the high-dimensional setting, Lin et al. (2018) proposed a screening approach to perform variable selection and established an error bound for the estimators, which allows $|\mathcal{A}|$ goes to infinity. The selected variables are then used to fit classic SIR. Furthermore, the resulting algorithm is shown to be consistent and achieved the optimal convergence rate under certain sparsity conditions when p is of order $o(n^2c^2)$, where c is the generalised signal-to-noise ratio. Tan et al. (2018) proposed a convex formulation for sparse SIR in the high-dimensional setting by adapting techniques from the sparse canonical correlation analysis. Their proposal estimates the central subspace directly and performs variable selection simultaneously. Moreover, the proposed method can be adapted for sufficient dimension reduction methods that can be formulated as generalised eigenvalue problems.

As mentioned in introduction, most literature mainly focus on SIR with consistency on variable selection. Qian et al. (2019) proposed methods under a unified minimum discrepancy framework with regularisation. Consistency results in both central subspace estimation and variable selection are established simultaneously for some famous SDR methods, including SIR, PFC and SAVE. More importantly, their approach allows many quantities such as the structural dimension, the number of important predictors and the number of slices to diverge with n. Unlike many high-dimensional SDR methods, their method did not necessarily require a sparsity condition on the predictor covariance matrix or the maximum eigenvalue of the predictor covariance matrix to be upper bounded. Furthermore, they developed a new algorithm that can efficiently solve a general class of high-dimensional sparse minimum discrepancy problems.

Many SDR methods can be rewritten as a minimisation problem using an objective function of the form

$$L_{1n}(\Gamma, \mathbf{V}) = \mathrm{tr}\left((\gamma_n - \boldsymbol{\Sigma}_n\Gamma\mathbf{V})^\top \Omega_n(\gamma_n - \boldsymbol{\Sigma}_n\Gamma\mathbf{V})\right), \tag{7}$$

where γ_n, $\boldsymbol{\Sigma}_n$ and Ω_n are sample estimates for the population matrices $\widetilde{\mathbf{M}}$, $\mathbf{W}$ and $\mathbf{W}^{-1}$. Here, $\widetilde{\mathbf{M}}$ is a $p \times l$ kernel matrix associated with a particular SDR method, where $d \le l \le p$, $\mathbf{W}$ is some $p \times p$ positive definite matrix, $\Gamma \in \mathbb{R}^{p\times d}$ and $\mathbf{V} \in \mathbb{R}^{d\times l}$ represent parameters to be estimated by minimisation of L_{1n}. The general form of (7) is an adaptation of the minimum discrepancy approach proposed by Cook and Ni (2005). To identify the correct sparsity structure of $\mathcal{S}_{Y\,|\,\mathbf{X}}$ under $p \gg n$ scenarios, Qian et al. (2019) proposed to adopt coordinate-independent regularisation approach and imposed the penalty $P_{\mathbf{V}}(\Gamma)$ with tuning parameter λ_n on (7) under the alternative constraint $\mathbf{V}\mathbf{V}^\top = \mathbf{I}_d$, given the objective function

$$\begin{aligned} L_{2n}(\Gamma, \mathbf{V}) = \tfrac{1}{2}\mathrm{tr}\left((\gamma_n - \boldsymbol{\Sigma}_n\Gamma\mathbf{V})^\top \Omega_n(\gamma_n - \boldsymbol{\Sigma}_n\Gamma\mathbf{V})\right) \\ + \lambda_n P_{\mathbf{V}}(\Gamma), \quad \text{subject to } \mathbf{V}\mathbf{V}^\top = \mathbf{I}_d. \end{aligned}$$

Given its minimiser $(\widehat{\Gamma}, \widehat{\mathbf{V}}) = \arg\min_{\Gamma,\mathbf{V}} L_{2n}(\Gamma, \mathbf{V})$, they simultaneously estimated $\mathcal{S}_{Y\,|\,\mathbf{X}}$ by $\mathbf{Span}(\Gamma)$ and estimated $\mathcal{A}_0$ by $\widehat{\mathcal{A}}_0 = \{1 \le j \le p : e_j^\top \widehat{\Gamma}\widehat{\Gamma}^\top e_j > 0\}$.

Tan et al. (2020) considered four loss functions

(i) General loss. $L_G(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = ||\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}^\top - \boldsymbol{\beta}\boldsymbol{\beta}^\top||$;
(ii) Projection loss. $L_P(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = ||\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\beta}}^\top\widehat{\boldsymbol{\beta}})^{-1}\widehat{\boldsymbol{\beta}}^\top - \boldsymbol{\beta}(\boldsymbol{\beta}^\top\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^\top||_F^2$;
(iii) Prediction loss. $L_X(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = \inf_{W\in\mathbb{R}^{p\times p}} ||\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}W)||_F^2$;
(iv) Correlation loss. $L_C(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = 1 - \frac{1}{d}\mathrm{Tr}[(\widehat{\boldsymbol{\beta}}^\top\boldsymbol{\Sigma}\widehat{\boldsymbol{\beta}})^{-1}(\widehat{\boldsymbol{\beta}}^\top\boldsymbol{\Sigma}\boldsymbol{\beta})(\boldsymbol{\beta}^\top\boldsymbol{\Sigma}\boldsymbol{\beta})^{-1}(\boldsymbol{\beta}^\top\boldsymbol{\Sigma}\widehat{\boldsymbol{\beta}})]$,

where $||\cdot||_F$ denotes the Frobenius norm of a matrix. Further, Tan et al. (2020) established the minimax lower bound for sparse SIR under general loss, projection loss and prediction loss. They proposed natural sparse SIR estimator and proved that the upper error bound associated with all four loss functions can match the minimax lower bound obtained, which implies that it is a rate-optimal estimator for sparse SIR. However, this optimal estimation is computational intractable. Then they developed the computational feasible counterpart for this natural sparse SIR estimator through convex relaxation. But their theoretical investigation suggested that such computational realisation for natural sparse SIR estimator cannot maintain the optimal estimation rate.

To further address this issue, they proposed a refined sparse SIR estimator. The refined sparse SIR estimator is also rate-optimal yet computational intractable. However, its computational feasible counterpart based on the adaptive estimation procedure is proven to be nearly rate-optimal. Compared to the Lasso-SIR (Lin et al., 2019), which was shown to be rate optimal only when $p = o(n^2)$, their sparse SIR approach is rate optimal even when $\log p = o(n)$. Therefore, their proposed sparse SIR estimator certainly enjoys a much wider range of applications. The reason why Lasso-SIR fails to work when $\log p = o(n)$ is that it requires the estimation of the eigenvalues and eigenvectors of the $p \times p$ non-sparse SIR kernel matrix. It is well known that the sample eigenvalues and eigenvectors are not even consistent when p/n has a nonzero limit as $n \to \infty$. In summary, the minimax lower bound obtained, the two rate-optimal yet computational infeasible estimators, the two corresponding computational tractable counterparts, and the theoretical upper bound of the four estimators under four-loss functions together provide a thorough understanding of sparse SIR. It is also worth noting that Lin et al. (2019) is just the first step of Tan et al. (2020). Bondell and Li (2009) demonstrated that $\mathrm{Supp}(\boldsymbol{\beta}) = \mathcal{A}$, then the sparse representation of SIR relies on $|\mathcal{A}|$, the number of truly relevant predictors, where $\mathrm{Supp}(\boldsymbol{\beta})$ denotes the support of $\boldsymbol{\beta}$. Assuming $|\mathcal{A}| \le s$, sparse SIR is further defined through seeking $\boldsymbol{\beta}$ such that

$$\begin{aligned}\boldsymbol{\beta} &= \operatorname*{arg\,max}_{\mathbf{B}\in\mathbb{R}^{p\times d}} \mathrm{Tr}(\mathbf{B}^\top \mathbf{M}\mathbf{B}) \quad \text{s.t. } \mathbf{B}^\top \boldsymbol{\Sigma}\mathbf{B} \\ &= \mathbf{I}_d \text{ and } |\mathrm{Supp}(\mathbf{B})| \le s. \qquad (8)\end{aligned}$$

