Statistica Sinica **32** (2022), 2297-2314
doi:https://doi.org/10.5705/ss.202022.0169

A REVIEW ON SLICED INVERSE REGRESSION, SUFFICIENT DIMENSION REDUCTION, AND APPLICATIONS

Ming-Yueh Huang and Hung Hung

Academia Sinica and National Taiwan University

Abstract: For effective dimension reduction (e.d.r.) in regression, the sliced inverse regression (SIR) is used to detect detailed structures of conditional distributions and reduce the dimensionality of covariates in a nonparametric manner. Subsequent analysis can then be based on features of a lower dimension, which assists with model interpretation and increases the estimation efficiency. The concept of e.d.r. has led to the framework of sufficient dimension reduction (SDR), with promising developments in various fields. Here, we first review the SIR and other estimation methods for SDR when a complete random sample with finite-dimensional covariates is available. Then, we discuss extensions and applications to cases with more complicated structures, including high-dimensional data and two types of incomplete data. Lastly, we emphasize the importance of SDR in modern statistical applications, and explain how conventional SDR methods need to adapt to different data structures in order to ensure good performance.

Key words and phrases: Dimension reduction, high-dimensional data, semiparametric statistics.

1. Introduction

Advancements in science and technology are resulting in data increasing in terms of both size and complexity. One characteristic of such complexity is the abundance of available covariates, which complicates the dependence relationship between the response variable and the covariates. To deal with multivariate or high-dimensional covariates, conventional statistical approaches use parametric models that simplify the task of estimation and inference. However, these approaches suffer from model misspecification, requiring a model diagnosis for each of the fitted models. Although nonparametric smoothing can directly estimate the conditional mean, or even the conditional distribution functions, without specifying parametric model assumptions, the large number of covariates leads to unstable estimates, owing to the curse of dimensionality.

Corresponding author: Ming-Yueh Huang, Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan. E-mail: myh0728@stat.sinica.edu.tw.

Influenced by the pioneering work of Li (1991) on effective dimension reduction (e.d.r.) in regression, the framework of sufficient dimension reduction (SDR, Cook (1998)) has yielded many powerful statistical methods for regression analysis and supervised learning. These methods can detect detailed structures of conditional distributions and reduce the dimension of covariates in a fully nonparametric manner. Let $Y \in \mathbb{R}$ be a response of interest and $\mathbf{X} \in \mathbb{R}^p$ be a vector of covariates. SDR aims to find d (with $0 \le d \le p$) linear indices $\beta_1^{\mathrm{T}}\mathbf{X}, \ldots, \beta_d^{\mathrm{T}}\mathbf{X}$ such that

$$Y \perp\!\!\!\perp \mathbf{X} \,|\, \mathbf{B}^{\mathrm{T}}\mathbf{X}, \tag{1.1}$$

where $\mathbf{B} = (\beta_1, \ldots, \beta_d)$ is a $p \times d$ parameter matrix (Cook (1998)). Note that (1.1) is equivalent to

$$\mathrm{P}(Y \le y \,|\, \mathbf{X} = \mathbf{x}) = \mathrm{P}(Y \le y \,|\, \mathbf{B}^{\mathrm{T}}\mathbf{X} = \mathbf{B}^{\mathrm{T}}\mathbf{x}), \tag{1.2}$$

for all $y \in \mathbb{R}$ (Zeng and Zhu (2010)). That is, conditional on $\mathbf{X}$, the distribution of Y is the same as that conditional on $\mathbf{B}^{\mathrm{T}}\mathbf{X}$. Thus, we can make an inference about Y given $\mathbf{X}$ based on $\mathbf{B}^{\mathrm{T}}\mathbf{X}$, without loss of information. If there exists $d < p$ such that (1.1) holds, dimension reduction is achieved using the non-trivial linear indices $\beta_1^{\mathrm{T}}\mathbf{X}, \ldots, \beta_d^{\mathrm{T}}\mathbf{X}$.

To identify the estimable parameters in SDR, note that (1.1) implies that $Y \perp\!\!\!\perp \mathbf{X} \,|\, (\mathbf{BA})^{\mathrm{T}}\mathbf{X}$ for all nonsingular $d \times d$ matrices $\mathbf{A}$. Because the column space of a matrix is invariant under nonsingular transformations, it suffices to identify the column space of B, which is called an SDR subspace. To reduce the dimension by as much as possible, practitioners are interested in the SDR subspace with the smallest dimension. Under mild conditions (Cook (1994, 1996, 1998)), the intersection of all SDR subspaces is still an SDR subspace. This unique, smallest SDR subspace is called the central subspace, and is denoted by $\mathcal{S}_{Y|\mathbf{X}}$. Note that the definition of a central subspace is intrinsically nonparametric. The conditional distribution function in (1.2) is completely fluent, and there is no need to specify particular parametric models. Therefore, we not require a parametric model diagnosis to validate validate (1.1), and SDR becomes particularly useful in exploratory data analysis.

