Statistica Sinica Preprint No: SS-2022-0112

Title	Slicing-free Inverse Regression in High-dimensional Sufficient Dimension Reduction
Manuscript ID	SS-2022-0112
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202022.0112
Complete List of Authors	Qing Mai, Xiaofeng Shao, Runmin Wang and Xin Zhang
Corresponding Authors	Qing Mai
E-mails	qmai@fsu.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0112

Slicing-free Inverse Regression in High-dimensional Sufficient Dimension Reduction

Qing Mai[1], Xiaofeng Shao[2], Runmin Wang[3] and Xin Zhang[1]

Florida State University[1]

University of Illinois at Urbana-Champaign[2]

Texas A&M University[3]

Abstract: The sliced inverse regression (SIR) is the most recognized method in sufficient dimension reduction. For high-dimensional multivariate applications, there is promising progress related to the theory and methods of a high-dimensional SIR. However, two problems remain in this context. First, choosing the number of slices in an SIR is difficult, and depends on the sample size, distributions of the variables, and other practical considerations. Second, extending the SIR from a univariate response to a multivariate response is not trivial. Targeting the same dimension reduction subspace as that of the SIR, we propose a new slicing-free method that provides a unified solution to sufficient dimension reduction for high-dimensional covariates and univariate or multivariate responses. We achieve this by adopting the martingale difference divergence matrix (MDDM) and penalized eigen-decomposition algorithms. To establish the consistency of our method for a high-dimensional predictor and a multivariate response, we develop a new concentration inequality for the sample MDDM around its population counterpart using U-statistics theory, which may be of independent interest. Simulations and a real-data analysis demonstrate the favorable finite-sample performance of the proposed method.

Key words and phrases: Multivariate response, Sliced inverse regression, Sufficient dimension reduction, U-statistic.

1. Introduction

Sufficient dimension reduction (SDR) is an important statistical analysis tool for data visualization, summary, and inference. SDR extracts low-rank projections of the predictors X that contain all information about the response Y, without prespecifying a parametric model. The semiparametric nature of SDR leads to great flexibility and convenience in practice. After performing SDR, we can model the conditional distributions of the response, given the lower-dimensional projected covariate, using existing parametric or nonparametric methods. A salient feature of SDR is that the low-rank projection space can be estimated accurately at a parametric rate, with the nonparametric part treated as an infinite-dimensional nuisance parameter. For example, in multi-index models, SDR is used to estimate the multiple projection directions, without estimating the unspecified link function.

A cornerstone of SDR is the sliced inverse regression (SIR), pioneered by Li (1991), who first discovered the connection between the low-rank projection space and the eigen-space of $\mathrm{cov}(\mathrm{E}(\mathbf{X} \mid Y))$, under suitable assumptions. An SIR is performed by slicing the response Y, and then aggregating the conditional mean of the predictor $\mathbf{X}$, given the response Y within each slice. For example, consider a univariate response Y. Slicing involves picking $K + 1$ constants $-\infty = a_0 < a_1 <$

$\ldots < a_K = \infty$, and defining a new random variable H, where $H = k$ if and only if $a_{k-1} < Y \leq a_k$. After a centering and standardization of the covariate, that is, $\mathbf{X} \to \widetilde{\mathbf{X}} = \mathbf{\Sigma}_{\mathbf{X}}^{-1/2}(\mathbf{X} - \mathrm{E}(\mathbf{X}))$, a simple eigen-decomposition can be used to find linear projections that explain most of the variability in the conditional expectation of the transformed predictor given the response across slices, that is, $\mathrm{cov}(\mathrm{E}(\widetilde{\mathbf{X}} \mid H))$. An important variation of the SIR is the sliced average variance estimation (Cook and Weisberg; 1991), which uses the conditional variance across slices. A key step in these inverse regression methods is the choice of the slicing scheme. If Y is sliced too coarsely, we may not be able to capture the full dependence of Y on the predictors, leading to significant bias in the estimation of $\mathrm{cov}(\mathrm{E}(\widetilde{\mathbf{X}} \mid Y))$. In contrast, if Y is sliced too finely, then the within-slice sample size becomes too small, leading to large variability in the estimation. Although Li (1991) and Hsing and Carroll (1992) show that SDR can still be consistent in a large sample even when the slicing scheme is chosen poorly, Zhu and Ng (1995) argue that the choice of slicing scheme is critical to achieve high estimation efficiency. However, to the best of our knowledge, there is little generally applicable guidance in the literature on how to choose a good slicing scheme.

Zhu et al. (2010) and Cook and Zhang (2014) show that it is beneficial to aggregate multiple slicing schemes, rather than relying on one, although their methods focus only on a univariate response, and in many real-life problems, multi-response data are common. Here, a component-wise analysis may not be sufficient, because it does not make full use of the component-wise dependence in the response. However, slicing a multivariate response is notoriously difficult, owing to the curse of dimensionality, a common problem in multivariate nonparametric smoothing. As the dimension for the response becomes moderately large, it becomes increasingly difficult to ensure that each slice contains a reasonable number of samples, and the estimation can be unstable in practice. Hence, it is highly desirable to develop new SDR methods that do not involve slicing.

An important line of research in the recent SDR literature is to develop SDR methods for data sets with high-dimensional covariates, as motivated by many contemporary applications. The idea of SDR is naturally attractive for high-dimensional data sets, because an effective reduction of the dimension in $\mathbf{X}$ allows us to use existing modeling and inference methods for low-dimensional covariates. However, most classical SDR methods are not directly applicable to the large p small n setting, where p is the dimension of $\mathbf{X}$ and n is the sample size. To overcome the challenges associated with high-dimensional covariates, several methods have been proposed. Lin et al. (2018) show that the SIR estimator is consistent if and only if $\lim p/n = 0$. When the dimension p is larger than n, they propose a diagonal thresholding screening SIR (DT-SIR) algorithm, and show that it is consistent in terms of recovering the dimension reduction space, under certain sparsity assumptions on both the covariance matrix of the predictors and the loadings of the directions. Lin et al. (2019) introduce a simple Lasso regression method that estimates the SDR space by constructing artificial response variables from the top eigenvectors of the estimated conditional covariance matrix. Tan et al. (2018a) propose a two-stage computational framework to solve the sparse generalized eigenvalue problem, which includes the high-dimensional SDR as a special case, and propose a truncated Rayleigh flow method (RIFLE) to estimate the leading generalized eigenvector; see also Lin et al. (2020) and Tan et al. (2018b). Although these methods provide valuable tools to tackle the high-dimensional SDR problem, they still rely on the SIR in their methodology and involve choosing a single slicing scheme, with little guidance on how to choose such a scheme. Consequently, these methods cannot be applied easily to data with a multivariate response, and the effect of the choice of slicing scheme is unclear.

In this article, we propose a novel slicing-free SDR method in the high-dimensional setting. Our proposal is inspired by a recent nonlinear dependence metric, called the martingale difference divergence matrix (MDDM, Lee and Shao; 2018). Lee and Shao (2018) developed the MDDM as a matrix-valued extension of the martingale difference divergence (MDD) of Shao and Zhang (2014), which measures the (conditional) mean dependence of a response variable given a covariate, and used it to reduce the dimension of a multivariate time series. As recently revealed by Zhang et al. (2020), at the population level, the eigenvectors (or generalized eigenvectors) of the MDDM are always contained in the central subspace. Building on these prior works, we propose using a penalized eigen-decomposition on the MDDM to perform SDR in high dimensions. When the covariance matrix of the predictor is the identity matrix, we use the truncated power method with hard thresholding to estimate the top-K eigenvectors of the MDDM. For a more general covariance structure, we apply the RIFLE algorithm (Tan et al.; 2018a) to the sample MDDM instead of to the sample SIR estimator of $\text{cov}(\text{E}(\mathbf{X} \mid Y))$. By using the sample MDDM, this approach is free of slicing, enabling us to treat univariate and multivariate responses in a unified way, and thus circumvent the practical difficulty of selecting the number of slices (especially for a multivariate response). From a theoretical perspective, we derive a concentration inequality for the sample MDDM around its population counterpart by using U-statistics theory, and obtain a rigorous nonasymptotic theoretical justification for the estimated central subspaces for both settings. The results of simulations and a real-data analysis confirm that the proposed penalized MDDM outperforms slicing-based methods in terms of estimation accuracy.

The rest of this paper is organized as follows. In Section 2, we give a brief review of the MDDM, and then present a new concentration inequality for the sample MDDM around its population counterpart. In Section 3, we present our general methodology of adopting the MDDM in both model-free and model-based SDR problems, where we establish population-level connections between the central subspace and the eigen-decomposition and the generalized eigen-decomposition of the MDDM. Algorithms for regularized eigen-decomposition and generalized eigen-decomposition problems are proposed in Sections 4.1 and 4.2, respectively. Theoretical properties are established in Section 5. Section 6 contains numerical studies. Finally, Section 7 concludes the paper. The Supplementary Material provides all additional technical details and numerical results.

2. The MDDM and its concentration inequality

Consider a pair of random vectors $\mathbf{V} \in \mathbb{R}^p$ and $\mathbf{U} \in \mathbb{R}^q$, such that $\mathrm{E}(\|\mathbf{U}\|^2 + \|\mathbf{V}\|^2) < \infty$. We use $\|\mathbf{U}\| = |\mathbf{U}|_q$ to denote the Euclidean norm in $\mathbb{R}^q$. Define

$$\mathrm{MDDM}(\mathbf{V} \mid \mathbf{U}) = -\mathrm{E}\left[\{\mathbf{V} - \mathrm{E}(\mathbf{V})\}\{\mathbf{V}' - \mathrm{E}(\mathbf{V}')\}^{\mathrm{T}}\|\mathbf{U} - \mathbf{U}'\|\right] \in \mathbb{R}^{p \times p},$$

where $(\mathbf{V}', \mathbf{U}')$ is an independent copy of $(\mathbf{V}, \mathbf{U})$. Lee and Shao (2018) established the following key properties of $\mathrm{MDDM}(\mathbf{V} \mid \mathbf{U})$: (i) it is symmetric and positive semi-definite; (ii) $\mathrm{E}(\mathbf{V} \mid \mathbf{U}) = \mathrm{E}(\mathbf{V})$, almost surely, is equivalent to $\mathrm{MDDM}(\mathbf{V} \mid \mathbf{U}) = 0$; (iii) for any $p \times d$ matrix $\mathbf{A}$, $\mathrm{MDDM}(\mathbf{A}^{\mathrm{T}}\mathbf{V} \mid \mathbf{U}) = \mathbf{A}^{\mathrm{T}}\mathrm{MDDM}(\mathbf{V} \mid \mathbf{U})\mathbf{A}$; (iv) there exist $p - d$ linearly independent combinations of $\mathbf{V}$ that are (conditionally) mean independent of $\mathbf{U}$ if and only if $\mathrm{rank}(\mathrm{MDDM}(\mathbf{V}|\mathbf{U})) = d$.