The above formulation of sparse SIR enjoys a similar fashion as that of sparse CCA (Gao et al., 2015). To get theoretical results, the following conditions are required.

(A1) the conditional mean $E\{(\mathbf{X} - E(\mathbf{X})|\boldsymbol{\beta}^\top \mathbf{X})\}$ is linear in $\mathbf{X}$;
(A2) $|\mathcal{A}_k|$ is bounded for $k = 1, \dots, d$;
(A3) the nonzero eigenvalues $\lambda_1, \dots, \lambda_d$ are distinct;
(A4) there exists a positive constant a_0 such that $0 < a_0 < 1/4, \log p/n \le a_0$ and $E(\exp[t\{\mathbf{X}_i - E(\mathbf{X}_i)\}^2]) \le K < \infty$, for $i = 1, \dots, p$, and all $|t| \le a_0$;
(A5) $D_0 = \max_{1\le i\le p} \sum_{j=1}^p |\sigma_{ij}|$ is bounded constant as $p \to \infty$;
(A6) the restricted isometry and restricted orthogonality constants $\delta_{2S_k}^{A^k}$ and $\theta_{S_k,2S_k}^{A^K}$ satisfy $\delta_{2S_k}^{A^k} + \theta_{S_k,2S_k}^{A^K} < 1$, where $A^k = \mathbf{M} - \lambda_k \boldsymbol{\Sigma}$.

See Yu et al. (2013) for more details.

Theorem 3.2: *Suppose that Conditions A1–A6 are satisfied, and $\zeta_k = C_0(\log p/n)^{1/2}$. Then*

$$||\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k||^2 \le \frac{4C^2 S_K}{(1 - \delta_{2S_k}^{A^k} - \theta_{S,2S}^{A^K})^2} \frac{\log p}{n},$$

with a probability greater than $1 - 58p^{-\tau}$ for some τ greater than $(\log p)^{-1} \log 58$, where $A^k, \delta_{2S_k}^{A^k}$ and $\theta_{S,2S}^{A^K}$ are defined in Condition A6.

Remark 3.2: Theorem 3.2 suggests that a small price can obtain a sparse solution, as the squared estimation error of the regularised estimation is optimal up to a factor of $\log p$. The consistency property of Lin et al. (2018), Tan et al. (2018) and Tan et al. (2020) are similar with Theorem 3.2, but the threshold for $||\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k||^2$ can be different.

4. The current literature of variable screening

Although there is a vast literature of applying sufficient dimension reduction for model-free selection, the result of developing screening consistency for the ultra-high dimensional setting is scant. Therefore many scholars are concentrated on investigating methods to achieve screening consistency.

4.1. Marginal utility

Yu, Dong, Shao (2016) proposed an approach called marginal SIR for model-free variable selection. Since $\mathbf{M}$ contains all the regression information between Y and $\mathbf{X}$, Yu, Dong, Shao (2016) considered the diagonal element of $\mathbf{M}$ as the marginal utility for the corresponding predictor. Specially, let e_k be the standard unit vector in $\mathbb{R}^p$ with 1 being the kth element and 0 otherwise. They considered the following utility for $\mathbf{X}_k$:

$$m_k = e_k^\top \boldsymbol{\Sigma}^{-1} \mathbf{M} \boldsymbol{\Sigma}^{-1} e_k. \qquad (9)$$

Yu, Dong, Shao (2016) refer to m_k as the population level marginal SIR utility. To apply Dantzig selector for the estimation of the marginal SIR utility m_k, they defined $p_\ell = E\{\mathbb{1}(Y \in \mathbf{J}_\ell)\}, \ell = 1, \dots, H$. Let $\mu_\ell = E\{\mathbf{X}\mathbb{1}(Y \in \mathbf{J}_\ell)\}$. Then $\mathbf{M}_{\mathrm{SIR}} = \sum_{\ell=1}^H p_\ell E(\mathbf{X} \mid Y \in \mathbf{J}_\ell) E(\mathbf{X}^\top \mid Y \in \mathbf{J}_\ell)$ can be written as $\mathbf{M}_{\mathrm{SIR}} = \sum_{\ell=1}^H \mu_\ell \mu_\ell^\top / p_\ell$. Therefore

$$m_k^{\mathrm{SIR}} = e_k^\top \left(\sum_{\ell=1}^H \nu_\ell \nu_\ell^\top / p_\ell \right) e_k, \quad \nu_\ell = \boldsymbol{\Sigma}^{-1} \mu_\ell.$$

The marginal utility m_k is estimated by

$$\widehat{m}_k^{\mathrm{SIR}} = e_k^\top \left(\sum_{\ell=1}^H \widehat{\nu}_\ell \widehat{\nu}_\ell^\top / \widehat{p}_\ell \right) e_k,$$

where $\widehat{\boldsymbol{\Sigma}} = \sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^\top / n$ and $\widehat{\mu}_\ell = \sum_{i=1}^n \mathbf{X}_i \mathbb{1}(Y_i \in \mathbf{J}_\ell)/n$. For a given threshold b_n, the active set $\mathcal{A}$

is estimated by including the predictors such that $\widehat{m}_k^{\mathrm{SIR}}$ exceeds b_n or $\widehat{\mathcal{A}} = \{k : \widehat{m}_k^{\mathrm{SIR}} \geq b_n\}$. Yu, Dong, Shao (2016) take an example that $\mathbf{X} = (X_1, \dots, X_p)^\top \sim \mathcal{N}(0, \mathbf{\Sigma})$. 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 $|i - j| > 1$, $1 \leq i, j \leq p$. Let $Y = \boldsymbol{\beta}^\top \mathbf{X} + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, 1)$ is independent of $\mathbf{X}$ and $\boldsymbol{\beta} = (1.2, -2, 0. \dots, 0)^\top$. Then the active set for the linear regression models is $\mathcal{A} = \{1, 2\}$. Consider five utilities for X_1: the marginal absolute Pearson correlation from Fan and Lv (2008), the marginal squared distance correlation utility from Li et al. (2012), the marginal fused Kolmogorov filter utility as defined in (5.3) of Yu, Dong, Shao (2016), the marginal independence SIR utility as defined in (5.1) of Yu, Dong, Shao (2016) and the marginal SIR utility as defined in (9). Unfortunately, the first four independence screening methods will fail to recover the active predictor X_1, only marginal SIR achieves desired result.