Another useful observation is that if we set $\mathbf{Z} = \mathbf{A}^{-1}(\mathbf{X}-\mathbf{c})$ and $\mathbf{z} = \mathbf{A}^{-1}(\mathbf{x}-\mathbf{c})$, where $\mathbf{A}$ is a $p \times p$ nonsingular symmetric constant matrix, and $\mathbf{c} \in \mathbb{R}^p$ is a constant vector, then $\mathrm{P}(Y \le y \,|\, \mathbf{B}^{\mathrm{T}}\mathbf{X} = \mathbf{B}^{\mathrm{T}}\mathbf{x}) = \mathrm{P}\{Y \le y \,|\, (\mathbf{AB})^{\mathrm{T}}\mathbf{Z} = (\mathbf{AB})^{\mathrm{T}}\mathbf{z}\}$. It follows immediately that $\mathcal{S}_{Y|\mathbf{X}} = \mathbf{A}^{-1}\mathcal{S}_{Y|\mathbf{Z}}$. This property ensures that centralizing and standardizing the covariates does not change the nature of the SDR problem. Thus, to simplify the presentation, we assume that $\mathrm{E}(\mathbf{X}) = \mathbf{0}$

and $\mathrm{var}(\mathbf{X}) = \mathbf{I}_p$ in Sections 2–3. However, we relax these conditions in Section 4 for further applications.

In the next section, we discuss the sliced inverse regression (SIR) and other approaches for estimating $\mathcal{S}_{Y|\mathbf{X}}$. In Section 3, we review how SDR studies have been extended to address high-dimensional and infinite-dimensional covariates. In some applications, the response variables may be incomplete, in which case no complete sample of $(Y, \mathbf{X})$ is available. Thus, in Section 4, we review SDR methods for two types of incomplete data, right-censoring and counterfactual causal modeling.

2. Approaches for Estimation

2.1. SIR and its extensions

Let $\{(Y_i, \mathbf{X}_i) : i = 1, \ldots, n\}$ be a random sample of $(Y, \mathbf{X})$. To estimate $\mathcal{S}_{Y|\mathbf{X}}$, we first estimate the dimension of the central subspace, denoted by d_0, and then estimate a $p \times d_0$ basis matrix, denoted by $\mathbf{B}_0$. In this subsection, we focus on inverse regression methods, which reverse the relation between the response and the covariates. Because the response variable of interest is univariate, we can estimate the conditional moments and distributions of $\mathbf{X}$ given $Y = y$ using standard nonparametric methods that are less affected by the curse of dimensionality. The most widely used method in this group is the SIR of Li (1991), who shows that

$$\mathrm{E}(\mathbf{X} \,|\, Y) \in \mathcal{S}_{Y|\mathbf{X}}$$

under a linearity condition that $\mathrm{E}(\mathbf{X} \,|\, \mathbf{B}^{\mathrm{T}}\mathbf{X})$ is linear in $\mathbf{B}^{\mathrm{T}}\mathbf{X}$, and applies a principal component analysis on the random vector $\mathrm{E}(\mathbf{X} \,|\, Y)$ to recover $\mathcal{S}_{Y|\mathbf{X}}$. Equivalently, we can derive a basis of $\mathcal{S}_{Y|\mathbf{X}}$ from the solution of

$$\underset{\Gamma^{\mathrm{T}}\Gamma = \mathbf{I}_{d_0}}{\operatorname{argmax}} \operatorname{tr}(\Gamma^{\mathrm{T}} \mathbf{K}_{\mathrm{SIR}} \Gamma), \tag{2.1}$$

where $\mathbf{K}_{\mathrm{SIR}}$ is the SIR kernel matrix $\mathrm{var}\{E(\mathbf{X} \,|\, Y)\}$ and Γ is a $p \times d_0$ matrix. In practice, the solution is formed by the d_0 leading eigenvectors of $\mathbf{K}_{\mathrm{SIR}}$.

Observing that an SIR may fail when $\mathrm{E}(\mathbf{X} \,|\, Y) \equiv \mathbf{0}$ for symmetrically distributed covariates, Cook and Weisberg proposed the sliced average variance estimation (SAVE), which uses the second-moment-based kernel matrix $\mathbf{K}_{\mathrm{SAVE}} = \mathrm{E}[\{\mathbf{I}_p - \mathrm{var}(\mathbf{X} \,|\, Y)\}^2]$ to replace $\mathbf{K}_{\mathrm{SIR}}$ in (2.1). In addition, Li and Wang (2007) proposed the directional regression (DR) by using the kernel matrix $\mathbf{K}_{\mathrm{DR}} = \mathrm{E}([2\mathbf{I}_p - \mathrm{E}\{(\mathbf{X} - \tilde{\mathbf{X}})^{\otimes 2} \,|\, Y, \tilde{Y}\}]^2)$, where $(\tilde{\mathbf{X}}, \tilde{Y})$ is an independent copy of $(\mathbf{X}, Y)$. Under an additional constant variance assumption, it is shown that the column

spaces of $\mathbf{K}_{\text{SAVE}}$ and $\mathbf{K}_{\text{DR}}$ are contained in $\mathcal{S}_{Y|\mathbf{X}}$. Thus, SAVE and DR can successfully recover $\mathcal{S}_{Y|\mathbf{X}}$.