Given a random sample of size n, that is, $(\mathbf{U}_k, \mathbf{V}_k)_{k=1}^n$, the sample estimate of $\mathrm{MDDM}(\mathbf{V} \mid \mathbf{U})$, denoted by $\mathrm{MDDM}_n(\mathbf{V} \mid \mathbf{U})$, is defined as

$$\mathrm{MDDM}_n(\mathbf{V} \mid \mathbf{U}) = -\frac{1}{n^2}\sum_{j,k=1}^{n}(\mathbf{V}_j - \overline{\mathbf{V}}_n)(\mathbf{V}_k - \overline{\mathbf{V}}_n)^{\mathrm{T}}|\mathbf{U}_j - \mathbf{U}_k|_q, \qquad (2.1)$$

where $\overline{\mathbf{V}}_n = n^{-1}\sum_{k=1}^n \mathbf{V}_k$ is the sample mean.

In the following, we present a concentration inequality for the sample MDDM around its population counterpart, which plays an instrumental role in our consistency proof for the proposed penalized MDDM method later. To this end, we let $\mathbf{V} = (V_1, \cdots, V_p)^{\mathrm{T}} \in \mathbb{R}^p$, and assume the following condition.

(C1) There exist two positive constants σ_0 and C_0 such that

$$\begin{aligned} &\sup_p \max_{1\leq j\leq p} \mathrm{E}\{\exp(2\sigma_0 V_j^2)\} \leq C_0, \\ &\mathrm{E}\{\exp(2\sigma_0 \|\mathbf{U}\|_q^2)\} \leq C_0. \end{aligned} \tag{2.2}$$

For a matrix $A = (a_{ij})$, we denote its max norm as $\|A\|_{max} = \max_{ij} |a_{ij}|$.

Theorem 1. *Suppose that Condition (C1) holds. There exists a positive integer $n_0 = n_0(\sigma_0, C_0, q) < \infty$, $\gamma = \gamma(\sigma_0, C_0, q) \in (0, 1/2)$, and a finite positive constant $D_0 = D_0(\sigma_0, C_0, q) < \infty$, such that when $n \geq n_0$ and $16 > \epsilon > D_0 n^{-\gamma}$, we have*

$$P(\|\mathrm{MDDM}_n(\mathbf{V}|\mathbf{U}) - \mathrm{MDDM}(\mathbf{V}|\mathbf{U})\|_{\max} > 12\epsilon) \leq 54p^2 \exp\left\{-\frac{\epsilon^2 n}{36\log^3(n)}\right\}.$$

The above bound is nonasymptotic and holds for all (n, p, ϵ), as long as the condition is satisfied. The exponent $\dfrac{\epsilon^2 n}{\log^3(n)}$ is from the use of a truncation argument, along with Hoeffding's inequality for U-statistics, and seems hard to improve. Nevertheless, we achieve an exponential-type bound under a uniform sub-Gaussian condition on both $\mathbf{V}$ and $\mathbf{U}$. This result may be of independent theoretical interest. For example, in the time series dimension reduction problem studied by Lee and Shao (2018), our Theorem 1 could potentially help extend their theory from low-dimensional multivariate time series to higher dimensions.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0112

3. Slicing-free Inverse Regression using the MDDM

3.1 Inverse regression subspace in SDR

SDR methods aim to identify the central subspace that preserves all information in the predictors. In this study, we consider the SDR problem of a multivariate response $\mathbf{Y} \in \mathbb{R}^q$ on a multivariate predictor $\mathbf{X} \in \mathbb{R}^p$. The central subspace $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ is defined as the intersection of all subspaces $\mathcal{S}$ such that $\mathbf{Y} \perp\!\!\!\perp \mathbf{X} \mid \mathbf{P}_{\mathcal{S}}\mathbf{X}$, where $\mathbf{P}_{\mathcal{S}}$ is the projection matrix onto $\mathcal{S}$. By construction, the central subspace $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ is the smallest dimension reduction subspace that contains all information in the conditional distribution of $\mathbf{Y}$ given $\mathbf{X}$. Many methods have been proposed for recovering the central subspace or a portion of the central subspace (Li; 1991; Cook and Weisberg; 1991; Bura and Cook; 2001; Chiaromonte et al.; 2002; Yin and Cook; 2003; Cook and Ni; 2005; Li and Wang; 2007; Zhou and He; 2008); see Li (2018) for a comprehensive review. Although the central subspace is well defined for both univariate and multivariate responses, most existing SDR methods consider the case with a univariate response, and an extension to a multivariate response is nontrivial.

The definition of a central subspace is not very constructive, because it requires taking the intersection of *all* subspaces $\mathcal{S} \subseteq \mathbb{R}^p$ such that $\mathbf{Y} \perp\!\!\!\perp \mathbf{X} \mid \mathbf{P}_{\mathcal{S}}\mathbf{X}$. It is difficult to estimate the central subspace without specifying a model between $\mathbf{Y}$ and $\mathbf{X}$. To achieve this, we often need additional assumptions, such as the linearity and the coverage conditions. The linearity condition requires that, for any basis of the central subspace $\boldsymbol{\beta}$, we must have that $\mathrm{E}(\mathbf{X} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})$ is linear in $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}$. The linearity condition is guaranteed if $\mathbf{X}$ is elliptically contoured, and allows us to connect the central subspace to the conditional expectation $\mathrm{E}(\mathbf{X} \mid \mathbf{Y})$. Define $\boldsymbol{\Sigma}_{\mathbf{X}}$ as the covariance of $\mathbf{X}$, and the *inverse regression subspace*

$$\mathcal{S}_{\mathrm{E}(\mathbf{X}|\mathbf{Y})} \equiv \mathrm{span}\{\mathrm{E}(\mathbf{X} \mid \mathbf{Y} = \mathbf{y}) - \mathrm{E}(\mathbf{X}) : \mathbf{y} \in \mathbb{R}^q \text{ such that } \mathrm{E}(\mathbf{X} \mid \mathbf{Y} = \mathbf{y}) \text{ exists}\}. \tag{3.1}$$

The following property is well known, and is often used to develop SDR methods.

Proposition 1. *Under the linearity condition, we have* $\mathcal{S}_{\mathrm{E}(\mathbf{X}|\mathbf{Y})} \subseteq \boldsymbol{\Sigma}_{\mathbf{X}}\mathcal{S}_{\mathbf{Y}|\mathbf{X}} \subseteq \mathbb{R}^p$.

The coverage condition further assumes that $\mathcal{S}_{\mathrm{E}(\mathbf{X}|\mathbf{Y})} = \boldsymbol{\Sigma}_{\mathbf{X}}\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$. It follows that we can estimate the central subspace by modeling the conditional expectation of $\mathbf{X}$. Indeed, many SDR methods approximate $\mathrm{E}(\mathbf{X} \mid \mathbf{Y})$. For example, the SIR slices the univariate Y into several categories, and estimates the mean of $\mathbf{X}$ within each slice. Most methods follow this slice-and-estimate procedure. The number of slices is important to the estimation. If there are too few slices, we may not be able to fully capture the dependence of $\mathbf{X}$ on Y; however, if there are too many slices, there are insufficient samples within each slice to allow an accurate estimation.

3.2 The MDDM in SDR

In this section, we lay the foundation for applying the MDDM to SDR. We show that the subspace spanned by the MDDM coincides with the inverse regression subspace in (3.1). In particular, we have Proposition 2, which is also used in Zhang et al. (2020), without a proof, in the context of a multivariate linear regression.

Proposition 2. *For multivariate* $\mathbf{X} \in \mathbb{R}^p$ *and* $\mathbf{Y} \in \mathbb{R}^q$*, assuming the existence of* $\mathrm{E}(\mathbf{X})$, $\mathrm{E}(\mathbf{X} \mid \mathbf{Y})$, *and* $\mathrm{MDDM}(\mathbf{X} \mid \mathbf{Y})$, *we have* $\mathcal{S}_{\mathrm{E}(\mathbf{X}|\mathbf{Y})} = \mathrm{span}\{\mathrm{MDDM}(\mathbf{X} \mid \mathbf{Y})\}$.

Therefore, the rank of $\mathrm{MDDM}(\mathbf{X} \mid \mathbf{Y})$ is the dimensionality of the inverse regression subspace, and the nontrivial eigenvectors of $\mathrm{MDDM}(\mathbf{X} \mid \mathbf{Y})$ contain all the information for $\mathcal{S}_{\mathrm{E}(\mathbf{X}|\mathbf{Y})}$. Combining Propositions 1 & 2, we immediately have that (i) under the linearity condition, $\mathbf{\Sigma}_{\mathbf{X}}^{-1}\mathrm{span}\{\mathrm{MDDM}(\mathbf{X} \mid \mathbf{Y})\} \subseteq \mathcal{S}_{\mathbf{Y}|\mathbf{X}}$, and (ii) under the linearity and coverage conditions, $\mathbf{\Sigma}_{\mathbf{X}}^{-1}\mathrm{span}\{\mathrm{MDDM}(\mathbf{X} \mid \mathbf{Y})\} = \mathcal{S}_{\mathbf{Y}|\mathbf{X}}$.

Henceforth, we assume both the linearity and coverage conditions, which are assumed either explicitly or implicitly in inverse regression-type dimension reduction methods (e.g., Li; 1991; Cook and Ni; 2005; Zhu et al.; 2010; Cook and Zhang; 2014). Then, the central subspace is related to the eigen-decomposition of $\mathrm{MDDM}(\mathbf{X} \mid \mathbf{Y})$. Specifically, we have the following scenarios.

If $\mathrm{cov}(\mathbf{X}) = \sigma^2\mathbf{I}_p$, for some $\sigma^2 > 0$, then obviously $\mathrm{span}\{\mathrm{MDDM}(\mathbf{X} \mid \mathbf{Y})\} = \mathcal{S}_{\mathbf{Y}|\mathbf{X}}$. This includes single-index and multiple-index models with uncorrelated predictors. Let K be the rank of $\mathrm{MDDM}(\mathbf{X} \mid \mathbf{Y})$. Then, the dimension of the central subspace is K, and the first K eigenvectors of $\mathrm{MDDM}(\mathbf{X} \mid \mathbf{Y})$ span the central subspace.