4.2. Trace pursuit

Yu, Dong, Zhu (2016) proposed trace pursuit as a novel approach for model-free variable selection. They first extended the classical stepwise regression in linear models and proposed an STP algorithm for model-free variable selection. Furthermore, they proposed the FTP algorithm. After finding a solution path by adding one predictor into the model at a time, a modified Bayesian information criterion (BIC, hereafter) provides a chosen model that is guaranteed to include all important predictors. Finally, the two-stage trace pursuit algorithm uses FTP for initial variable screening.

For working index set $\mathcal{F}$ and index $j \in \mathcal{F}^c$, if we want to test

$$H_0 : Y \perp\!\!\!\perp X_j \mid \mathbf{X}_\mathcal{F} \text{ vs. } H_a : \\ Y \text{ is not independent of } X_j \text{ given } \mathbf{X}_\mathcal{F}. \qquad (10)$$

For any index set $\mathcal{F}$, denote $\mathbf{X}_\mathcal{F} = \{X_i : i \in \mathcal{F}\}$, $\mathrm{var}(\mathbf{X}_\mathcal{F}) = \mathbf{\Sigma}_\mathcal{F}$. Taking SIR as an example, denote $\mathbf{M}^{\mathrm{SIR}} = \mathbf{\Sigma}_\mathcal{F}^{-1/2} \mathbf{M}_{\mathrm{SIR}} \mathbf{\Sigma}_\mathcal{F}^{-1/2}$. Recall that $\mathcal{A}$ denotes the active index set satisfying $Y \perp\!\!\!\perp \mathbf{X}_{\mathcal{A}^c} \mid \mathbf{X}_\mathcal{A}$, and $\mathcal{I} = \{1, \dots, p\}$ denotes the full index set. It is worth noting that, if the assumption **(W1)** holds true, then for any index set $\mathcal{F}$ such that $\mathcal{A} \subseteq \mathcal{F} \subseteq \mathcal{I}$, $\mathrm{tr}(\mathbf{M}^{\mathrm{SIR}}_\mathcal{A}) = \mathrm{tr}(\mathbf{M}^{\mathrm{SIR}}_\mathcal{F}) = \mathrm{tr}(\mathbf{M}^{\mathrm{SIR}})$. It suggests that $\mathrm{tr}(\mathbf{M}^{\mathrm{SIR}}_\mathcal{F})$ can be used to capture the strength of relationship between Y and $\mathbf{X}_\mathcal{F}$. Denote $\mathcal{F} \cup j$ as the index set of j together with all the indices in $\mathcal{F}$. Given that $\mathbf{X}_\mathcal{F}$ is already in the model, then trace difference $\mathrm{tr}(\mathbf{M}^{\mathrm{SIR}}_{\mathcal{F}\cup j}) - \mathrm{tr}(\mathbf{M}^{\mathrm{SIR}}_\mathcal{F})$ can be used to test the contribution of the additional variable X_j to Y. The idea of using trace difference is similar to the extra sums of squares test in the classical multiple linear regression setting. The following subset LCM assumption is required in Yu, Dong, Zhu (2016),

$$E(X_j \mid \mathbf{X}_\mathcal{F}) \text{ is a linear function of } \mathbf{X}_\mathcal{F} \text{ for any } \mathcal{F} \\ \subset \mathcal{I} \text{ and } j \in \mathcal{F}^c. \qquad (11)$$

Furthermore, they also provided the STP algorithm, that is

(a) Initialisation. Set the initial working set to be $\mathcal{F} = \emptyset$.

(b) Forward addition. Find index $a_\mathcal{F}$ such that

$$a_\mathcal{F} = \arg\max_{j \in \mathcal{F}^c} \mathrm{tr}(\widehat{\mathbf{M}}^{\mathrm{SIR}}_{\mathcal{F}\cup j}).$$

If $T^{\mathrm{SIR}}_{a_\mathcal{F} \mid \mathcal{F}} = n\{\mathrm{tr}(\widehat{\mathbf{M}}^{\mathrm{SIR}}_{\mathcal{F}\cup a_\mathcal{F}}) - \mathrm{tr}(\widehat{\mathbf{M}}^{\mathrm{SIR}}_\mathcal{F})\} > c^{\mathrm{SIR}}$, update $\mathcal{F}$ to be $\mathcal{F} \cup a_\mathcal{F}$.

(c) Backward deletion. Find index $d_\mathcal{F}$ such that

$$d_\mathcal{F} = \arg\max_{j \in \mathcal{F}} \mathrm{tr}(\widehat{\mathbf{M}}^{\mathrm{SIR}}_{\mathcal{F}\backslash j}).$$

If $T^{\mathrm{SIR}}_{d_\mathcal{F} \mid \mathcal{F}\backslash d_\mathcal{F}} = n\{\mathrm{tr}(\widehat{\mathbf{M}}^{\mathrm{SIR}}_\mathcal{F}) - \mathrm{tr}(\widehat{\mathbf{M}}^{\mathrm{SIR}}_{\mathcal{F}\backslash d_\mathcal{F}})\} < c_{\mathrm{SIR}}$, update $\mathcal{F}$ to be $\mathcal{F} \backslash d_\mathcal{F}$.

(d) Repeat steps (b) and (c) until no predictors can be added or deleted.

The test for SAVE and DR can be defined in a parallel fashion if the following CCV assumption together with the subset LCM (11) assumption holds true

$$\mathrm{var}(X_j \mid \mathbf{X}_\mathcal{F}) \text{ is nonrandom for any } \mathcal{F} \subseteq \mathcal{I} \text{ and } j \in \mathcal{F}^c.$$

Li and Dong (2020) had a recent extension of trace pursuit to matrix-valued predictors. Suppose the response variable $Y \in \mathbb{R}$ and the predictor $\mathbf{X} \in \mathbb{R}^{p \times q}$ have the following general relationship:

$$Y = g(\mathbf{X}) + \varepsilon, \qquad (12)$$

where $g : \mathbb{R}^{p \times q} \to \mathbb{R}$ is an unknown function, ε is independent of $\mathbf{X}$, and $E(\varepsilon) = 0$. Assume that $\mathbf{X}$ follows the matrix normal distribution, which is denoted as $\mathbf{X} \sim \mathcal{N}_{p,q}(\mu, \mathbf{U}, \mathbf{V})$ with $\mu \in \mathbb{R}^{p \times q}$, $\mathbf{U} \in \mathbb{R}^{p \times p}$ and $\mathbf{V} \in \mathbb{R}^{q \times q}$. Then, the row covariance matrix is $\mathbf{U} = E\{(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^\top\}/\mathrm{tr}(\mathbf{V})$, and the column covariance matrix is $\mathbf{V} = E\{(\mathbf{X} - \boldsymbol{\mu})^\top (\mathbf{X} - \boldsymbol{\mu})\}/\mathrm{tr}(\mathbf{U})$.