In practice, these methods require estimating the conditional moments of $\mathbf{X}$ given $\mathbf{Y}$, and usually involve a slice/bandwidth selection problem. To avoid using these tuning parameters and to improve the estimation accuracy, Zhu et al. (2010) introduced the discretization–expectation method and Zhu, Zhu and Feng (2010) proposed a class of cumulative slicing estimations similar to the SIR, SAVE, and DR. To determine d_0, Li (1991) showed that the nonzero eigenvalues of the estimated kernel matrix for $\mathbf{K}_{\text{SIR}}$ follow a χ^2 distribution, and hence proposed a sequential χ^2 test. Overall, the inverse regression-type methods are easily implemented and have solid theoretical theory. In addition, Chen and Li (1998) showed that the standard errors are available for the estimated central subspace, and provide graphical and diagnostic tools to enhance the analysis.

Several extensions of the SIR have been proposed to address different statistical problems. For measurement error regression, Carroll and Li (1992) showed that if one proceeds with the usual analysis, ignoring measurement error, then the SIR yields estimates that consistently estimate the true regression parameter, up to a constant of proportionality. For discriminant analysis, Chen and Li (2001) connected the SIR to Fisher's canonical variates, showing several ways of generalizing Fisher's linear discriminant analysis for better exploration and exploitation of nonlinear data patterns. Lastly, Li et al. (2003) and Lue (2009) extended the SIR for multivariate response regression.

2.2. Gradient-based methods

The inverse regression methods are elegant and relatively simple to implement, but are restricted by additional assumptions on the covariates, such as the linearity condition and the constant variance assumption. A simple way to relax these assumptions is to use the higher-order derivatives of the regression function. Let $H(\mathbf{x})$ be the Hessian matrix of the regression function $m(\mathbf{x}) = \mathrm{E}(Y \mid \mathbf{X} = \mathbf{x})$ at $\mathbf{x}$, and define the principal Hessian directions (pHd's) $\beta_1, \ldots, \beta_p$ as the eigenvectors of the matrix $H(\mathbf{x})\mathrm{var}(\mathbf{X})$. Li (1992) showed that the pHd's with nonzero eigenvalues are in the central subspace $\mathcal{S}_{Y|\mathbf{X}}$. Under a normal distribution assumption on $\mathbf{X}$, the average Hessian matrix $\mathrm{E}\{H(\mathbf{X})\}$ is equivalent to the weighted covariance $\mathrm{E}[\{Y - \mathrm{E}(Y)\}\{\mathbf{X} - \mathrm{E}(\mathbf{X})\}\{\mathbf{X} - \mathrm{E}(\mathbf{X})\}^T]$ from Stein's lemma. Suppose the weighted covariance is estimated by $\widehat{\mathbf{K}}_{\text{pHd}}$. Then, the eigenvectors of $\widehat{\mathbf{K}}_{\text{pHd}}$ with nonzero eigenvalues can be used to recover $\mathcal{S}_{Y|\mathbf{X}}$. Li, Lue and Chen (2000) provide an extension to tree-structured regression.

Instead of imposing normal distribution assumptions, Xia et al. (2002) proposed the minimum average variance estimation (MAVE) by observing that (1.2) implies $m(\mathbf{x}) = \mathrm{E}(Y \,|\, \tilde{\mathbf{B}}_0^T\mathbf{X} = \tilde{\mathbf{B}}_0^T\mathbf{x})$, for some $\mathrm{span}(\tilde{\mathbf{B}}_0) \subset \mathrm{span}(\mathbf{B}_0)$, and

$$\frac{\mathrm{dE}(Y \,|\, \mathbf{X} = \mathbf{x})}{\mathrm{d}\mathbf{x}} = \tilde{\mathbf{B}}_0^{\mathrm{T}} \frac{\partial \mathrm{E}(Y \,|\, \tilde{\mathbf{B}}_0^T\mathbf{X} = \tilde{\mathbf{B}}_0^T\mathbf{x})}{\partial(\tilde{\mathbf{B}}_0^T\mathbf{x})} \in \mathrm{span}(\tilde{\mathbf{B}}_0),$$

for all $\mathbf{x} \in \mathbb{R}^p$. In fact, $\mathrm{span}(\tilde{\mathbf{B}}_0)$ is the smallest linear subspace that satisfies $Y \perp\!\!\!\perp \mathrm{E}(Y \,|\, \mathbf{X}) \,|\, \tilde{\mathbf{B}}_0^{\mathrm{T}}\mathbf{X}$, and is called the central mean subspace and is denoted by $\mathcal{S}_{\mathrm{E}(Y \,|\, \mathbf{X})}$. By using a local linear regression to estimate the gradients $\mathrm{dE}(Y \,|\, \mathbf{X} = \mathbf{x})/\mathrm{d}\mathbf{x}$, the MAVE estimates $\tilde{\mathbf{B}}_0$ by solving