If $\mathrm{cov}(\mathbf{X} \mid \mathbf{Y}) = \sigma^2\mathbf{I}_p$, for some $\sigma^2 > 0$, then we have $\mathbf{\Sigma}_{\mathbf{X}} = \sigma^2\mathbf{I}_p + \mathrm{cov}\{\mathrm{E}(\mathbf{X} \mid \mathbf{Y})\}$. Because $\mathrm{span}[\mathrm{cov}\{\mathrm{E}(\mathbf{X} \mid \mathbf{Y})\}] = \mathcal{S}_{\mathrm{E}(\mathbf{X}|\mathbf{Y})}$, we can show that $\mathcal{S}_{\mathbf{Y}|\mathbf{X}} = \mathbf{\Sigma}_{\mathbf{X}}^{-1}\mathrm{span}\{\mathrm{MDDM}(\mathbf{X} \mid \mathbf{Y})\} = \mathrm{span}\{\mathrm{MDDM}(\mathbf{X} \mid \mathbf{Y})\}$. To see this, let $\mathrm{cov}\{\mathrm{E}(\mathbf{X} \mid \mathbf{Y})\} = \mathbf{U}\mathbf{U}^{\mathrm{T}}$, for some

$\mathbf{U} \in \mathbb{R}^{p\times K}$. Then, $\mathrm{span}(\mathbf{U}) = \mathrm{span}[\mathrm{cov}\{\mathrm{E}(\mathbf{X} \mid \mathbf{Y})\}] = \mathrm{span}\{\mathrm{MDDM}(\mathbf{X} \mid \mathbf{Y})\}$, and we may also write $\mathrm{MDDM}(\mathbf{X} \mid \mathbf{Y}) = \mathbf{U}\boldsymbol{\Psi}\mathbf{U}^{\mathrm{T}}$, for some symmetric positive-definite matrix $\boldsymbol{\Psi} \in \mathbb{R}^{K\times K}$. The result follows by applying the Woodbury matrix identity to $\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} = (\sigma^2\mathbf{I}_p + \mathbf{U}\mathbf{U}^{\mathrm{T}})^{-1} = \sigma^{-2}\mathbf{I}_p - \sigma^{-2}\mathbf{U}(\sigma^2\mathbf{I}_K + \mathbf{U}^{\mathrm{T}}\mathbf{U})^{-1}\mathbf{U}^{\mathrm{T}}$. The nontrivial eigenvectors of $\mathrm{MDDM}(\mathbf{X} \mid \mathbf{Y})$ again span the central subspace.

For a general covariance structure, the d-dimensional central subspace $\mathcal{S}_{\mathbf{Y}|\mathbf{X}} = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\mathrm{span}\{\mathrm{MDDM}(\mathbf{X} \mid \mathbf{Y})\}$ can be obtained by using a generalized eigen-decomposition. Specifically, consider the generalized eigenvalue problem

$$\mathrm{MDDM}(\mathbf{X} \mid \mathbf{Y})\mathbf{v}_i = \varphi_i\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{v}_i,\ \ \varphi_i \geq 0,\ \ \mathbf{v}_i \in \mathbb{R}^p, \tag{3.2}$$

where $\mathbf{v}_i^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{v}_j = 0$, for $i \neq j$. Then, similarly to Li (2007) and Chen et al. (2010), it is straightforward to show that the generalized eigenvector spans the central subspace, $\mathcal{S}_{\mathbf{Y}|\mathbf{X}} = \mathrm{span}(\mathbf{v}_1, \ldots, \mathbf{v}_K)$.

Existing works on SDR often focus on the eigen-decomposition or the generalized eigen-decomposition of $\mathrm{cov}\{\mathrm{E}(\mathbf{X} \mid Y = y)\}$, where nonparametric estimates of $\mathrm{E}(\mathbf{X} \mid Y = y)$ are obtained by slicing the support of the univariate response Y. In contrast, the MDDM approach requires no tuning parameter selection (i.e., specifying a slicing scheme). Moreover, a high-dimensional theoretical study of the MDDM is easier, and does not require additional assumptions on the conditional mean function $\mathrm{E}(\mathbf{X} \mid \mathbf{Y})$, such as smoothness in the empirical mean function of $\mathbf{X}$ given Y (e.g., sliced stable condition in Lin et al. 2018).

3.3 The MDDM for model-based SDR

Thus far, we have discussed model-free SDR. Another important research area in SDR is model-based methods, which provide valuable insights when using an inverse regression estimation under the assumption that the conditional distribution of $\mathbf{X} \mid \mathbf{Y}$ is normal. In this section, we consider the principal fitted component (PFC) model, which is discussed in detail in Cook and Forzani (2009) and Cook (2007), and generalize it from a univariate response to a multivariate response. We argue that the (generalized) eigen-decomposition of the MDDM is potentially advantageous to likelihood-based approaches under the PFC model. This is somewhat surprising, but reasonable, considering that the advantages of the MDDM over least squares and likelihood-based estimations are demonstrated in Zhang et al. (2020) for multivariate linear models.

Let $\mathbf{X}_{\mathbf{y}} \sim \mathbf{X} \mid (\mathbf{Y} = \mathbf{y})$ denote the conditional variable. Then, the PFC model is

$$\mathbf{X}_{\mathbf{y}} = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\nu}_{\mathbf{y}} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(0, \boldsymbol{\Delta}), \tag{3.3}$$

where $\boldsymbol{\Gamma} \in \mathbb{R}^{p\times K}$, for $K < p$, is a nonstochastic orthogonal matrix, and $\boldsymbol{\nu}_{\mathbf{y}} \in \mathbb{R}^{K}$ is the latent variable that depends on $\mathbf{y}$. Then, the latent variable $\boldsymbol{\nu}_{\mathbf{y}}$ is fitted as $\boldsymbol{\nu}_{\mathbf{y}} = \boldsymbol{\alpha}\mathbf{f}_{\mathbf{y}}$, with some user-specified functions $\mathbf{f}_{\mathbf{y}} = (f_1(\mathbf{y}), \ldots, f_m(\mathbf{y}))^{\mathrm{T}} \in \mathbb{R}^{m}$, for $m \geq K$, which maps a q-dimensional response to an m-dimensional response. In the univariate PFC model, $q = 1$, so the m functions can be viewed as an expansion of the response (similar to slicing). For our multivariate extensions of the PFC model,

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0112

there is no requirement of $m \geq q$. The PFC model can be written as

$$\mathbf{X}_y = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\alpha}\mathbf{f}_{\mathbf{y}} + \boldsymbol{\varepsilon}, \tag{3.4}$$

where $\boldsymbol{\Gamma}$ and $\boldsymbol{\alpha}$ are estimated similarly to the multivariate reduced-rank regression, with $\mathbf{X} \in \mathbb{R}^p$ being the response and $\mathbf{f}_{\mathbf{y}} \in \mathbb{R}^m$ being the predictor. Finally, the central subspace under this PFC model is $\boldsymbol{\Delta}^{-1}\mathrm{span}(\boldsymbol{\Gamma})$, which simplifies to $\mathrm{span}(\boldsymbol{\Gamma})$ if we further assume the isotropic error (i.e., isotropic PFC model) $\boldsymbol{\Delta} = \mathrm{cov}(\mathbf{X} \mid \mathbf{Y}) = \sigma^2\mathbf{I}_p$.

For the PFC model, our MDDM approach is the same as the model-free MDDM counterpart, and has two main advantages over the likelihood-based PFC estimation: (i) there is no need to specify the functions $\mathbf{f}_{\mathbf{y}}$, and thus no risk of misspecification, and (ii) extensions to high-dimensional settings are much more straightforward. Moreover, under the isotropic PFC model, the central subspace $\mathcal{S}_{\mathbf{Y}|\mathbf{X}} = \mathrm{span}(\boldsymbol{\Gamma})$ is exactly the first K eigenvectors of $\mathrm{MDDM}(\mathbf{X} \mid \mathbf{Y})$.

4. Estimation

4.1 Penalized decomposition of the MDDM

Based on the results in the last section, we can use the penalized eigen-decomposition of the MDDM to estimate the central subspace in a high dimension when the covariance $\boldsymbol{\Sigma}_{\mathbf{X}}$ or the conditional covariance $\mathrm{cov}(\mathbf{X} \mid \mathbf{Y})$ is proportional to the identity matrix $\mathbf{I}_p$. Here, we construct such an estimate. Note that the penalized decomposition of the MDDM we develop here is immediately applicable to the dimension reduction of multivariate stationary time series in Lee and Shao (2018). However, this is beyond the scope of this study. Moreover, it is well known that $\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}$ is not easy to estimate in high dimensions. Then, even for a general covariance structure, the eigen-decomposition of the MDDM provides an estimate of the inverse regression subspace (though it may differ from the central subspace) that is useful for exploratory data analysis (e.g., detecting and visualizing a nonlinear mean function).

As such, we estimate the eigenvectors of $\mathrm{MDDM}(\mathbf{X} \mid \mathbf{Y})$. We assume that $\mathrm{MDDM}(\mathbf{X} \mid \mathbf{Y})$ has K nontrivial eigenvectors, denoted by $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$, respectively. We use the shorthand notation $\mathbf{M} = \mathrm{MDDM}(\mathbf{X} \mid \mathbf{Y})$. In addition, note that, given the first $k-1$ eigenvectors, $\boldsymbol{\beta}_k$ is the top eigenvector of $\mathbf{M}_k$, where $\mathbf{M}_k = \mathbf{M} - \sum_{l<k}(\boldsymbol{\beta}_l^{\mathrm{T}}\mathbf{M}\boldsymbol{\beta}_l)\boldsymbol{\beta}_l\boldsymbol{\beta}_l^{\mathrm{T}}$.

It is well known that the eigenvectors cannot be estimated accurately in high dimensions without additional assumptions. We adopt the popular sparsity assumption that many entries in $\boldsymbol{\beta}_k$ are zero. To estimate these sparse eigenvectors, denote $\widehat{\mathbf{M}}_1 = \mathrm{MDDM}_n(\mathbf{X} \mid \mathbf{Y})$, where the sample MDDM_n is defined in (2.1). We find $\widehat{\boldsymbol{\beta}}_k$, for $k = 1, \ldots, K$, as follows:

$$\widehat{\boldsymbol{\beta}}_k = \arg\max_{\boldsymbol{\beta}} \boldsymbol{\beta}^{\mathrm{T}}\widehat{\mathbf{M}}_k\boldsymbol{\beta} \text{ s.t. } \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta} = 1, \|\boldsymbol{\beta}\|_0 \leq s, \tag{4.1}$$

where $\widehat{\mathbf{M}}_1 = \mathrm{MDDM}_n(\mathbf{X} \mid \mathbf{Y})$, $\widehat{\mathbf{M}}_k = \widehat{\mathbf{M}}_1 - \sum_{l<k}\delta_l\widehat{\boldsymbol{\beta}}_l\widehat{\boldsymbol{\beta}}_l^{\mathrm{T}}$, for $k > 1$, with $\delta_l = \widehat{\boldsymbol{\beta}}_l^{\mathrm{T}}\widehat{\mathbf{M}}_1\widehat{\boldsymbol{\beta}}_l$, and s is a tuning parameter.

We solve the above problem by combining the truncated power method with hard thresholding. For a vector $\mathbf{v} \in \mathbb{R}^p$ and a positive integer s, denote v_s^* as the sth largest value of $|v_j|$, for $j = 1, \ldots, p$. The hard-thresholding operator is $\mathrm{HT}(\mathbf{v}, s) = (v_1 I(|v_1| \geq v_s^*), \ldots, v_p I(|v_p| \geq v_s^*))^{\mathrm{T}}$, which sets the $p - s$ elements in $\mathbf{v}$ to zero. We solve (4.1) using Algorithm 1, where the initialization $\widehat{\boldsymbol{\beta}}_1^{(0)}$ may be randomly generated. Note that Yuan and Zhang (2013) proposed Algorithm 1 to perform a principal component analysis using the penalized eigen-decomposition on the sample covariance.