Let $\mathcal{I}_{row} = \{1, \dots, p\}$ be the full index set of rows and $\mathbf{X}_{j,\cdot}$ be the jth row of $\mathbf{X}$ for $j = 1, \dots, p$. Define the active row set $\mathcal{A}$ as

$$\mathcal{A} = \{j \in \mathcal{I}_{row} : Y \text{ depends on } \mathbf{X}_{j,\cdot} \text{ in model (12)}\}.$$

Similarly, let $\mathcal{I}_{col} = \{1, \dots, q\}$ be the full index set of columns and $\mathbf{X}_{\cdot,k}$ be the kth column of $\mathbf{X}$ for $k = 1, \dots, q$. Define the active column set $\mathcal{B}$ as

$$\mathcal{B} = \{k \in \mathcal{I}_{col} : Y \text{ depends on } \mathbf{X}_{\cdot,k} \text{ in model (12)}\}.$$

Based on the active row and column predictors, model (12) can be expressed as

$$Y = g^*(\mathbf{X}_{\mathcal{A},\mathcal{B}}) + \varepsilon,$$

where $g^* : \mathbb{R}^{|\mathcal{A}| \times |\mathcal{B}|} \to \mathbb{R}$ with $|\cdot|$ denoting the cardinality of a set, and $\mathbf{X}_{\mathcal{A},\mathcal{B}}$ denotes the submatrix of $\mathbf{X}$ that contains the active rows indexed by $\mathcal{A}$ and the active columns indexed by $\mathcal{B}$. Note that Y depends on $\mathbf{X}$ only through $\mathbf{X}_{\mathcal{A},\mathcal{B}}$. Li and Dong (2020) introduced procedures to recover the active row set $\mathcal{A}$ in detail. Let $\mathbf{X}_{j,\cdot}$, $j = 1, \ldots, p$, be the jth row of $\mathbf{X}$ and $\mathbf{X}_{-j,\cdot} \in \mathbb{R}^{(p-1)\times q}$ be the matrix that includes all but the jth row of $\mathbf{X}$. To test the importance of $\mathbf{X}_{j,\cdot}$, they considered the following row hypotheses:

$$H_{0,\{j\}}^{row} : Y \perp\!\!\!\perp \mathbf{X} \mid \mathbf{X}_{-j,\cdot} \text{ vs. } H_{a,\{j\}}^{row} : \\ Y \text{ is not independent of } \mathbf{X} \text{ given } \mathbf{X}_{-j,\cdot}. \qquad (13)$$

Under the null hypothesis, $H_{0,\{j\}}^{row}$, the response Y depends on $\mathbf{X}$ only through $\mathbf{X}_{-j,\cdot}$. In the special case of $q = 1$, $\mathbf{X}$ becomes a p-dimensional vector, and (13) is equivalent to testing the importance of one component of $\mathbf{X}$ given the other $p-1$ predictors. This special case is known as the marginal coordinate test (Cook, 2004). Let $\mathbf{U}_{-j,-j} \in \mathbb{R}^{(p-1)\times(p-1)}$ be the submatrix of $\mathbf{U}$ that excludes the jth row and the jth column of $\mathbf{U}$. Define the following quantity:

$$\delta_j^{row} = \mathrm{tr}(\mathbf{M}) - \mathrm{tr}(\mathbf{M}_{-j,\cdot}),$$

where $\mathbf{M} = \mathbf{U}^{-1}E(\mathbf{X}Y)\mathbf{V}^{-1}E^{\top}(\mathbf{X}Y)$ and $\mathbf{M}_{-j,\cdot} = \mathbf{U}_{-j,-j}^{-1} E(\mathbf{X}_{-j,\cdot}Y)\mathbf{V}^{-1}E^{\top}(\mathbf{X}_{-j,\cdot}Y)$. This trace difference δ_j^{row} is the key quantity to test the importance of the jth row of $\mathbf{X}$, which is same as Yu, Dong, Zhu (2016). Note that $\delta_j^{row} = 0$ under $H_{0,\{j\}}^{row}$.

To develop the screening consistency for ultrahigh dimensional setting, Zhu et al. (2011) proposed a novel variable screening procedure under a unified model framework, which covers a wide variety of commonly used parametric and semiparametric models. They assumed that $E(\mathbf{X}_i) = 0$ and $\mathrm{var}(x_i) = 1$ for $i = 1, \ldots, p$ and $\Omega(Y) = E\{\mathbf{X}F(Y \mid \mathbf{X})\}$ for ease of explanation. It then follows by the law of iterated expectations that $\Omega(Y) = \mathrm{cov}\{\mathbf{X}, \mathbb{1}(Y < y)\}$. Let $\Omega_i(Y)$ be the ith element of $\Omega(Y)$, and defined as

$$\omega_i = E\{\Omega_i^2(Y)\}, \ldots, \quad i = 1, \ldots, p.$$

Then ω_i is to serve as the population quantity of our proposed marginal utility measure for predictor ranking. Intuitively, one can see that, if x_i and Y are independent, then x_i and the indicator function $\mathbb{1}(Y \le y)$ change independently. Consequently, $\omega_i = 0$. On the other hand, if x_i and Y are related, then ω_i must be positive. For ease of presentation, they assumed that the sample predictors are all standardised; that is, $n^{-1}\sum_{j=1}^{n}\mathbf{X}_{ji} = 0$ and $n^{-1}\sum_{j=1}^{n}\mathbf{X}_{ji}^2 = 1$ for $i = 1, \ldots, p$. A natural estimator of ω_i is

$$\tilde{\omega}_i = \frac{1}{n}\sum_{j=1}^{n}\left\{\frac{1}{n}\sum_{k=1}^{n}\mathbf{X}_{ki}\mathbb{1}(Y_k < Y_j)\right\}^2, \quad i = 1, \ldots, p,$$

where $\mathbf{X}_{ki}$ denotes the kth element of x_i. The new method does not require imposing a specific model structure on regression functions, and thus is particularly appealing to ultrahigh-dimensional regressions. They showed that, with the number of predictors growing at an exponential rate of the sample size, the proposed procedure possesses consistency in ranking, which is both useful in its own right and can lead to consistency in selection. Lin et al. (2017) first introduced a large class of models depending on the smallest non-zero eigenvalue λ of the kernel matrix of SIR, then the determination of the minimax rate for estimating the central space over two classes is derived, which is the first paper that studied the minimax estimation of sparse SIR. Furthermore, they showed that the estimator based on the SIR procedure converges at rate $d \wedge ((sd + s\log(ep/s))/(n\lambda))$, which is the optimal rate for the single index models and multiple index models with fixed structural dimension d, fixed $s = |\mathcal{A}|$ and λ. However, Lin et al. (2017) only considered the projection loss (Li & Wang, 2007). More importantly, their theoretical study is actually based on the assumption that covariance matrix is diagonal.

Before discussing the consistency property, we need some conditions. Taking Yu, Dong, Shao (2016) as an example,

(C1) The coverage condition: $\mathbf{Span}\{\boldsymbol{\Sigma}E(\mathbf{X} \mid Y \in \mathbf{J}_\ell)\ell = 1, \ldots, H\} = \mathcal{S}_{Y\mid\mathbf{X}}$.