$$\min_{\mathbf{a},\mathbf{b},\tilde{\mathbf{B}}} \sum_{i=1}^{n}\sum_{j=1}^{n} \{Y_i - a_j - \mathbf{b}_j^{\mathrm{T}}\tilde{\mathbf{B}}^{\mathrm{T}}(\mathbf{X}_i - \mathbf{X}_j)\}^2 K_h(\tilde{\mathbf{B}}^{\mathrm{T}}\mathbf{X}_i - \tilde{\mathbf{B}}^{\mathrm{T}}\mathbf{X}_j),$$

where $\mathbf{a} = (a_1, \ldots, a_n)$, $\mathbf{b} = (\mathbf{b}_1, \ldots, \mathbf{b}_n)$, $K_h(\cdot) = K(\cdot/h)/h$, K is a kernel function, and h is a bandwidth. To recover $\mathcal{S}_{Y|\mathbf{X}}$, Zhu and Zeng (2006) showed that $\mathcal{S}_{Y|\mathbf{X}} = \sum_g \mathcal{S}_{\mathrm{E}\{g(Y) \,|\, \mathbf{X}\}}$, and used Fourier transforms to aggregate sufficient central mean subspaces. Xia (2007) adapted the idea of the MAVE by using $K_{h^*}(Y - y)$ as an induced response and proposed the density based MAVE (dMAVE) by aggregating the MAVE criteria over different values of y. To further avoid the smoother in the induced response, Wang and Xia (2008) used $I(Y \le y)$ as another induced response, proposing the sliced regression (SR). These estimators are further unified by an ensemble of MAVEs introduced by Yin and Li (2011).

To estimate d_0 using these nonparametric gradient-based methods, researchers often use information criterion-based methods (Zhu, Miao and Peng (2006); Ma and Zhang (2015)), which view SDR as a problem of selecting a model from a series of nested semiparametric models. In a similar spirit, Wang and Xia (2008) used the cross-validation for the unknown link function to select d_0. In contrast to inverse regression-type methods, gradient-based methods require neither the linearity condition nor the constant variance assumption. However, they usually involve at least d_0-dimensional nonparametric smoothing estimators, and are not straightforward to implement in practice.

2.3. Semiparametric methods

The semiparametric methods view (1.2) as a semiparametric distribution regression model under a given d_0. The parameter matrix $\mathbf{B}_0$ is the finite-dimensional major parameter, and the link function $F(\mathbf{u}) = \mathrm{P}(Y \le y \,|\, \mathbf{B}_0^{\mathrm{T}}\mathbf{X} = \mathbf{u})$

is an infinite-dimensional nuisance parameter. Applying the geometric tools of Bickel et al. (1998) and Tsiatis (2006), Ma and Zhu (2012) derived the complete family of influence functions, and subsequently obtained the complete class of regular $n^{1/2}$-consistent estimators (van der Vaart (1998)). More precisely, the class of unnormalized influence functions is

$$\{\mathbf{g}(Y,\mathbf{X}) - \mathrm{E}\{\mathbf{g}(Y,\mathbf{X}) \,|\, Y, \mathbf{B}_0^{\mathrm{T}}\mathbf{X}\} : \mathrm{E}\{\mathbf{g}(Y,\mathbf{X}) \,|\, \mathbf{X}\} = \mathrm{E}\{\mathbf{g}(Y,\mathbf{X}) \,|\, \mathbf{B}_0^{\mathrm{T}}\mathbf{X}\}\}. \quad (2.2)$$

The form of these influence functions offers many ways to construct consistent estimating equations. For example, for any functions $\mathbf{f}(y,\mathbf{u})$ and $\alpha(\mathbf{x})$, the function

$$\mathbf{g}(Y,\mathbf{X}) = [\mathbf{f}(Y,\mathbf{B}_0^T\mathbf{X}) - \mathrm{E}\{\mathbf{f}(Y,\mathbf{B}_0^T\mathbf{X}) \,|\, \mathbf{B}_0^T\mathbf{X}\}][\alpha(\mathbf{X}) - \mathrm{E}\{\alpha(\mathbf{X}) \,|\, \mathbf{B}_0^T\mathbf{X}\}]$$

satisfies $\mathrm{E}\{\mathbf{g}(Y,\mathbf{X}) \,|\, \mathbf{B}_0^T\mathbf{X}\} = 0$. Thus, a regular estimator can be obtained by solving the sample version of the estimating equation

$$\sum_{i=1}^{n}[\mathbf{f}(Y_i,\mathbf{B}_0^T\mathbf{X}_i) - \mathrm{E}\{\mathbf{f}(Y_i,\mathbf{B}_0^T\mathbf{X}_i) \,|\, \mathbf{B}_0^T\mathbf{X}_i\}][\alpha(\mathbf{X}_i) - \mathrm{E}\{\alpha(\mathbf{X}_i) \,|\, \mathbf{B}_0^T\mathbf{X}_i\}] = 0. \quad (2.3)$$