In our algorithm, we require a prespecified sparsity level s and subspace dimension K. In terms of theory, we show that our estimators for $\boldsymbol{\beta}_k$, for $k = 1, \ldots, K$, are all consistent for their population counterparts when the sparsity s is sufficiently large (i.e., larger than the population sparsity level) and the number of directions K is no bigger than the true dimension of the central subspace. Therefore, our method is flexible in the sense that the prespecified s and K do not have to be exactly correct. In practice, especially in exploratory data analysis, the number of sequentially extracted directions is often set to be small (i.e., $K = 1, 2,$ or 3). Determining the true central subspace dimension is a separate and important research topic in SDR (e.g., Bura and Yang; 2011; Luo and Li; 2016), and is beyond the scope of this study. Moreover, the prespecified sparsity level s combined with ℓ_0-regularization is potentially convenient for post-dimension reduction inference (Kim et al.; 2020), as in the post-selection inference of a canonical correlation analysis over subsets of variables with prespecified cardinalities (McKeague and Zhang; 2020).

Algorithm 1 Penalized eigen-decomposition of MDDM.

```
1. Input: $s, K, \widehat{\mathbf{M}}_1 = \widehat{\mathbf{M}} = \mathrm{MDDM}_n(\mathbf{X} \mid \mathbf{Y})$.
2. Initialize $\widehat{\boldsymbol{\beta}}_1^{(0)}$.
3. For $k = 1, \ldots, K$, do
    (a) Iterate over $t$ until convergence:
        i. Set $\widehat{\boldsymbol{\beta}}_k^{(t)} = \widehat{\mathbf{M}}_k \widehat{\boldsymbol{\beta}}_k^{(t-1)}$.
        ii. If $\|\widehat{\boldsymbol{\beta}}_k^{(t)}\|_0 \leq s$, set
                $$\widehat{\boldsymbol{\beta}}_k^{(t)} = \frac{\widehat{\boldsymbol{\beta}}_k^{(t)}}{\|\widehat{\boldsymbol{\beta}}_k^{(t)}\|_2};$$
            else
                $$\widehat{\boldsymbol{\beta}}_k^{(t)} = \frac{\mathrm{HT}(\widehat{\boldsymbol{\beta}}_k^{(t)}, s)}{\|\mathrm{HT}(\widehat{\boldsymbol{\beta}}_k^{(t)}, s)\|_2}$$
    (b) Set $\widehat{\boldsymbol{\beta}}_k = \widehat{\boldsymbol{\beta}}_k^{(t)}$ at convergence and $\widehat{\mathbf{M}}_{k+1} = \widehat{\mathbf{M}}_k - \widehat{\boldsymbol{\beta}}_k^{\mathrm{T}} \widehat{\mathbf{M}} \widehat{\boldsymbol{\beta}}_k \cdot \widehat{\boldsymbol{\beta}}_k \widehat{\boldsymbol{\beta}}_k^{\mathrm{T}}$.
4. Output $\widehat{\mathcal{S}}_{\mathbf{Y}|\mathbf{X}} = \mathrm{span}(\widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_K)$.
```

As pointed out by a referee, other sparse principal component analysis (PCA) methods can potentially be applied to decompose the MDDM. We choose to extend the algorithm in Yuan and Zhang (2013) to facilitate computation and theoretical development. For computationally efficient sparse PCA methods such as Zou et al. (2006); Witten et al. (2009), their theoretical properties are unfortunately unknown. Hence, we expect the theoretical study of their MDDM-variants to be very challenging. On the other hand, for the theoretically justified sparse PCA methods such as Vu and Lei (2013); Cai et al. (2013), the computation is less efficient.

4.2 Generalized eigenvalue problems with the MDDM

Now, we consider the general (arbitrary) covariance structure $\boldsymbol{\Sigma}_{\mathbf{X}}$. We continue to use $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$ to denote the nontrivial eigenvectors of $\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\text{span}(\text{MDDM}(\mathbf{X} \mid \mathbf{Y}))$ so that the central subspace is spanned by $\boldsymbol{\beta}$. Again, we assume that these eigenvectors are sparse. In principle, we could assume that $\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}$ is also sparse, and construct its estimate accordingly. However, $\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}$ is a nuisance parameter for our ultimate goal, and additional assumptions on it may unnecessarily limit the applicability of our method. Hence, we take a different approach.

To avoid estimating $\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}$, we note that $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$ can also be viewed as the generalized eigenvectors defined as follows, which is equivalent to (3.2):

$$\boldsymbol{\beta}_k = \arg\max_{\boldsymbol{\beta}} \boldsymbol{\beta}^{\mathrm{T}}\mathbf{M}\boldsymbol{\beta}, \text{ s.t. } \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\beta} = 1, \boldsymbol{\beta}_l^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\beta} = 0 \text{ for any } l < k. \tag{4.2}$$

Directly solving the generalized eigen-decomposition problem in (4.2) is not easy if we want to impose further penalties, because it is difficult to satisfy the orthogonality constraints. Therefore, we consider another form for (4.2) that does not involve the orthogonal constraints. This alternative form is based on the following lemma.

Lemma 1. *Let* $\lambda_j = \boldsymbol{\beta}_j^{\mathrm{T}}\mathbf{M}\boldsymbol{\beta}_j$ *and* $\mathbf{M}_k = \mathbf{M} - \boldsymbol{\Sigma}_{\mathbf{X}}(\sum_{j<k}\lambda_j\boldsymbol{\beta}_j\boldsymbol{\beta}_j^{\mathrm{T}})\boldsymbol{\Sigma}_{\mathbf{X}}$. *We have*

$$\boldsymbol{\beta}_k = \arg\max_{\boldsymbol{\beta}} \boldsymbol{\beta}^{\mathrm{T}}\mathbf{M}_k\boldsymbol{\beta}, \qquad s.t.\ \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\beta} = 1. \tag{4.3}$$

Motivated by Lemma 1, we consider the penalized problem that $\boldsymbol{\beta}_k = \arg\max_{\boldsymbol{\beta}} \boldsymbol{\beta}^{\mathrm{T}}\widehat{\mathbf{M}}_k\boldsymbol{\beta}$ such that $\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\beta} = 1, \|\boldsymbol{\beta}\|_0 \leq s$, where $\widehat{\mathbf{M}}_1 = \text{MDDM}_n(\mathbf{X} \mid \mathbf{Y})$ and $\widehat{\mathbf{M}}_k = \widehat{\mathbf{M}}_1 -$

$\widehat{\mathbf{\Sigma}}_{\mathbf{X}}\left(\sum_{l<k}\delta_l\widehat{\boldsymbol{\beta}}_l\widehat{\boldsymbol{\beta}}_l^{\mathrm{T}}\right)\widehat{\mathbf{\Sigma}}_{\mathbf{X}}$, for $k>1$, with $\delta_l=\widehat{\boldsymbol{\beta}}_l^{\mathrm{T}}\widehat{\mathbf{M}}\widehat{\boldsymbol{\beta}}_l$, and s is a tuning parameter. We adopt the RIFLE algorithm of Tan et al. (2018a) to solve this problem; see Algorithm 2. In our simulation studies, we consider a randomly generated initial value $\widehat{\boldsymbol{\beta}}_1^{(0)}$ and a fixed step size $\eta=1$, and observe reasonably good performance.

Although Algorithm 2 is a generalization of the RIFLE algorithm of Tan et al. (2018a), there are important differences between the two. On the one hand, the RIFLE algorithm extracts only the first generalized eigenvector, whereas Algorithm 2 is capable of estimating multiple generalized eigenvectors by properly deflating the MDDM. In SDR problems, the central subspace often has a structural dimension greater than one, and it is necessary to find more than one generalized eigenvector. Hence, Algorithm 2 is potentially more useful than the RIFLE algorithm, in practice. On the other hand, the usefulness of the RIFLE algorithm has been demonstrated in several statistical applications, including sparse sliced inverse regression. Here, Algorithm 2 decomposes the MDDM, which is the first time the penalized generalized eigenvector problem has been used to perform SDR in a slicing-free manner in high dimensions. A brief analysis of the computation complexity is included in Section S3 in the Supplementary Material.

5. Theoretical properties

In this section, we consider the theoretical properties of the generalized eigenvectors of $(\mathrm{MDDM}(\mathbf{X}\mid\mathbf{Y}),\mathbf{\Sigma}_{\mathbf{X}})$. Recall that if we know that $\mathbf{\Sigma}_{\mathbf{X}}=\mathbf{I}$, the generalized

Algorithm 2 Generalized eigen-decomposition of MDDM.

```
1. Input: $s, K, \widehat{\mathbf{M}}_1 = \widehat{\mathbf{M}}$, and step size $\eta > 0$.
2. Initialize $\widehat{\boldsymbol{\beta}}_1^{(0)}$.
3. For $k = 1, \ldots, K$, do
    (a) Iterate over $t$ until convergence:
        i. Set $\rho^{(t-1)} = \dfrac{(\widehat{\boldsymbol{\beta}}_k^{(t-1)})^{\mathrm{T}} \widehat{\mathbf{M}}_k \widehat{\boldsymbol{\beta}}_k^{(t-1)}}{(\widehat{\boldsymbol{\beta}}_k^{(t-1)})^{\mathrm{T}} \widehat{\boldsymbol{\Sigma}}_{\mathbf{X}} \widehat{\boldsymbol{\beta}}_k^{(t-1)}}$.
        ii. $\mathbf{C} = \mathbf{I} + (\eta/\rho^{(t-1)}) \cdot (\widehat{\mathbf{M}}_k - \rho^{(t-1)} \widehat{\boldsymbol{\Sigma}}_{\mathbf{X}})$
        iii. $\widetilde{\boldsymbol{\beta}}_k^{(t)} = \mathbf{C} \widehat{\boldsymbol{\beta}}_k^{(t-1)} / \|\mathbf{C} \widehat{\boldsymbol{\beta}}_k^{(t-1)}\|_2$.
        iv. $\widehat{\boldsymbol{\beta}}_k^{(t)} = \dfrac{\mathrm{HT}(\widetilde{\boldsymbol{\beta}}_k, s)}{\|\mathrm{HT}(\widetilde{\boldsymbol{\beta}}_k, s)\|_2}$
    (b) Set $\widetilde{\boldsymbol{\beta}}_k = \widehat{\boldsymbol{\beta}}_k^{(t)}$ at convergence and scale it to obtain $\widehat{\boldsymbol{\beta}}_k = \dfrac{\widetilde{\boldsymbol{\beta}}_k}{\sqrt{\widetilde{\boldsymbol{\beta}}_k^{\mathrm{T}} \widehat{\boldsymbol{\Sigma}}_{\mathbf{X}} \widetilde{\boldsymbol{\beta}}_k}}$.
    (c) Set $\widehat{\mathbf{M}}_{k+1} = \widehat{\mathbf{M}}_k - \widehat{\boldsymbol{\Sigma}}_{\mathbf{X}} \widehat{\boldsymbol{\beta}}_k^{\mathrm{T}} \widehat{\mathbf{M}} \widehat{\boldsymbol{\beta}}_k \cdot \widehat{\boldsymbol{\beta}}_k \widehat{\boldsymbol{\beta}}_k^{\mathrm{T}} \widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}$.
4. Output $\widehat{\mathcal{S}}_{\mathbf{Y}|\mathbf{X}} = \mathrm{span}(\widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_K)$.
```