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

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

(C4) There exists $0 < g < 1 - 2 \times r$ such that $f^2 s^2 \log p = O_p(n^g)$, where s is the cardinality of $\mathcal{A}$ and r is specified in condition (C5).

(C5) There exists $0 < a_2 < \infty$ and $r \le 1/2$ such that $\min_{k\in\mathcal{A}} m_k > 2a_2 n^{-r}$.

More details please refer to Yu, Dong, Shao (2016).

Theorem 4.1: *Assume above conditions hold, then the shrinkage estimator* $\widehat{\boldsymbol{\beta}}$ *satisfies consistency in variable selection,*

$$\Pr(\widehat{\mathcal{A}} \supseteq \mathcal{A}) \to 1.$$

Theorem 4.1 is given in Yu, Dong, Shao (2016), Yu, Dong, Zhu (2016), Lin et al. (2017), and Zhu et al. (2011).

5. Minimax rate

Recently, an impressive range of penalised SIR methods has been proposed to estimate the central subspace in a sparse fashion. However, few of them considered the sparse sufficient dimension reduction from a decision-theoretical point of view. To address this issue, Tan et al. (2020) established the minimax rates of convergence for estimating the sparse SIR directions under various commonly used loss functions in the literature of sufficient dimension reduction. Lin et al. (2019) introduced a simple Lasso regression method to obtain an estimator of the sufficient dimension reduction space, which is only the first step of Tan et al. (2020). Moreover, Tan et al. (2020) discovered the possible trade-off between statistical guarantee and computational performance for sparse SIR and proposed an adaptive estimation scheme for sparse SIR which is computationally tractable and rate optimal under the condition that $\log p = o(n)$, which is weaker than Lin et al. (2019).

As we can see that, the kernel matrix $\mathbf{M}_{\text{SIR}}$ can be estimated as

$$\widehat{\mathbf{M}}_{\text{SIR}} = \sum_{\ell=1}^{H} \widehat{\mu}_\ell \widehat{\mu}_\ell^\top / \widehat{p}_\ell.$$

Then it is natural to estimate $\boldsymbol{\beta}$ via replacing $\mathbf{M}$ and $\boldsymbol{\Sigma}$ in (8) by their sample estimators, which yields

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{\text{SIR}} &= \operatorname*{arg\,max}_{\mathbf{B}\in\mathbb{R}^{p\times d}} \operatorname{Tr}(\mathbf{B}^\top \widehat{\mathbf{M}}_{\text{SIR}} \mathbf{B}) \text{ s.t. } \mathbf{B}^\top \widehat{\boldsymbol{\Sigma}} \mathbf{B} \\ &= \mathbf{I}_d \text{ and } |\operatorname{Supp}(\mathbf{B})| \le s, \qquad (14)\end{aligned}$$

The solution $\widehat{\boldsymbol{\beta}}_{\text{SIR}}$ in (14) is called the natural sparse SIR estimator. The following theorem establishes the lower bound and upper bound of the four loss functions for the natural sparse SIR estimator.

Theorem 5.1: *Assume $n\lambda^2 \ge C_0 \log \frac{ep}{s}$ for some sufficiently large constant C_0. Then there exist positive constants C and c_0 such that*

$$\inf_{\widehat{\boldsymbol{\beta}}} \sup_{\mathbb{P}\in\mathcal{P}} E_{\mathbb{P}} L_G(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \ge C \frac{s\log(ep/s)}{n\lambda^2} \wedge c_0,$$

$$\inf_{\widehat{\boldsymbol{\beta}}} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}\left\{ L_P(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \ge C \frac{s\log(ep/s)}{n\lambda^2} \wedge c_0 \right\} \ge 0.8,$$

$$\inf_{\widehat{\boldsymbol{\beta}}} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}\left\{ L_X(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \ge C \frac{s\log(ep/s)}{n\lambda^2} \wedge c_0 \right\} \ge 0.8,$$

where $\mathcal{P} = \mathcal{P}(n, H, s, p, d, \lambda; K, m)$.

Theorem 5.2: *Assume that $\frac{s\log(ep/s)}{n\lambda^2} \le c$ for some small constant $c \in (0,1)$. Then for any $C' > 0$, there exists a positive constant C such that*

$$\begin{aligned}&L_G(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \vee L_P(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \vee L_X(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \vee L_C(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \\ &\quad \le C \frac{s\log(ep/s)}{n\lambda^2}\end{aligned}$$

with probability greater than $1 - 2\exp(-C'(s + \log (ep/s)))$ uniformly over $\mathcal{P} \in P(n, H, s, p, d, \lambda; K, m)$.

Since SIR can be rewritten as a least-square formulation, they finally proposed an adaptive estimation scheme for sparse SIR which is computationally tractable and rate optimal. More details about the adaptive sparse SIR estimator can refer to Tan et al. (2020).

6. Further investigation

6.1. Marginal utility

Motivated by Yu, Dong, Shao (2016), we can extend their method to SAVE and DR. Let $\varphi_\ell = E\{\mathbf{X}\mathbf{X}^\top \mathbb{1}(Y \in \mathbf{J}_\ell)\}$. Then $\mathbf{M}_{\text{SAVE}} = \sum_{\ell=1}^{H} p_\ell \{\boldsymbol{\Sigma} - \operatorname{var}(\mathbf{X} \mid Y \in \mathbf{J}_\ell)\}^2$ can be written as $\mathbf{M}_{\text{SAVE}} = \sum_{\ell=1}^{H} p_\ell \{\boldsymbol{\Sigma} - \varphi_\ell/p_\ell + \mu_\ell\mu_\ell^\top/p_\ell^2\}^2$. Therefore

$$\begin{aligned}m_k^{\text{SAVE}} &= e_k^\top \boldsymbol{\Sigma}^{-1} \left(\sum_{\ell=1}^{H} p_\ell \{\boldsymbol{\Sigma} - \varphi_\ell/p_\ell + \mu_\ell\mu_\ell^\top/p_\ell^2\}^2 \right) \\ &\quad \times \boldsymbol{\Sigma}^{-1} e_k.\end{aligned}$$

The marginal utility m_k is estimated by

$$\begin{aligned}\widehat{m}_k^{\text{SAVE}} &= e_k^\top \widehat{\boldsymbol{\Sigma}}^{-1} \left(\sum_{\ell=1}^{H} \widehat{p}_\ell \{\widehat{\boldsymbol{\Sigma}} - \widehat{\varphi}_\ell/\widehat{p}_\ell + \widehat{\mu}_\ell\widehat{\mu}_\ell^\top/\widehat{p}_\ell^2\}^2 \right) \\ &\quad \times \widehat{\boldsymbol{\Sigma}}^{-1} e_k,\end{aligned}$$

where $\widehat{\varphi}_\ell = \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^\top \mathbb{1}(Y_i \in \mathbf{J}_\ell)/n$. For a given threshold b_n, the active set $\mathcal{A}$ is estimated by including the predictors such that $\widehat{m}_k^{\text{SAVE}}$ exceeds b_n or $\widehat{\mathcal{A}} = \{k : \widehat{m}_k^{\text{SAVE}} \ge b_n\}$.