The estimating equation (2.3) has a double-robustness property, allowing a misspecification of either $\mathrm{E}\{\mathbf{f}(Y_i,\mathbf{B}_0^T\mathbf{X}_i) \,|\, \mathbf{B}_0^T\mathbf{X}_i\}$ or $\mathrm{E}\{\alpha(\mathbf{X}_i) \,|\, \mathbf{B}_0^T\mathbf{X}_i\}$. More precisely, for any function $\mathbf{h}(\mathbf{B}_0^T\mathbf{x})$, the estimation equations

$$\sum_{i=1}^{n}[\mathbf{f}(Y_i,\mathbf{B}_0^T\mathbf{X}_i) - \mathbf{h}(\mathbf{B}_0^T\mathbf{X}_i)][\alpha(\mathbf{X}_i) - \mathrm{E}\{\alpha(\mathbf{X}_i) \,|\, \mathbf{B}_0^T\mathbf{X}_i\}] = 0 \quad (2.4)$$

and

$$\sum_{i=1}^{n}[\mathbf{f}(Y_i,\mathbf{B}_0^T\mathbf{X}_i) - \mathrm{E}\{\mathbf{f}(Y_i,\mathbf{B}_0^T\mathbf{X}_i) \,|\, \mathbf{B}_0^T\mathbf{X}_i\}][\alpha(\mathbf{X}_i) - \mathbf{h}(\mathbf{B}_0^T\mathbf{X}_i)] = 0 \quad (2.5)$$

both yield consistent estimators. From (2.4)–(2.5), several popular existing SDR methods can be shown to be special cases of the semiparametric estimation family. For example, to obtain the SIR as a semiparametric estimator, we choose $\mathbf{f}(y,\mathbf{B}_0^T\mathbf{x}) = \mathrm{E}(\mathbf{X} \,|\, Y = y)$ and $\alpha(\mathbf{x}) = \mathbf{x}^T$. The linearity condition promises a parametric form $\mathrm{E}\{\alpha(\mathbf{X}) \,|\, \mathbf{B}_0^T\mathbf{X}\} = \mathbf{X}^T\mathbf{P}$, for some parameter matrix $\mathbf{P}$. By further choosing $\mathbf{h}(\mathbf{B}_0^T\mathbf{x}) = 0$, the population version of (2.4) yields $\mathrm{E}\{\mathrm{E}(\mathbf{X} \,|\, Y)\mathbf{X}^T\}(I_p - \mathbf{P}) = 0$, or equivalently, $\mathbf{K}_{\mathrm{SIR}}\mathrm{var}(\mathbf{X} \,|\, \mathbf{B}_0^T\mathbf{X}) = 0$. The solution coincides with the maximizer of (2.1).

From the form of the influence functions in (2.2), Ma and Zhu (2013) derived the efficient score from (2.2), given by

$$\mathbf{S}_{\text{eff}}(Y, \mathbf{X}; \mathbf{B}_0) = \text{vec}\left[\{\mathbf{X} - \text{E}(\mathbf{X} \mid \mathbf{B}_0^{\text{T}}\mathbf{X})\}\frac{\partial \log f(Y \mid \mathbf{B}_0^{\text{T}}\mathbf{X})}{\partial(\mathbf{B}_0^{\text{T}}\mathbf{X})}\right],$$

where $f(y \mid \mathbf{u})$ is the conditional density of Y given $\mathbf{B}_0^{\text{T}}\mathbf{X} = \mathbf{u}$. The corresponding semiparametric efficient estimator for $\mathbf{B}_0$ can be obtained by solving the estimating equation

$$\sum_{i=1}^{n} \mathbf{S}_{\text{eff}}(Y_i, \mathbf{X}_i; \mathbf{B}) = 0,$$

with respect to $\mathbf{B}$. That is, the resulting estimator has the smallest asymptotic variance-covariance matrix among all regular $n^{1/2}$-consistent estimators.

3. Extensions to High-Dimensional Data

3.1. Sufficient variable selection

Variable selection is a common procedure when analyzing high-dimensional data sets that screens out irrelevant variables, allowing the subsequent analysis to be based on a small subset of X. Variable selection shares the spirit of SDR, but specifically targets on the subset of X that preserves the information about the association between (X, Y). This motivates the idea of sufficient variable selection (SVS), which aims to identify a matrix $\mathbf{V}$ such that

$$Y|X \overset{d}{=} Y|\mathbf{V}^{\text{T}}X, \tag{3.1}$$

where $\mathbf{V}^{\text{T}} = [\mathbf{I}_r, \mathbf{0}_{r\times(p-r)}]\Pi$, for some column permutation matrix Π when multiplying on the right side of a matrix. The intersection of the span of all such $\mathbf{V}$, denoted by $\mathcal{S}^{\mathbf{V}}_{Y|\mathbf{X}}$, is called the central variable selection space, and is the target of SVS. It is obvious that $\mathcal{S}_{Y|\mathbf{X}} \subseteq \mathcal{S}^{\mathbf{V}}_{Y|\mathbf{X}}$, owing to the specific structure of $\mathbf{V}$ required in SVS. Moreover, it can be shown that $\mathcal{S}^{\mathbf{V}}_{Y|\mathbf{X}}$ exists and is unique if $\mathcal{S}_{Y|\mathbf{X}}$ exists.