eigenvectors reduce to eigenvectors, and can be estimated using Algorithm 1. If we do not have any information about $\boldsymbol{\Sigma}_{\mathbf{X}}$, we can find the generalized eigenvectors using Algorithm 2. Either way, we let $\boldsymbol{\beta}_k$, for $k = 1, \ldots, K$, be the first K (generalized) eigenvectors of MDDM($\mathbf{X} \mid \mathbf{Y}$). Throughout the proof, we let C denote a generic constant that can vary based on the context. We show the consistency of $\widehat{\boldsymbol{\beta}}_k$ by proving that $\eta_k = |\sin \Theta(\widehat{\boldsymbol{\beta}}_k, \boldsymbol{\beta}_k)| \leq Cs\epsilon$. We assume that K is fixed, and $s\epsilon \leq 1$. Recall that we define $\lambda_j = \boldsymbol{\beta}_j^{\mathrm{T}} \mathbf{M} \boldsymbol{\beta}_j$ as the (generalized) eigenvalue. Further, define

$d = \max_{k=1}^{K}\{\|\boldsymbol{\beta}_k\|_0\}$. When we study Algorithm 1 or Algorithm 2, we assume that $s = d + 2s'$, where $s' = Cd$ for a sufficiently large C. To apply the concentration inequalities for the MDDM, we restate Condition (C1) in terms of $\mathbf{X}$ and $\mathbf{Y}$ as Condition (C1$'$), along with other suitable conditions:

(C1$'$) There exist two positive constants σ_0 and C_0 such that $\mathrm{E}\{\exp(2\sigma_0\|\mathbf{Y}\|_q^2)\} \le C_0$ and $\sup_p \max_{1\le j\le p} \mathrm{E}\{\exp(2\sigma_0 X_j^2)\} \le C_0$.

(C2) There exist $\Delta > 0$ such that $\min_{k=1,\ldots,K}(\lambda_k - \lambda_{k+1}) \ge \Delta$.

(C3) There exist constants U, L that do not depend on n, p such that $L \le \lambda_K \le \lambda_1 \le U$.

(C4) As $n \to \infty$, $dn^{-1/2}(\log p)^{1/2}(\log n)^{3/2} \to 0$.

Condition (C2) guarantees that the eigenvectors are well defined. Condition (C3) imposes bounds on the eigenvalues of $\mathbf{M}$. Researchers often impose similar assumptions on the covariance matrix to achieve consistent estimation. Condition (C4) restricts the growth rate of p, d with respect to n. Note that d is the population sparsity level of $\boldsymbol{\beta}_k$, and s is the user-specified sparsity level in Algorithms 1 and 2. If we fix d, the dimension is allowed to grow at the rate $\log p = o(n\log^{-3} n)$. When we allow d to diverge, we require it to diverge more slowly than $\left\{n/(\log p \log^3 n)\right\}^{\frac{1}{2}}$.

We present the nonasymptotic results for Algorithm 1 in the following theorem, where the constants $D_1, D_2, \sigma_0, \gamma, C_0$ are defined previously in Theorem 1 under Condition (C1).

Theorem 2. *Assume that Conditions (C1′), (C2), and (C3) hold, and* $\mathbf{\Sigma}_{\mathbf{X}} = \mathbf{I}$. *Further, assume that there exists* $\theta \in (0, 1/2)$ *such that, for* $k = 1, \ldots, K$, *we have* $(\widehat{\boldsymbol{\beta}}_k^0)^{\mathrm{T}} \boldsymbol{\beta}_k \geq 2\theta$, *and*

$$\mu \quad = \quad \sqrt{[1 + 2\{(\frac{d}{s'})^{1/2} + \frac{d}{s'}\}]\{1 - 0.5\theta(1+\theta)(1 - (\gamma^*)^2)\}} \quad < \quad 1, \quad (5.1)$$

where $\gamma^* = \dfrac{\lambda_K - \frac{3}{4}\Delta}{\lambda_K - \frac{1}{4}\Delta}$. *Then, there exists a positive integer* $n_0 = n_0(\sigma_0, C_0, q) < \infty$, $\gamma = \gamma(\sigma_0, C_0, q) \in (0, 1/2)$, *and a finite positive* $D_0 = D_0(\sigma_0, C_0, q)$, *such that when* $n > n_0$, *we have* $D_0 n^{-\gamma} < \dfrac{\Delta}{4s}$, *and for any* $D_0 n^{-\gamma} < \epsilon < \min\{\dfrac{\Delta}{4s}, \theta\}$, *with a probability greater than* $1 - 54p^2 \exp\left\{-\dfrac{\epsilon^2 n}{36 \log^3 n}\right\}$,

$$|\sin \mathbf{\Theta}(\widehat{\boldsymbol{\beta}}_k, \boldsymbol{\beta}_k)| \leq Cs\epsilon, \quad k = 1, \ldots, K. \quad (5.2)$$

Let $n^{-1/2}(\log p)^{1/2} \log^{3/2} n \ll \epsilon \ll d^{-1}$. Then, Theorem 2 directly implies the following asymptotic result that justifies the consistency of our estimator.

Corollary 1. *Assume that Conditions (C1′) and (C2)–(C4) hold. Suppose there exists* $\gamma > 0$ *such that* $d \leq s \ll \min\{n^{\gamma}, \dfrac{n^{\frac{1}{2}}}{(\log p)^{\frac{1}{2}} (\log n)^{\frac{3}{2}}}\}$. *Under the conditions in Theorem 2, the quantities* $|\sin \mathbf{\Theta}(\widehat{\boldsymbol{\beta}}_k, \boldsymbol{\beta}_k)| \to 0$ *with probability tending to one, for* $k = 1, \ldots, K$.

Corollary 1 reveals that, without specifying a model, Algorithm 1 can achieve consistency when p grows at an exponential rate of n. To be exact, we can allow $\log p = o\{n/(d^2 \log^3 n)\}$. Here, the theoretical results are established for the output of Algorithm 1, instead of the solution of the optimization problem (4.1). Note that it is possible for there to be a difference between the theoretical optimal solution of (4.1) and the estimate we use in practice, because the optimization problem is nonconvex, and, numerically, we might not achieve the global maximum. Thus, it might be more meaningful to study the property of the estimate obtained as the output of Algorithm 1. The above theorem guarantees that the estimate we use in practice has the desired theoretical properties.

Although our rate in Theorem 2 is not as high as that of a sparse SIR, as established very recently by Lin et al. (2018) when $\mathbf{\Sigma}_{\mathbf{X}} = \mathbf{I}$ and for general $\mathbf{\Sigma}_{\mathbf{X}}$ by Lin et al. (2019), and by Tan et al. (2020), we have some unique advantages over these proposals. For simplicity, we assume that d is fixed in the subsequent discussion. First, both SIR methods require an estimation of within-slice means, rather than the MDDM. As shown in Theorem 1, the MDDM converges to its population counterpart at a slower rate than the sample within-slice mean does. However, by adopting the MDDM, we no longer need to determine the slicing scheme, and we do not encounter the curse of dimensionality when slicing a multivariate response. Second, Lin et al. (2018) only achieve the optimal rate when $p = o(n^2)$, and cannot handle ultrahigh dimensions. In contrast, Algorithm 1 allows p to diverge at an exponential rate of n, and is more suitable for ultrahigh-dimensional data. Third, although Tan et al. (2020) achieve consistency when $\log p = o(n)$, their model assumptions are much more restrictive. For example, they assume that Y is categorical and $\mathbf{X}$ is normal within each slice of Y, and they randomly split the data set to form independent batches to facilitate their proofs, which is not done in their numerical studies. The theoretical properties for their proposal are unclear beyond the (conditionally) Gaussian model and without sample splitting. In contrast, our method makes no model assumption between $\mathbf{X}$ and Y, and our theory requires no sample splitting. Thus, our results are more widely applicable, and we obtain good rates. Furthermore, unlike the theory in Tan et al. (2020), our theoretical result characterizes the same method we use in practice. Moreover, the convergence rate of our method has an additional factor of $\log^3(n)$ compared to Tan et al. (2020), which grows at a slow rate of n that only imposes mild restriction on the dimensionality. For example, for any positive constant $\xi \in (0,1)$, if $\log p = O(n^{1-\xi})$, our method is consistent. In this sense, although we cannot handle the optimal dimensionality of $\log p = o(n)$, the gap is very small.

Next, we consider the penalized generalized eigen-decomposition in Algorithm 2. We assume that the step size η satisfies $\eta\lambda_{\max}(\boldsymbol{\Sigma}_{\mathbf{X}}) < 1/2$, and

$$\sqrt{[1 + 2\{(\frac{d}{s'})^{1/2} + \frac{d}{s'}\}][1 - \frac{\eta\lambda_{\min}(\boldsymbol{\Sigma}_{\mathbf{X}})(1 - \frac{\lambda_2}{\lambda_1})}{16\kappa(\boldsymbol{\Sigma}_{\mathbf{X}}) + 16\frac{\lambda_2}{\lambda_1}}]} < 1, \tag{5.3}$$

where $\lambda_{\max}(\boldsymbol{\Sigma}_{\mathbf{X}})$, $\lambda_{\min}(\boldsymbol{\Sigma}_{\mathbf{X}})$, and $\kappa(\boldsymbol{\Sigma}_{\mathbf{X}})$ are the largest eigenvalue, smallest eigenvalue, and condition number of $\boldsymbol{\Sigma}_{\mathbf{X}}$, respectively. The nonasymptotic results are as follows.