Next, we consider marginal DR with the Dantzig selector. Then

$$\begin{aligned}\mathbf{M}_{\text{DR}} = \sum_{\ell=1}^{H} 2p_\ell [&E\{E^2(\mathbf{X}\mathbf{X}^\top)\} \\ &+ E^2\{E(\mathbf{X} \mid Y \in \mathbf{J}_\ell) E(\mathbf{X}^\top \mid Y \in \mathbf{J}_\ell)\} \\ &+ E\{E(\mathbf{X}^\top \mid Y \in \mathbf{J}_\ell) E(\mathbf{X} \mid Y \in \mathbf{J}_\ell)\} \\ &\times E\{E(\mathbf{X} \mid Y \in \mathbf{J}_\ell) E(\mathbf{X}^\top \mid Y \in \mathbf{J}_\ell)\}] - 2\boldsymbol{\Sigma}\end{aligned}$$

can be written as

$$\begin{aligned}\mathbf{M}_{\text{DR}} &= 2\sum_{\ell=1}^{H} p_\ell \left(\varphi_\ell - \boldsymbol{\Sigma}\right)^2 + 2\left(\sum_{\ell=1}^{H} \mu_\ell\mu_\ell^\top/p_\ell^2 \right)^2 \\ &\quad + 2\left(\sum_{\ell=1}^{H} \mu_\ell^\top\mu_\ell/p_\ell^2 \right)\left(\sum_{\ell=1}^{H} \mu_\ell\mu_\ell^\top/p_\ell^2 \right).\end{aligned}$$

Therefore

$$\begin{aligned}m_k^{\text{DR}} = 2e_k^\top \boldsymbol{\Sigma}^{-1} \Bigg\{ &\sum_{\ell=1}^{H} p_\ell \left(\varphi_\ell - \boldsymbol{\Sigma}\right)^2 \\ &+ \left(\sum_{\ell=1}^{H} \mu_\ell\mu_\ell^\top/p_\ell^2 \right)^2 \\ &+ \left(\sum_{\ell=1}^{H} \mu_\ell^\top\mu_\ell/p_\ell^2 \right)\left(\sum_{\ell=1}^{H} \mu_\ell\mu_\ell^\top/p_\ell^2 \right) \Bigg\} \boldsymbol{\Sigma}^{-1} e_k.\end{aligned}$$

The marginal utility m_k is estimated by

$$\begin{aligned}\widehat{m}_k^{\mathrm{DR}} = 2e_k^\top \widehat{\boldsymbol{\Sigma}}^{-1}\Bigg\{&\sum_{\ell=1}^{H} \widehat{p}_\ell\left(\widehat{\varphi}_\ell - \widehat{\boldsymbol{\Sigma}}\right)^2 \\ &+ \left(\sum_{\ell=1}^{H} \widehat{\mu}_\ell\widehat{\mu}_\ell^\top/\widehat{p}_\ell^2\right)^2 \\ &+ \left(\sum_{\ell=1}^{H} \widehat{\mu}_\ell^\top\widehat{\mu}_\ell/\widehat{p}_\ell^2\right)\left(\sum_{\ell=1}^{H} \widehat{\mu}_\ell\widehat{\mu}_\ell^\top/\widehat{p}_\ell^2\right)\Bigg\}\widehat{\boldsymbol{\Sigma}}^{-1}e_k.\end{aligned}$$

For a given threshold b_n, the active set $\mathcal{A}$ is estimated by including the predictors such that $\widehat{m}_k^{\mathrm{DR}}$ exceeds b_n, or $\widehat{\mathcal{A}} = \{k : \widehat{m}_k^{\mathrm{DR}} \geq b_n\}$. Following the proof of Yu, Dong, Shao (2016), we can expect the marginal SAVE and DR with the Dantzig selector to achieve selection consistency.

6.2. Minimax rate

Motivated by Tan et al. (2020), we can further investigate the natural sparse SAVE estimator and upper error bound. Let $E_n\mathbf{X} = \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i$ and $\widehat{\boldsymbol{\Sigma}} = \frac{1}{n-1}\sum_{i=1}^n(\mathbf{X}_i - E_n\mathbf{X})(\mathbf{X}_i - E_n\mathbf{X})^\top$ be the sample mean and sample covariance of $\mathbf{X}$, then the SAVE kernel matrix $\mathbf{M}$ is estimated as

$$\widehat{\mathbf{M}}_{\mathrm{SAVE}} = \sum_{\ell=1}^{H}\widehat{p}_\ell\{\widehat{\boldsymbol{\Sigma}} - \widehat{\varphi}_\ell/\widehat{p}_\ell + \widehat{\mu}_\ell\widehat{\mu}_\ell^\top/\widehat{p}_\ell^2\}^2.$$

Similarly, the DR kernel matrix is estimated as

$$\begin{aligned}\widehat{\mathbf{M}}_{\mathrm{DR}} = {}& 2\sum_{\ell=1}^{H}\widehat{p}_\ell\left(\widehat{\varphi}_\ell - \widehat{\boldsymbol{\Sigma}}\right)^2 + 2\left(\sum_{\ell=1}^{H}\widehat{\mu}_\ell\widehat{\mu}_\ell^\top/\widehat{p}_\ell\right)^2 \\ &+ 2\left(\sum_{\ell=1}^{H}\widehat{\mu}_\ell^\top\widehat{\mu}_\ell/\widehat{p}_\ell^2\right)\left(\sum_{\ell=1}^{H}\widehat{\mu}_\ell\widehat{\mu}_\ell^\top/\widehat{p}_\ell^2\right).\end{aligned}$$

Then it is natural to estimate $\boldsymbol{\beta}$ via replacing $\mathbf{M}$ and $\boldsymbol{\Sigma}$ in (8) by their sample estimators, which yields

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{\mathrm{SAVE}} &= \mathop{\arg\max}_{\mathbf{B}\in\mathbb{R}^{p\times d}} \operatorname{Tr}(\mathbf{B}^\top\widehat{\mathbf{M}}_{\mathrm{SAVE}}\mathbf{B}) \text{ s.t. } \mathbf{B}^\top\widehat{\boldsymbol{\Sigma}}\mathbf{B} \\ &= \mathbf{I}_d \text{ and } |\mathrm{Supp}(\mathbf{B})| \leq s, \\ \widehat{\boldsymbol{\beta}}_{\mathrm{DR}} &= \mathop{\arg\max}_{\mathbf{B}\in\mathbb{R}^{p\times d}} \operatorname{Tr}(\mathbf{B}^\top\widehat{\mathbf{M}}_{\mathrm{DR}}\mathbf{B}) \text{ s.t. } \mathbf{B}^\top\widehat{\boldsymbol{\Sigma}}\mathbf{B} \\ &= \mathbf{I}_d \text{ and } |\mathrm{Supp}(\mathbf{B})| \leq s\end{aligned} \tag{15}$$

The solution $\widehat{\boldsymbol{\beta}}_{\mathrm{SAVE}}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{DR}}$ in (15) are called the natural sparse SAVE and DR estimator. The following theorem establishes the lower bound and upper bound of the four loss functions for the natural sparse SAVE and DR estimator.