The close connection between $\mathcal{S}_{Y|\mathbf{X}}$ and $\mathcal{S}^{\mathbf{V}}_{Y|\mathbf{X}}$ suggests that SDR methods do provide information about regarding $\mathcal{S}^{\mathbf{V}}_{Y|\mathbf{X}}$. In particular, extensions of SDR methods to achieve SVS are generally based on the fact that

$$\mathbf{e}_j \notin \mathcal{S}^{\mathbf{V}}_{Y|\mathbf{X}} \Longleftrightarrow \mathbf{V}_{(j)} = 0, \quad j = 1, \ldots, p, \tag{3.2}$$

where $\mathbf{e}_j$ has a one in the jth element, and zero elsewhere, $\mathbf{V}_{(j)}$ is the jth row of

$\mathbf{V}$, and $\mathbf{V}$ is a basis of $\mathcal{S}^{\mathbf{V}}_{Y|\mathbf{X}}$. This implies that we can use an estimate of $\mathcal{S}_{Y|\mathbf{X}}$ with row-sparsity to estimate $\mathcal{S}^{\mathbf{V}}_{Y|\mathbf{X}}$. To implement this idea, Chen, Zou and Cook (2010) proposed the coordinate-independent sparse sufficient dimension reduction estimator (CISE). The population version of the CISE is defined as the solution of

$$\max_{\mathbf{V}:\mathbf{V}^{\mathrm{T}}\Sigma\mathbf{V}=I} \operatorname{tr}(\mathbf{V}^{\mathrm{T}}\mathbf{M}\mathbf{V}) - \sum_{j=1}^{p} \lambda_j \|\mathbf{V}_{(j)}\|_2, \tag{3.3}$$

where $\mathbf{M}$ is a method-specific kernel matrix satisfying $\operatorname{span}(\mathbf{M}) = \Sigma\mathcal{S}_{Y|\mathbf{X}}$, and $\lambda_j > 0$ is a penalty parameter. Note that the maximization problem $\max_{\mathbf{V}:\mathbf{V}^{\mathrm{T}}\Sigma\mathbf{V}=I} \operatorname{tr}(\mathbf{V}^{\mathrm{T}}\mathbf{M}\mathbf{V})$ is equivalent to the generalized eigenvalue problem commonly used in SDR methods. The group-lasso-type penalty $\|\mathbf{V}_{(j)}\|_2$ forces the solution of (3.3) to yield $\mathbf{V}_{(j)} = 0$, for some j. Let $\widehat{\mathbf{V}}$ be the sample version of the solution of (3.3). Then, we can estimate $\mathcal{S}^{\mathbf{V}}_{Y|\mathbf{X}}$ by

$$\{\mathbf{e}_j : \widehat{\mathbf{V}}_{(j)} \neq 0, j = 1, \ldots, p\}. \tag{3.4}$$

In contrast to the CISE, which uses a penalized estimation criterion, Yu, Dong and Shao (2016) extend the SDR methods to rank the importance of variables, and thus achieve SVS. For a given method-specific kernel matrix $\mathbf{M}$, let $s_j = \mathbf{e}_j^{\mathrm{T}}\Sigma^{-1}\mathbf{M}\Sigma^{-1}\mathbf{e}_j$, for $j = 1, \ldots, p$. Yu, Dong and Shao (2016) show that

$$\mathbf{e}_j \in \mathcal{S}^{\mathbf{V}}_{Y|\mathbf{X}} \Longleftrightarrow s_j > 0, \quad j = 1, \ldots, p. \tag{3.5}$$

Thus, s_j provides information about the importance of X_j in the association between (Y, X). Let $\widehat{s}_j$ be the sample version of s_j. The authors propose estimating $\mathcal{S}^{\mathbf{V}}_{Y|\mathbf{X}}$ by

$$\{\mathbf{e}_j : \widehat{s}_j > \delta, j = 1, \ldots, p\}, \tag{3.6}$$

where δ is a predetermined positive critical value.

We conclude this section by explaining the difference between SVS and conventional variable selection. In general, variable selection methods (such as the sure independence screening (SIS) of Fan and Lv (2008) or the distance correlation (DC) of Székely, Rizzo and Bakirov (2007)) can only guarantee *screening consistency*, that is, a relevant variable is identified with probability tending to one, but irrelevant variables may also be falsely identified with positive probability. On the other hand, owing to the definition of $\mathcal{S}^{\mathbf{V}}_{Y|\mathbf{X}}$, SVS can achieve *selection consistency*, that is, relevant variables and only relevant variables are identified with probability tending to one, provided that $\operatorname{span}(\mathbf{M}) = \Sigma\mathcal{S}_{Y|\mathbf{X}}$. Thus, with the assistance of SDR, one can expect SVS to be more accurate than conven-

tional variable selection methods, especially in the sense of excluding irrelevant variables.