Theorem 3. *Assume that Conditions (C1′), (C2), and (C3) hold. Suppose there exists $\gamma \in (0, 1/2)$ such that $d \leq s = o(n^{\gamma})$, and there exists a constant $\theta(\kappa(\boldsymbol{\Sigma}_{\mathbf{X}}), \lambda_{\max}(\boldsymbol{\Sigma}_{\mathbf{X}}), \Delta, \lambda_1, \lambda_K, \eta)$*

$\in (0,1)$ such that $\frac{(\widehat{\boldsymbol{\beta}}_k^0)^{\mathrm{T}}\boldsymbol{\beta}_k}{\|\widehat{\boldsymbol{\beta}}_k^0\|_2} \geq 1-\theta$. Then, there exists a positive integer $n_0 = n_0(s_0, C_0) < \infty$ and four finite positive constants $D_0 = D_0(\gamma, \sigma_0, C_0) \in (0,\infty)$, $D_1 = D_1(C_0) \in (0,\infty)$, $D_2 = D_2(\sigma_0, C_0) \in (0,\infty)$, and $\epsilon_0 = \epsilon_0(\lambda_1, \lambda_2, \lambda_{\min}(\boldsymbol{\Sigma}), \Delta)$ such that for any ϵ that satisfies $s\epsilon < \epsilon_0$ and $D_0 n^{-\gamma} < \epsilon \leq 1$, with a probability greater than $1 - D_1 p^2 n \exp\{-D_2\epsilon^2 n/\log^3 n\}$, we have $|\sin\Theta(\widehat{\boldsymbol{\beta}}_k, \boldsymbol{\beta}_k)| \leq Cs\epsilon$, for $k = 1, \ldots, K$.

Theorem 3 is proved by showing that $\widehat{\mathbf{M}}_k$ and $\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}$ are close to their counterparts in the sense that $\mathbf{u}^{\mathrm{T}}\widehat{\mathbf{M}}_k\mathbf{u}$ and $\mathbf{u}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}\mathbf{u}$ are close to $\mathbf{u}^{\mathrm{T}}\mathbf{M}_k\mathbf{u}$ and $\mathbf{u}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{u}$, respectively, for any $\mathbf{u}$ with only s nonzero elements. Then, because Algorithm 2 is a generalization of the RIFLE algorithm [Tan et al. (2018a)], some properties of the latter allow us to establish the consistency of Algorithm 2. By comparison, our proofs are significantly more involved than that in Tan et al. (2018a), because we have to estimate K generalized eigenvectors, instead of just the first one. We need to carefully control the error bounds to guarantee that the estimation errors do not accumulate to a higher order beyond the first generalized eigenvector.

Analogous to Corollary 1, we can easily obtain asymptotic consistency results by translating Theorem 3.

Corollary 2. *Assume that Conditions (C1)–(C4) hold. Suppose there exists $\gamma \in (0, 1/2)$ such that $d \leq s \ll \min\{n^\gamma, \frac{n^{\frac{1}{2}}}{(\log p \log^3 n)^{\frac{1}{2}}}\}$. Under the conditions in Theorem 3, the quantities $|\sin\boldsymbol{\Theta}(\widehat{\boldsymbol{\beta}}_k, \boldsymbol{\beta}_k)| \to 0$ with a probability tending to one, for $k = 1, \ldots, K$.*

Corollary 2 shows that Algorithm 2 produces consistent estimates of the generalized eigenvectors $\boldsymbol{\beta}_k$ even when p grows at an exponential rate of the sample size n, and thus is suitable for ultrahigh-dimensional problems. Similarly to Corollary 1, Corollary 2 has no gap between the theory and the numerical outputs, because it concerns the outputs of Algorithm 2. Note that the dimensionality in Corollary 2 is the same as that in Corollary 1. Thus, with a properly chosen step size η, the penalized generalized eigen-decomposition is intrinsically no more difficult than the penalized eigen-decomposition. However, if we have knowledge about $\boldsymbol{\Sigma}_{\mathbf{X}}$ being the identity matrix, it is still beneficial to exploit such information and use Algorithm 1, because Algorithm 1 does not involve the step size and is more convenient in practice. Furthermore, although Algorithm 2 does not achieve the same rate of convergence as recent sparse SIR proposals, it has many practical and theoretical advantages, just as for Algorithm 1, as discussed earlier.

Finally, note that our theoretical studies require conditions on the initial value. Specifically, we require the initial value to be non-orthogonal to the truth. This is a common technical condition for iterative algorithms; see Yuan and Zhang (2013) and Tan et al. (2018a), for example. Such conditions do not seem critical for our algorithms to work in practice. In our numerical studies, we use randomly generated initial values, and the performance of our methods appears to be competitive.

6. Numerical studies

6.1 Simulations

We compare our slicing-free approaches with state-of-the-art high-dimensional extensions of SIR estimators. We consider both univariate and multivariate response settings. Specifically, for the univariate response simulations, we include Rifle-SIR (Tan et al.; 2018a) and Lasso-SIR (Lin et al.; 2019) as the two main competitors; for the multivariate response simulations, we mainly compare our method with the projective resampling approach to SIR (PR-SIR, Li et al.; 2008), which is a computationally expensive method that repeatedly projects the multivariate response to one-dimensional subspaces. For Rifle-SIR, we adopt the Rifle algorithm to estimate the leading eigenvector of the sample matrix $\mathrm{cov}\{\mathrm{E}(\mathbf{X} \mid Y)\}$ based on slicing. In addition, we include the oracle-SIR as a benchmark method, where we perform a SIR on the subset of truly relevant variables (hence, a low-dimensional estimation problem). For all these SIR-based methods, we include two different slicing schemes by setting the number of slices to be 3 and 10, where 3 is the minimal number of slices required to obtain our two-dimensional central subspace, and 10 is a typical choice in the literature. To evaluate the performance of these SDR methods, we use the subspace estimation error, defined as $\mathcal{D}(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = \|\mathbf{P}_{\widehat{\boldsymbol{\beta}}} - \mathbf{P}_{\boldsymbol{\beta}}\|_F / \sqrt{2K}$, where $\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta} \in \mathbb{R}^{p \times K}$ are the estimated and the true basis matrices, respectively, of the central subspace, and $\mathbf{P}_{\widehat{\boldsymbol{\beta}}}, \mathbf{P}_{\boldsymbol{\beta}} \in \mathbb{R}^{p \times p}$ are the corresponding projection matrices. This subspace estimation error is always between zero and one, and a small value indicates a good estimation.

First, we consider the following six models for a univariate response regression: $\mathcal{M}_1$ and $\mathcal{M}_2$ are single-index models (i.e., $K = 1$); $\mathcal{M}_3$–$\mathcal{M}_5$ are multiple-index models (i.e., $K = 2$); and $\mathcal{M}_6$ is an isotropic PFC model with $K = 1$. Specifically,

$$\mathcal{M}_1 : Y = (\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}) + \sin(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}) + \epsilon, \quad \mathcal{M}_2 : Y = 2\arctan(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}) + 0.1(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X})^3 + \epsilon,$$

$$\mathcal{M}_3 : Y = \frac{\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}}{0.5 + (1.5 + \boldsymbol{\beta}_2^{\mathrm{T}}\mathbf{X})^2} + 0.2\epsilon, \quad \mathcal{M}_4 : Y = \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X} + (\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}) \cdot (\boldsymbol{\beta}_2^{\mathrm{T}}\mathbf{X}) + 0.3\epsilon,$$

$$\mathcal{M}_5 : Y = sign(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}) \cdot \log(|\boldsymbol{\beta}_2^{\mathrm{T}}\mathbf{X} + 5|) + 0.2\epsilon, \quad \mathcal{M}_6 : \mathbf{X} = 2\boldsymbol{\beta}_1 \exp(Y)/3 + 0.5\boldsymbol{\epsilon},$$

where $\mathbf{X} \sim N_p(0, \boldsymbol{\Sigma}_{\mathbf{X}})$ and $\epsilon \sim N(0,1)$ for $\mathcal{M}_1$–$\mathcal{M}_5$, and $Y \sim N(0,1)$ and $\boldsymbol{\epsilon} \sim N_p(0, \mathbf{I}_p)$ for the isotropic PFC model ($\mathcal{M}_6$). The sparse directions in the central subspace $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathbb{R}^p$ are orthogonal, because we let the first $s = 6$ elements in $\boldsymbol{\beta}_1$ and elements 6 - 12 in $\boldsymbol{\beta}_2$ be $1/\sqrt{6}$ (all other elements are zero). For $\mathcal{M}_1$–$\mathcal{M}_5$, we consider both the independent predictor setting with $\boldsymbol{\Sigma}_{\mathbf{X}} = \mathbf{I}_p$ and the correlated predictor setting with an auto-regressive correlation that $\Sigma_X(i,j) = 0.5^{|i-j|}$, for $i, j = 1, 2, ..., p$. For each model setting, we vary the sample size $n \in \{200, 500, 800\}$ and predictor dimension $p \in \{200, 500, 800, 1200, 2000\}$, and simulate 1000 independent data sets.

For our method, we apply the generalized eigen-decomposition algorithm (Algorithm 2) in all six models (even when the covariance of $\mathbf{X}$ is the identity matrix). In the single-index models $\mathcal{M}_1$ and $\mathcal{M}_2$, we use a random initialization ($\widehat{\boldsymbol{\beta}}^{(0)}$ is generated randomly from p-dimensional standard normal) for our algorithm and Rifle-SIR to demonstrate their robustness to initialization. The step size in the algorithm is simply fixed as

$\eta = 1$. For the more challenging multiple-index models, $\mathcal{M}_3 - \mathcal{M}_6$, we consider the best-case scenarios for each method. Therefore, the true parameter $\boldsymbol{\beta}$ is used as the initial value, and an optimal $\eta \in \{0.1, 0.2, \ldots, 1.0\}$ is selected from a separate training sample with 400 observations. The results based on 1000 replications for $n = 200$ and $p = 800$ are summarized in Table 1; the remaining results can be found in the Supplementary Material. Overall, the slicing-free MDDM approach is much more accurate than existing SIR-based methods. It is almost as accurate as the oracle-SIR. Moreover, it is clear that SIR-type methods are rather sensitive to the choice of the number of slices.

Next, we consider three multivariate response models, where the response dimension is $q = 4$. These three models are a multivariate linear model, a single-index heteroschedastic error model, and an isotropic PFC model. The predictors satisfy $\mathbf{X} \sim N_p(0, \mathbf{I}_p)$ in the following two forward regression models. Therefore, we apply Algorithm 1 for our method under models $\mathcal{M}_7$ and $\mathcal{M}_8$. For the isotropic PFC model $\mathcal{M}_9$, where $\mathbf{X} \mid \mathbf{Y} \sim N_p(\boldsymbol{\beta} f(\mathbf{Y}), \mathbf{I}_p)$, we still apply Algorithm 2, to be consistent with the univariate case. For the projective resampling methods, PR-SIR and PR-Oracle-SIR, we generate a sufficiently large number of $n \log(n)$ random projections so that the PR methods reach their fullest potential.