Theorem 6.1: *Assume $n\lambda^2 \geq C_0 \log\frac{ep}{s}$ for some sufficiently large constant C_0. Then there exist positive constants C and c_0 such that*

$$\begin{aligned}&\inf_{\widehat{\boldsymbol{\beta}}}\sup_{\mathbb{P}\in\mathcal{P}} E_{\mathbb{P}}L_G(\widehat{\boldsymbol{\beta}},\boldsymbol{\beta}) \geq C\frac{s\log(ep/s)}{n\lambda^2}\wedge c_0, \\ &\inf_{\widehat{\boldsymbol{\beta}}}\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}\left\{L_P(\widehat{\boldsymbol{\beta}},\boldsymbol{\beta}) \geq C\frac{s\log(ep/s)}{n\lambda^2}\wedge c_0\right\} \geq 0.8, \\ &\inf_{\widehat{\boldsymbol{\beta}}}\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}\left\{L_X(\widehat{\boldsymbol{\beta}},\boldsymbol{\beta}) \geq C\frac{s\log(ep/s)}{n\lambda^2}\wedge c_0\right\} \geq 0.8,\end{aligned}$$

where $\mathcal{P} = \mathcal{P}(n, H, s, p, d, \lambda; K, m)$.

Theorem 6.2: *Assume that $\frac{s\log(ep/s)}{n\lambda^2} \leq c$ for some small constant $c \in (0,1)$. Then for any $C' > 0$, there exists a positive constant C such that*

$$\begin{aligned}&L_G(\widehat{\boldsymbol{\beta}},\boldsymbol{\beta}) \vee L_P(\widehat{\boldsymbol{\beta}},\boldsymbol{\beta}) \vee L_X(\widehat{\boldsymbol{\beta}},\boldsymbol{\beta}) \vee L_C(\widehat{\boldsymbol{\beta}},\boldsymbol{\beta}) \\ &\quad \leq C\frac{s\log(ep/s)}{n\lambda^2}\end{aligned}$$

with probability greater than $1 - 2\exp(-C'(s + \log(ep/s)))$ uniformly over $\mathcal{P} \in P(n, H, s, p, d, \lambda; K, m)$. In which $\widehat{\boldsymbol{\beta}}$ is constructed in (15).

Following by Tan et al. (2020), $\widehat{\boldsymbol{\beta}}$ in (15) is rate optimal under general loss, projection loss and prediction loss. Moreover, the natural sparse SAVE estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{SAVE}}$ and DR estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{DR}}$ can also be regarded as one optimal estimator for the SAVE directions and DR directions. However, the estimation procedure (15) depends on the unknown sparsity parameter s and is computationally infeasible as it involves exhaustive search over all $\mathbf{B} \in \mathbb{R}^{p\times d}$ subject to the sparsity constraint. Tan et al. (2020) defined a refined sparse SIR estimator based on that SIR can be viewed as transformation-based projection pursuit. Since SAVE and DR cannot be rewritten as a least-square formulation, we do not define refined sparse SAVE and DR estimator.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research is supported by the National Natural Science Foundation of China Grant 11971170, the 111 project B14019 and the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning.

Notes on contributors

Lu Li is currently a Ph.D student at School of Statistics, East China Normal University.

Dr Xuerong Meggie Wen is currently an associate professor of Statistics at Dept. of Mathematics and Statistics, Missouri University of Science and Technology.

Dr Zhou Yu is a Professor of Statistics at School of Statistics, East China Normal Univercity.

References

Bondell, H. D., & Li, L. (2009). Shrinkage inverse regression estimation for model–free variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *71*(1), 287–299. https://doi.org/10.1111/j.1467-9868.2008.00686.x.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics*, *37*(4), 373–384. https://doi.org/10.1080/00401706.1995.10484371

Candes, E., & Tao, T. (2007). The Dantzig selector: Statistical estimation when *p* is much larger than *n*. *The Annals of Statistics*, *35*(6), 2313–2351. https://doi.org/10.1214/009053606000001523

Chen, J., Stern, M., Wainwright, M. J., & Jordan, M. I. (2017). Kernel feature selection via conditional covariance minimization. In *Advances in Neural Information Processing Systems* (pp. 6946–6955).

Chen, X., Zou, C., & Cook, R. D. (2010). Coordinate-independent sparse sufficient dimension reduction and variable selection. *The Annals of Statistics*, *38*(6), 3696–3723. https://doi.org/10.1214/10-AOS826

Cook, R. D. (1996). Graphics for regressions with a binary response. *Journal of the American Statistical Association*, *91*(435), 983–992. https://doi.org/10.1080/01621459.1996.10476968

Cook, R. D. (1998). *Regression graphics*. Wiley.

Cook, R. D. (2004). Testing predictor contributions in sufficient dimension reduction. *The Annals of Statistics*, *32*(3), 1062–1092. https://doi.org/10.1214/009053604000000292

Cook, R. D., & Forzani, L. (2008). Principal fitted components for dimension reduction in regression. *Statistical Science*, *23*(4), 485–501. https://doi.org/10.1214/08-STS275

Cook, R. D., & Forzani, L. (2009). Likelihood-based sufficient dimension reduction. *Journal of the American Statistical Association*, *104*(485), 197–208. https://doi.org/10.1198/jasa.2009.0106

Cook, R. D., & Ni, L. (2005). Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. *Journal of the American Statistical Association*, *100*(470), 410–428. https://doi.org/10.1198/016214504000001501

Cook, R. D., & Weisberg, S. (1991). Sliced inverse regression for dimension reduction: Comment. *Journal of the American Statistical Association*, *86*(414), 328–332. https://doi.org/10.2307/2290564.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*(456), 1348–1360. https://doi.org/10.1198/016214501753382273

Fan, J., & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *70*(5), 849–911. https://doi.org/10.1111/rssb.2008.70.issue-5

Fukumizu, K., Bach, F. R., & Jordan, M. I. (2009). Kernel dimension reduction in regression. *The Annals of Statistics*, *37*(4), 1871–1905. https://doi.org/10.1214/08-AOS637

Fukumizu, K., & Leng, C. (2014). Gradient-based kernel dimension reduction for regression. *Journal of the American Statistical Association*, *109*(505), 359–370. https://doi.org/10.1080/01621459.2013.838167

Fung, W. K., He, X., Liu, L., & Shi, P. (2002). Dimension reduction based on canonical correlation. *Statistica Sinica*, *12*(2002), 1093–1113. https://www.jstor.org/stable/24307017.

Gao, C., Ma, Z., Ren, Z., & Zhou, H. H. (2015). Minimax estimation in sparse canonical correlation analysis. *The Annals of Statistics*, *43*(5), 2168–2197. https://doi.org/10.1214/15-AOS1332

Gretton, A., Bousquet, O., Smola, A., & Scholkopf, B. (2005). Measuring statistical dependence with Hilbert-Schmidt norms. In *International Conference on Algorithmic Learning Theory* (pp. 63–77). Springer.

Li, K. C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, *86*(414), 316–327. https://doi.org/10.1080/01621459.1991.10475035

Li, K. C. (1992). On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. *Journal of the American Statistical Association*, *87*(420), 1025–1039. https://doi.org/10.1080/01621459.1992.10476258

Li, K. C. (2000). High dimensional data analysis via the SIR/PHD approach. *Lecture Note in Progress*.