3.2. SDR under the high-dimensional setting

High-dimensional data sets with $p \gg n$ are frequently encountered in modern statistical applications. While there appears to be more regression information available to explore the increased complexity of differentiating the information from the noisy regressors, such data sets have impleded finding answers to scientific questions. Although SDR reduces the dimension of the covariates without losing information, most SDR methods cannot be applied directly when $p \gg n$. For example, inverse regression-based SDR methods can be formulated as a generalized eigenvalue problem:

$$\mathbf{K}\beta_j = \lambda_j \beta_j \quad \text{with} \quad \mathbf{K} = \Sigma^{-1}\mathbf{M}, \quad j = 1 \ldots, p, \tag{3.7}$$

where Σ is the $p \times p$ covariance matrix of X, and M is a $p \times p$ method-specific symmetric kernel matrix satisfying the property $\text{span}(\mathbf{M}) = \Sigma\mathcal{S}_{Y|\mathbf{X}}$. It can be shown that the leading d eigenvectors $\beta = [\beta_1, \ldots, \beta_d]$ provide an estimate of a basis of $\mathcal{S}_{Y|\mathbf{X}}$. For the choice of $\mathbf{M}$, SIR uses $\mathbf{M}_{\text{SIR}} = \text{cov}\{E(\mathbf{X}|Y)\}$, SAVE uses $\mathbf{M}_{\text{SAVE}} = \Sigma^{1/2}E[\{\mathbf{I} - \Sigma^{-1/2}\text{cov}(\mathbf{X}|Y)\Sigma^{-1/2}\}^2]\Sigma^{1/2}$, and DR uses $\mathbf{M}_{\text{DR}} = \Sigma^{1/2}\text{E}[\{2\mathbf{I} - \Sigma^{-1/2}\text{cov}(\mathbf{X} - \mathbf{X}'|Y, Y')\Sigma^{-1/2}\}^2]\Sigma^{1/2}$, with $(\mathbf{X}', Y')$ a random copy of $(\mathbf{X}, Y)$. Because (3.7) involves the inverse of Σ and the sample covariance matrix $\widehat{\mathbf{X}}$ is singular for $p \geq n$, conventional SDR methods cannot be applied directly to estimate $\mathcal{S}_{Y|\mathbf{X}}$.

One group of high-dimensional SDR methods assumes knowledge of an orthonormal $\mathcal{E}_{p\times r}$ with $r \ll p$, such that $\mathbf{B} = \mathcal{E}\Gamma$, for some $\Gamma_{r\times d}$, with $d < r$. That is, $\text{span}(\mathbf{B}) \subseteq \text{span}(\mathcal{E})$ and, hence, $\mathcal{S}_{Y|\mathbf{X}} = \mathcal{E}S_{Y|\mathcal{E}^{\text{T}}\mathbf{X}}$. To estimate $\mathcal{S}_{Y|\mathbf{X}}$, it suffices to estimate Γ based on $(Y, \mathcal{E}^{\text{T}}X)$ which bypasses the large-p-small-n problem. For example, a common strategy is to reduce the dimension of X using PCA, which implicitly assumes that $\mathcal{E}$ is a basis of the span of the leading eigenvectors of Σ. Note that the PCA-based $\mathcal{E}$ depends on the information of $\mathbf{X}$ only. To use both the information of $(\mathbf{X}, Y)$, Li, Cook and Tsai (2007) borrow the idea of a partial least square to modify the SIR, proposing the partial inverse regression (PIR), which estimates $\mathcal{S}_{Y|\mathbf{X}}$ by

$$\widehat{\mathbf{B}}_{\text{PIR}} = \mathcal{E}\widehat{\Gamma}_{\text{SIR}}, \tag{3.8}$$

where $\widehat{\Gamma}_{\text{SIR}}$ is the SIR estimator of $S_{Y|\mathcal{E}^{\text{T}}\mathbf{X}}$ based on $(Y, \mathcal{E}^{\text{T}}\mathbf{X})$, $\mathcal{E}$ is chosen as the Krylov sequence $[\nu, \Sigma\nu, \ldots, \Sigma^{q-1}\nu]$, for some $q > 0$, and ν is the eigenvector

of $\mathbf{M}_{\text{SIR}}$. Later, Cook, Li and Chiaromonte (2007) extended the PIR to general SDR methods, also investigating how to determine q theoretically.

In contrast to the above-mentioned methods, in which $\mathcal{E}$ is estimated from the data, another group of methods overcome the large-p-small-n problem by implementing conventional SDR methods on subsets of $\mathbf{X}$ (with size r), that is, $\mathcal{E}^{\mathrm{T}} = [\mathbf{I}_r, \mathbf{0}_{r\times(p-r)}]\Pi$, for some column permutation matrix Π when multiplying on the right side of a matrix. Yin and Hilafu (2015) propose the sequential SDR (seq-SDR). Let $\mathbf{X}$ be partitioned into two disjoint sets $\mathbf{X}_{(1)}$ and $\mathbf{X}_{(2)}$. The main idea of the seq-SDR is based on the following two statements:

(i) $\mathbf{X}_{(1)} \perp (\mathbf{X}_{(2)}, Y)|\mathbf{B}_{(1)}^{\mathrm{T}}\mathbf{X}_{(1)}$

(ii) $\mathbf{X}_{(1)} \perp \mathbf{X}_{(2)}|(\mathbf{B}_{(1)}^{\mathrm{T}}\mathbf{X}_{(1)}, Y)$ and $\mathbf{X}_{(1)} \perp Y|\mathbf{B}_{(1)}^{\mathrm{T}}\mathbf{X}_{(1)}$,

which imply that

$$\mathbf{X}_{(1)} \perp Y|(\mathbf{B}_{(1)}^{\mathrm{T}}\mathbf{X}_{(1)}, \mathbf{X}_{(2)}). \tag{3.9}$$