$\mathcal{M}_7$: $Y_1 = \boldsymbol{\beta}_1^{\mathrm{T}} X + \epsilon_1$, $Y_2 = \boldsymbol{\beta}_2^{\mathrm{T}} \mathbf{X} + \epsilon_2$, $Y_3 = \epsilon_3$, and $Y_4 = \epsilon_4$. The errors $(\epsilon_1, \ldots, \epsilon_4)$ are independent standard normal, except for $\mathrm{cov}(\epsilon_1, \epsilon_2) = -0.5$. For this model, the central subspace is spanned by $\boldsymbol{\beta}_1 = (1, 0, 0, 0, ..., 0)^{\mathrm{T}}$ and

$\boldsymbol{\Sigma}_{\mathbf{X}}$		MDDM		Oracle-SIR(3)		Oracle-SIR(10)		Rifle-SIR(3)		Rifle-SIR(10)		LassoSIR(3)		LassoSIR(10)	
		Error	SE	Error	SE	Error	SE	Error	SE	Error	SE	Error	SE	Error	SE
$\mathbf{I}_p$	$\mathcal{M}_1$	10.1	0.1	12.5	0.1	10.3	0.1	25.2	1.0	53.7	1.4	37.9	0.4	59.9	0.7
	$\mathcal{M}_2$	10.3	0.1	13.1	0.1	10.6	0.1	26.1	1.0	54.7	1.4	40.1	0.4	61.5	0.7
	$\mathcal{M}_3$	17.7	0.2	40.8	0.2	27.7	0.2	71.3	0.0	71.2	0.0	76.5	0.2	85.0	0.2
	$\mathcal{M}_4$	23.0	0.2	45.8	0.3	36.4	0.3	71.9	0.0	71.6	0.0	85.2	0.2	91.5	0.2
	$\mathcal{M}_5$	30.8	0.6	28.8	0.2	22.1	0.1	71.6	0.0	71.2	0.0	71.2	0.3	81.3	0.3
AR	$\mathcal{M}_1$	18.7	0.3	21.0	0.2	17.6	0.2	34.7	0.8	39.8	1.1	35.3	0.3	35.5	0.3
	$\mathcal{M}_2$	14.2	0.2	20.7	0.2	14.8	0.2	33.1	0.7	33.6	1.1	34.6	0.3	30.5	0.3
	$\mathcal{M}_3$	25.2	0.3	44.6	0.2	34.1	0.2	71.5	0.0	71.3	0.0	54.8	0.2	47.1	0.3
	$\mathcal{M}_4$	59.1	0.5	75.1	0.2	69.9	0.3	81.0	0.2	78.7	0.2	89.7	0.2	92.1	0.2
	$\mathcal{M}_5$	46.2	0.6	46.4	0.2	35.5	0.2	73.8	0.1	72.4	0.0	66.5	0.2	61.4	0.3
PFC	$\mathcal{M}_6$	34.6	0.6	48.9	0.5	33.4	0.5	40.1	0.7	30.8	0.6	70.7	0.0	70.7	0.0

Table 1: Averaged subspace estimation errors and the corresponding standard errors (after multiplied by 100) for univariate response models ($n = 200, p = 800$).

$\boldsymbol{\beta}_2 = (0, 2, 1, 0, ..., 0)^{\mathrm{T}}$.

$\mathcal{M}_8$: $Y_1 = \exp(\epsilon_1)$ and $Y_i = \epsilon_i$, for $i = 2, 3, 4$, where $(\epsilon_1, \dots, \epsilon_4)$ are independent standard normal, except for $\mathrm{cov}(\epsilon_1, \epsilon_2) = \sin(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})$. For this model, the central subspace is $\boldsymbol{\beta} = (0.8, 0.6, 0, 0, \dots, 0)^{\mathrm{T}}$. Note that, marginally, each response is independent of $\mathbf{X}$.

$\mathcal{M}_9$: $\mathbf{X} = \boldsymbol{\beta}\left(\frac{1}{3}\sin(Y_1) + \frac{2}{3}\exp(Y_2) + Y_3\right) + \boldsymbol{\epsilon}$, where $\boldsymbol{\beta} = (1/\sqrt{6} \cdot 1_6, 0_{p-6})$, and $\boldsymbol{\epsilon} \sim N(0, \mathbf{I}_p)$. Hence, $\mathcal{S}_{\mathbf{Y}|\mathbf{X}} = \mathrm{span}(\boldsymbol{\beta})$.

Again, we consider various sample sizes and predictor dimension setups, each with 1000 replicates. We summarize the subspace estimation errors in Table 2. For $p = 800$ and 1200, the results are gathered in Section S1 in the Supplementary Material. It is clear that the proposed MDDM approach is much better than PR-SIR, and also improves much faster than PR-SIR does when we increase the sample size. Note too that the MDDM method perform better in inverse regression models, such as the isotropic PFC model, than it does in forward regression models, such as the linear model and index models. This finding is more apparent in the multivariate response simulations than in the univariate response simulations. This is expected, because the MDDM directly targets the inverse regression subspace, which is more directly driven by the response in the isotropic PFC models.

6.2 Real-Data Illustration

In this section, we use our method to analyze the NCI-60 data set (Shoemaker; 2006) that contains microRNA expression profiles and cancer drug activity measurements on the NCI-60 cell lines. The multivariate response is the cancer drug activities of $q = 15$ drugs; the predictor is $p = 365$ different microRNA; the sample size is $n = 60$.

First, we examine the predictive performance of our method on 500 random training—testing sample splits; each time, we randomly pick five observations to form the test set. We consider $K = 5$ for all methods. For the MDDM, we include both the eigen-decomposition (Algorithm 1) and the generalized eigen-decomposition

		$n = 100$						$n = 200$						$n = 400$					
		$p = 100$		$p = 200$		$p = 400$		$p = 100$		$p = 200$		$p = 400$		$p = 100$		$p = 200$		$p = 400$	
		Error	SE	Error	SE	Error	SE	Error	SE	Error	SE	Error	SE	Error	SE	Error	SE	Error	SE
$\mathcal{M}_7$	MDDM	37.1	0.5	39.8	0.5	42.5	0.5	24.0	0.4	25.3	0.4	26.9	0.4	16.1	0.3	17.3	0.3	18.6	0.3
	PR-Oracle-SIR(3)	12.6	0.2	12.2	0.2	12.0	0.2	8.8	0.1	8.5	0.1	8.7	0.1	5.9	0.1	5.8	0.1	5.8	0.1
	PR-Oracle-SIR(10)	16.2	0.3	15.7	0.3	15.6	0.3	9.6	0.2	9.4	0.2	95.2	0.2	6.0	0.1	6.0	0.1	6.0	0.1
	PR-SIR(3)	79.9	0.1	88.2	0.1	93.5	0.0	67.9	0.1	79.3	0.1	87.8	0.0	54.6	0.1	67.6	0.1	79.0	0.0
	PR-SIR(10)	83.5	0.1	90.6	0.1	94.9	0.0	70.1	0.1	81.6	0.1	90.1	0.1	55.3	0.1	68.2	0.1	80.2	0.1
$\mathcal{M}_8$	MDDM	79.4	0.9	85.8	0.8	90.0	0.7	55.9	1.2	61.0	1.2	68.4	1.2	27.1	0.9	30.3	1.0	31.0	1.0
	PR-Oracle-SIR(3)	40.9	0.9	41.3	0.9	41.4	0.9	26.0	0.7	24.9	0.7	25.0	0.6	14.9	0.4	14.9	0.4	15.0	0.4
	PR-Oracle-SIR(10)	44.1	0.9	43.8	0.9	43.5	0.9	25.1	0.6	23.7	0.6	24.1	0.6	13.1	0.3	13.0	0.3	13.2	0.3
	PR-SIR(3)	99.3	0.0	99.7	0.0	99.8	0.0	99.2	0.0	99.7	0.0	99.8	0.0	98.8	0.0	99.6	0.0	99.8	0.0
	PR-SIR(10)	99.3	0.0	99.7	0.0	99.9	0.0	99.1	0.0	99.6	0.0	99.8	0.0	98.4	0.1	99.6	0.0	99.8	0.0
$\mathcal{M}_9$	MDDM	15.3	0.3	15.4	0.3	15.7	0.3	9.9	0.1	10.1	0.1	10.0	0.1	7.1	0.1	7.2	0.1	7.1	0.1
	PR-Oracle-SIR(3)	15.2	0.2	15.2	0.2	14.9	0.2	10.5	0.1	10.6	0.1	10.5	0.1	7.5	0.1	7.6	0.1	7.4	0.1
	PR-Oracle-SIR(10)	13.8	0.2	13.9	0.2	13.6	0.2	9.4	0.1	9.7	0.1	9.6	0.1	6.8	0.1	6.8	0.1	6.7	0.1
	PR-SIR(3)	58.5	0.2	72.3	0.2	84.0	0.2	44.6	0.1	58.2	0.1	71.4	0.1	33.1	0.1	44.6	0.1	57.9	0.1
	PR-SIR(10)	54.8	0.2	68.5	0.2	80.6	0.2	41.1	0.2	54.3	0.2	67.7	0.2	30.2	0.1	41.0	0.1	54.2	0.1

Table 2: Averaged subspace estimation errors and the corresponding standard errors (after multiplying by 100) for multivariate response models.

(Algorithm 2). To distinguish between the two versions of the MDDM, we have "MDDM-ID" for the eigen-decomposition approach, because it implicitly assumes that the covariance of $\mathbf{X}$ or the conditional covariance of $\mathbf{X} \mid \mathbf{Y}$ is a constant times the identity matrix. We use random initial values, and choose the sparsity level to be $s = 25$ in the way described in Section S2 in the Supplementary Material. Then, the five reduced predictors $\boldsymbol{\beta}_k^{\mathrm{T}}\mathbf{X}$, for $k = 1, \ldots, 5$, are fed into a generalized additive

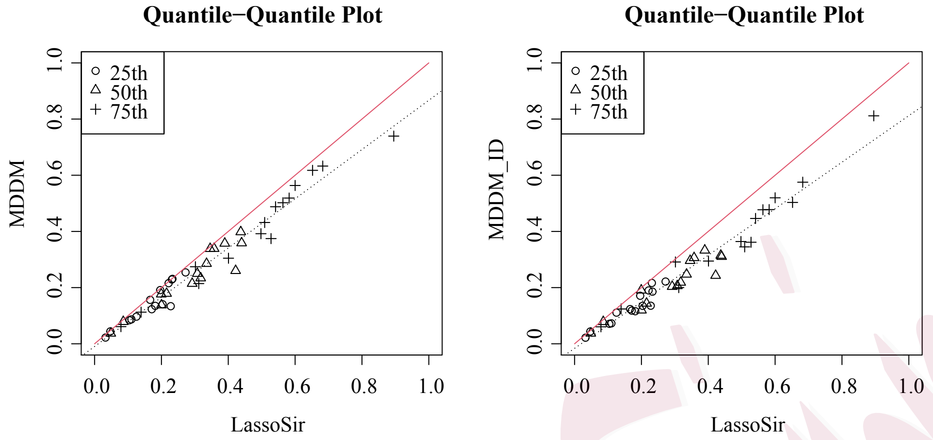

Figure 1: Quantile–quantile plots for prediction error comparisons between MDDM and Lasso-SIR (left panel), and between MDDM-ID and Lasso-SIR (right panel). Each point corresponds to the prediction mean squared errors for one of the $q = 15$ response variables, where different shapes represents different quantiles.