Li, L. (2007). Sparse sufficient dimension reduction. *Biometrika*, *94*(3), 603–613. https://doi.org/10.1093/biomet/asm044

Li, Z., & Dong, Y. (2020). Model free variable selection with matrix-valued predictors. *Journal of Computational and Graphical Statistics*, *27*, 1–11. https://doi.org/10.1080/10618600.2020.1806854

Li, L., & Nachtsheim, C. J. (2006). Sparse sliced inverse regression. *Technometrics*, *48*(4), 503–510. https://doi.org/10.1198/004017006000000129

Li, B., & Wang, S. (2007). On directional regression for dimension reduction. *Journal of the American Statistical Association*, *102*(479), 997–1008. https://doi.org/10.1198/016214507000000536

Li, L., & Yin, X. (2008). Sliced inverse regression with regularizations. *Biometrics*, *64*(1), 124–131. https://doi.org/10.1111/j.1541-0420.2007.00836.x

Li, R., Zhong, W., & Zhu, L. (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association*, *107*(499), 1129–1139. https://doi.org/10.1080/01621459.2012.695654

Lin, Q., Li, X., Huang, D., & Liu, J. S. (2017). On the optimality of sliced inverse regression in high dimensions. arXiv preprint arXiv:1701.06009.

Lin, Q., Zhao, Z., & Liu, J. (2019). Sparse sliced inverse regression via Lasso. *Journal of the American Statistical Association*, *114*(528), 1726–1739. https://doi.org/10.1080/01621459.2018.1520115

Lin, Q., Zhao, Z., & Liu, J. S. (2018). On consistency and sparsity for sliced inverse regression in high dimensions. *The Annals of Statistics*, *46*(2), 580–610. https://doi.org/10.1214/17-AOS1561

Ma, Y., & Zhu, L. (2012). A semiparametric approach to dimension reduction. *Journal of the American Statistical Association*, *107*(497), 168–179. https://doi.org/10.1080/01621459.2011.646925

Ma, Y., & Zhu, L. (2013a). A review on dimension reduction. *International Statistical Review*, *81*(1), 134–150. https://doi.org/10.1111/insr.2013.81.issue-1

Ma, Y., & Zhu, L. (2013b). Efficient estimation in sufficient dimension reduction. *The Annals of Statistics*, *41*(1), 250–268. https://doi.org/10.1214/12-AOS1072

Ma, Y., & Zhu, L. (2014). On estimation efficiency of the central mean subspace. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*(5), 885–901. https://doi.org/10.1111/rssb.2014.76.issue-5

Ni, L., Cook, R. D., & Tsai, C. L. (2005). A note on shrinkage sliced inverse regression. *Biometrika*, *92*(1), 242–247. https://doi.org/10.1093/biomet/92.1.242

Qian, W., Ding, S., & Cook, R. D. (2019). Sparse minimum discrepancy approach to sufficient dimension reduction with simultaneous variable selection in ultrahigh dimension. *Journal of the American Statistical Association*, *114*(527), 1277–1290. https://doi.org/10.1080/01621459.2018.1497498

Tan, K., Shi, L., & Yu, Z. (2020). Sparse SIR: Optimal rates and adaptive estimation. *The Annals of Statistics*, *48*(1), 64–85. https://doi.org/10.1214/18-AOS1791

Tan, K. M., Wang, Z., Zhang, T., Liu, H., & Cook, R. D. (2018). A convex formulation for high-dimensional sparse sliced inverse regression. *Biometrika*, *105*(4), 769–782. https://doi-org.ezproxy.uky.edu/10.1093/biomet/asy049.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

Wang, H., & Xia, Y. (2008). Sliced regression for dimension reduction. *Journal of the American Statistical Association*, *103*(482), 811–821. https://doi.org/10.1198/016214508000000418

Wu, Y., & Li, L. (2011). Asymptotic properties of sufficient dimension reduction with a diverging number of predictors. *Statistica Sinica*, *21*(2), 707. https://doi.org/10.5705/ss.2011.v21n2a

Xia, Y., Tong, H., Li, W. K., & Zhu, L. X. (2002). An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(3), 363–410. https://doi.org/10.1111/rssb.2002.64.issue-3

Yin, X., & Cook, R. D. (2002). Dimension reduction for the conditional kth moment in regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(2), 159–175. https://doi.org/10.1111/rssb.2002.64.issue-2

Yin, X., & Cook, R. D. (2003). Estimating central subspaces via inverse third moments. *Biometrika*, *90*(1), 113–125. https://doi.org/10.1093/biomet/90.1.113

Yin, X., & Hilafu, H. (2015). Sequential sufficient dimension reduction for large p, small n problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *77*(4), 879–892. https://doi.org/10.1111/rssb.2015.77.issue-4

Yin, X., Li, B., & Cook, R. D. (2008). Successive direction extraction for estimating the central subspace in a multiple-index regression. *Journal of Multivariate Analysis*, *99*(8), 1733–1757. https://doi.org/10.1016/j.jmva.2008.01.006

Yu, Z., Dong, Y., & Shao, J. (2016). On marginal sliced inverse regression for ultrahigh dimensional model-free feature selection. *The Annals of Statistics*, *44*(6), 2594–2623. https://doi.org/10.1214/15-AOS1424

Yu, Z., Dong, Y., & Zhu, L. X. (2016). Trace pursuit: A general framework for model-free variable selection. *Journal of the American Statistical Association*, *111*(514), 813–821. https://doi.org/10.1080/01621459.2015.1050494

Yu, Z., Zhu, L., Peng, H., & Zhu, L. (2013). Dimension reduction and predictor selection in semiparametric models. *Biometrika*, *100*(3), 641–654. https://doi.org/10.1093/biomet/ast005

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *68*(1), 49–67. https://doi.org/10.1111/rssb.2006.68.issue-1

Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, *38*(2), 894–942. https://doi.org/10.1214/09-AOS729

Zhou, J., & He, X. (2008). Dimension reduction based on constrained canonical correlation and variable filtering. *The Annals of Statistics*, *36*(4), 1649–1668. https://doi.org/10.1214/07-AOS529

Zhu, L. P., Li, L., Li, R., & Zhu, L. X. (2011). Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, *106*(496), 1464–1475. https://doi.org/10.1198/jasa.2011.tm10563

Zhu, L., Miao, B., & Peng, H. (2006). On sliced inverse regression with high-dimensional covariates. *Journal of the American Statistical Association*, *101*(474), 630–643. https://doi.org/10.1198/016214505000001285

Zhu, L. P., & Zhu, L. X. (2009a). On distribution–weighted partial least squares with diverging number of highly correlated predictors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *71*(2), 525–548. https://doi.org/10.1111/rssb.2009.71.issue-2

Zhu, L. P., & Zhu, L. X. (2009b). Nonconcave penalized inverse regression in single-index models with high dimensional predictors. *Journal of Multivariate Analysis*, *100*(5), 862–875. https://doi.org/10.1016/j.jmva.2008.09.003

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418–1429. https://doi.org/10.1198/016214506000000735