model for each drug. Finally, we evaluate the mean squared prediction error based on the test sample. The Rifle-SIR can only estimate a one-dimensional subspace, which did not yield an accurate prediction in this data set. Hence, for comparison, we compute five leading directions from the Lasso-SIR. The 25th, 50th, and 75th percentiles of the squared prediction errors for each of the 15 responses for all three models are obtained, and we construct quantile–quantile plots in Figure 1. The red line is the $y = x$ line, and the black dashed line is a simple linear regression fit for the results indicated by the y-axis label against that indicated by the x-axis. Clearly, for all the quantiles and for all the response variables, the MDDM results (MDDM or

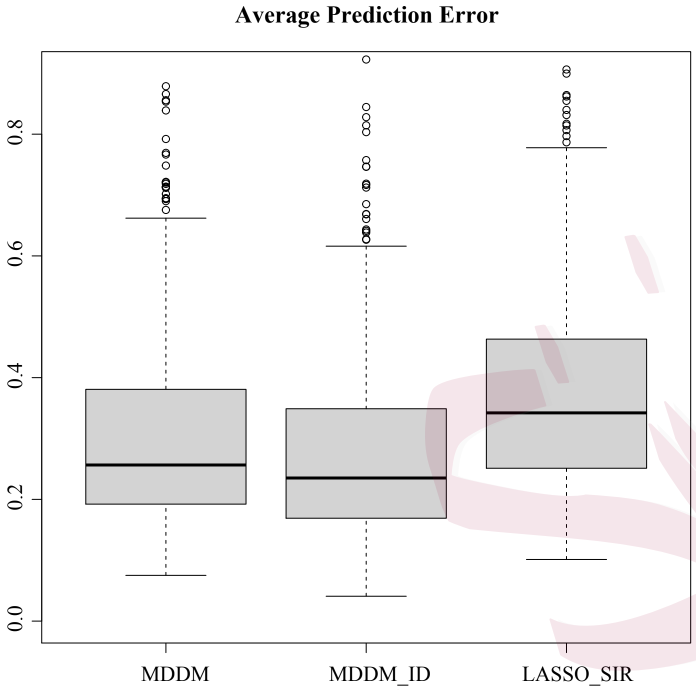

Figure 2: The averaged prediction error over 500 training-testing sample splits and over $q = 15$ response variables.

MDDM-ID) are better than those of Lasso-SIR in terms of prediction. In addition, we construct side-by-side box plots of the prediction error averaged over all response variables in Figure 2 to evaluate the overall improvement. Interestingly, the MDDM-ID is slightly better than the MDDM approach. This is likely because of the small sample size. With a training sample of size 55, the sample covariance of $p = 365$ variables is difficult to estimate accurately. We include additional real-data analysis results in Section S2 in the Supplementary Material.

7. Discussion

We have proposed a slicing-free high-dimensional SDR method based on a penalized eigen-decomposition of a sample MDDM. Our proposal is motivated by the usefulness of the MDDM for dimension reduction, and yields a relatively straightforward implementation of the recently developed RIFLE algorithm (Tan et al.; 2018a) by simply replacing the slicing-based estimator with the sample MDDM. Our methodology and implementation involve no slicing, and treats univariate and multivariate responses in a unified fashion. Theoretical support and finite-sample investigations provide convincing evidence that the MDDM is a very competitive alternative to SIR, and may be used as a surrogate for an SIR-based estimator in many related SDR problems.

As with most SDR methods, our proposal requires the linearity condition, the violation of which can make SDR very challenging. Existing works that relax the linearity condition are often practically difficult, owing to excessive computational costs, and cannot be easily extended to high dimensions (Cook and Nachtsheim; 1994; Ma and Zhu; 2012). One potentially useful approach is to transform data before SDR to alleviate obvious violations of the linearity assumption (Mai and Zou; 2015). In addition, we observe from our simulation studies that the RIFLE algorithm requires choosing several tuning parameters, such as the step size and the initial value, and that the optimization error could depend on these tuning parameters in a nontrivial way. Further investigation on the optimization error and data-driven choices for these tuning parameters are desirable, and are left for future research.

As pointed out by a referee, many SDR methods beyond SIR involve slicing. It will be interesting to study how to perform them in a slicing-free fashion as well. For example, Cook and Weisberg (1991) attempt to perform dimension reduction by estimating the conditional covariance of $\mathbf{X}$, while Yin and Cook (2003) consider the conditional third moment. These methods slice the response to estimate the conditional moments. In the future, one can develop slicing-free methods to estimate these higher-order moments and conduct SDR.

Supplementary Material

The online Supplementary Material provides additional simulation results and proofs.

Acknowledgements

The authors are grateful to the Editor, Associate Editor, and anonymous referees, whose suggestions led great improvement of this work. The authors contributed equally to this work and are listed in alphabetical order. Mai and Zhang's research was supported, in part, by the National Science Foundation grant CCF-1908969. Shao's research was supported, in part, by the National Science Foundation grant DMS-1607489.

References

Bura, E. and Cook, R. D. (2001). Extending sliced inverse regression: The weighted chi-squared test, *Journal of the American Statistical Association* **96**: 996–1003.

Bura, E. and Yang, J. (2011). Dimension estimation in sufficient dimension reduction: a unifying approach, *Journal of Multivariate analysis* **102**(1): 130–142.

Cai, T. T., Ma, Z. and Wu, Y. (2013). Sparse pca: Optimal rates and adaptive estimation, *Annals of Statistics* **41**: 3074–3110.

Chen, X., Zou, C., Cook, R. D. et al. (2010). Coordinate-independent sparse sufficient dimension reduction and variable selection, *The Annals of Statistics* **38**(6): 3696–3723.

Chiaromonte, R., Cook, R. D. and Li, B. (2002). Sufficient dimension reduction in regressions with categorical predictors, *The Annals of Statistics* **30**: 475–497.

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

Cook, R. D. and Forzani, L. (2009). Principal fitted components for dimension reduction in regression, *Statistical Science* **485**: 485–501.

Cook, R. D. and Nachtsheim, C. J. (1994). Reweighting to achieve elliptically contoured covariates in regression, *Journal of the American Statistical Association* **89**(426): 592–599.

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

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

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

Hsing, T. and Carroll, R. (1992). An asymptotic theory for sliced inverse regression, *Ann. Statist* **20**: 1040–1061.

Kim, K., Li, B., Yu, Z., Li, L. et al. (2020). On post dimension reduction statistical inference, *Annals of Statistics* **48**(3): 1567–1592.

Lee, C. E. and Shao, X. (2018). Martingale difference divergence matrix and its application to dimension reduction for stationary multivariate time series, *J. Amer. Statist. Assoc.* **113**: 216–229.

Li, B. (2018). *Sufficient Dimension Reduction, Methods and Applications with R*, CRC Press.

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

Li, B., Wen, S. and Zhu, L. (2008). On a projective resampling method for dimension reduction with multivariate responses, *Journal of the American Statistical Association* **103**(483): 1177–1186.

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

Li, L. (2007). Sparse sufficient dimension reduction, *Biometrika* **94**(3): 603–613.

Lin, Q., Li, X., Huang, D. and Liu, J. (2020). On the optimality of sliced inverse regression in high dimensions, *Annals of Statistics, to appear* .

Lin, Q., Zhao, Z. and Liu, J. (2018). On consistency and sparsity for sliced inverse regression in high dimension, *Annals of Statistics* **46**(2): 580–610.

Lin, Q., Zhao, Z. and Liu, J. S. (2019). Sparse sliced inverse regression via lasso, *Journal of the American Statistical Association* **114**: 1726–1739.

Luo, W. and Li, B. (2016). Combining eigenvalues and variation of eigenvectors for order determination, *Biometrika* **103**(4): 875–887.

Ma, Y. and Zhu, L. (2012). A semiparametric approach to dimension reduction, *Journal of the American Statistical Association* **107**(497): 168–179.

Mai, Q. and Zou, H. (2015). Nonparametric variable transformation in sufficient dimension reduction, *Technometrics* **57**(1): 1–10.

McKeague, I. W. and Zhang, X. (2020). Significance testing for canonical correlation analysis in high dimensions, *arXiv preprint arXiv:2010.08673* .

Shao, X. and Zhang, J. (2014). Martingale difference correlation and its use in high dimensional variable screening, *J. Amer. Statist. Assoc.* **109**: 1302–1318.

Shoemaker, R. H. (2006). The nci60 human tumour cell line anticancer drug screen, *Nature Reviews Cancer* **6**(10): 813–823.

Tan, K. M., Wang, Z., Liu, H. and Zhang, T. (2018a). Sparse Generalized Eigenvalue Problem: Optimal Statistical Rates via Truncated Rayleigh Flow, *Journal of the Royal Statistical Society Series B: Statistical Methodology* **80**(5): 1057–1086.

Tan, K. M., Wang, Z., Zhang, T., Liu, H. and Cook, R. D. (2018b). A convex formulation for high-dimensional sparse sliced inverse regression, *Biometrika* **105**(4): 769–782.

Tan, K., Shi, L. and Yu, Z. (2020). Sparse sir: Optimal rates and adaptive estimation, *The Annals of Statistics* **48**(1): 64–85.

Vu, V. Q. and Lei, J. (2013). Minimax sparse principal subspace estimation in high dimensions, *Annals of Statistics* **41**: 2905–2947.

Witten, D. M., Tibshirani, R. and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis, *Biostatistics* **10**(3): 515–534.

Yin, X. and Cook, R. D. (2003). Estimating central subspaces via inverse third moments, *Biometrika* **90**: 113–125.

Yuan, X.-T. and Zhang, T. (2013). Truncated power method for sparse eigenvalue problems, *Journal of Machine Learning Research* **14**(Apr): 899–925.

Zhang, X., Lee, C. E. and Shao, X. (2020). Envelopes in multivariate regression models with nonlinearity and heteroscedasticity, *Biometrika* **107**: 965–981.

Zhou, J. and He, X. (2008). Dimension reduction based on constrained canonical correlation and variable filtering, *The Annals of Statistics* **36**: 1649–1668.

Zhu, L. and Ng, K. W. (1995). Asymptotics of sliced inverse regression, *Statist. Sinica* **5**: 727–736.

Zhu, L.-P., Zhu, L.-X. and Feng, Z.-H. (2010). Dimension reduction in regressions through cumulative slicing estimation, *Journal of the American Statistical Association* **105**(492): 1455–1466.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0112

Zou, H., Hastie, T. and Tibshirani, R. (2006). Sparse principal component analysis, *Journal of computational and graphical statistics* **15**(2): 265–286.

Qing Mai, Florida State University
E-mail: qmai@fsu.edu
Xiaofeng Shao, University of Illinois at Urbana-Champaign
E-mail: xshao@illinois.edu
Runmin Wang, Texas A&M University
E-mail: runminw@tamu.edu
Xin Zhang, Florida State University
E-mail: xzhang8@fsu.edu