Once we have (3.9), we can replace $\mathbf{X}$ with the lower-dimensional $(\mathbf{B}_{(1)}^{\mathrm{T}}\mathbf{X}_{(1)}, \mathbf{X}_{(2)})$ without losing information. Note that both (i) and (ii) involve the SDR problem with multivariate responses, and the projective resampling SDR method of Li, Wen and Zhu (2008) can be applied to estimate $\mathbf{B}_{(1)}$, provided that the dimension of $\mathbf{X}_{(1)}$ is smaller than n. The implementation procedure for the seq-SDR is as follows:

1. Divide $\mathbf{X}$ into $(\mathbf{X}_{(1)}, \mathbf{X}_{(2)})$, and implement conventional SDR methods to estimate $\mathbf{B}_{(1)}$ based on either (i) or (ii).

2. Replace $\mathbf{X}$ with $(\mathbf{B}_{(1)}^{\mathrm{T}}\mathbf{X}_{(1)}, \mathbf{X}_{(2)})$ and repeat step 1 until the dimension of $\mathbf{X}$ cannot be reduced further.

Hilafu and Yin (2017) proposed seq-PIR, which incorporates the idea of PIR into the seq-SDR in order to tackle the problem of highly correlated $\mathbf{X}$.

An important issue when implementing seq-SDR and seq-PIR is the choice of the partitioning of $\mathbf{X}$ into $(\mathbf{X}_{(1)}, \mathbf{X}_{(2)})$, which affects the analysis results. As an alternative, Hilafu (2017) proposes the random SIR (rSIR), which does not require that we partition $\mathbf{X}$. The implementation procedure for the rSIR is as follows:

1. For each randomly generated $\mathcal{E}_\ell$, for $\ell = 1, \cdots, N$, implement SIR based on $(Y, \mathcal{E}_\ell^T\mathbf{X})$ to obtain $\widetilde{\Gamma}_\ell$, and transform back to $\widetilde{\mathbf{B}}_\ell = \mathcal{E}_\ell\widetilde{\Gamma}_\ell$ as an estimate of $\mathcal{S}_{Y|\mathbf{X}}$.

2. The final estimate of $\mathcal{S}_{Y|\mathbf{X}}$ is $\widehat{\mathbf{B}}_{\text{rSIR}}$, with the jth element being

$$\widehat{\mathbf{B}}_{\text{rSIR},j} = \frac{1}{N_j}\sum_{j=1}^{N}\widehat{\mathbf{B}}_{\ell,j}, \quad j = 1\ldots, p, \tag{3.10}$$

where $\widehat{\mathbf{B}}_{\ell,j}$ is the j-th element of $\widehat{\mathbf{B}}_{\ell}$, and N_j is the number of times that $\mathbf{X}_j$ is selected by $\{\mathcal{E}_\ell\}_{\ell=1}^{N}$.

A similar idea is adopted by Hung and Huang (2019), who proposed the integrated random partition SDR (iRP-SDR):

1. Randomly divide $\mathbf{X}$ into disjoint subsets. Calculate the DC between each subset and Y, and identify those subsets with leading DC values to form $\mathbf{X}_{(\ell)}$ (which corresponds to a certain choice of $\mathcal{E}_\ell$ such that $\mathbf{X}_{(\ell)} = \mathcal{E}_\ell^{\mathrm{T}}\mathbf{X}$), with dimension $r < n$.

2. Based on $(Y, \mathbf{X}_{(\ell)})$, implement a conventional SDR method (e.g., SIR) to obtain $\widehat{\Gamma}_\ell$ and the associated eigenvalue matrix $\widehat{\Lambda}_\ell$, and transform back to $\widehat{\mathbf{B}}_\ell = \mathcal{E}_\ell\widehat{\Gamma}_\ell$ as an estimate of $\mathcal{S}_{Y|\mathbf{X}}$.

3. The final estimate of $\mathcal{S}_{Y|\mathbf{X}}$ is obtained as the leading eigenvectors of the integrated kernel matrix $\widehat{\mathbf{K}} = (1/N)\sum_{\ell=1}^{N}\widehat{\mathbf{B}}_\ell\widehat{\Lambda}_\ell\widehat{\mathbf{B}}_\ell^{\mathrm{T}}\widehat{\mathbf{S}}$.

Neither rSIR nor iRP-SDR depend on a subset partition, but at the cost of requiring greater computational time due to multiple replicates. In addition, the integration method used to form the final estimate of $\mathcal{S}_{Y|\mathbf{X}}$ is not unique, which may affect the performance of the method.

3.3. Functional data

When the covariates are collected following time or other continuous indices, they can be treated as functional data with an infinite dimension. To illustrate this idea, let $\mathcal{T}$ be a compact interval, and let the covariate X be a random function from the real separable Hilbert space $H = L^2(\mathcal{T})$, endowed with inner product $\langle f, g\rangle = \int_{\mathcal{T}} f(t)\overline{g}(t)dt$ and norm $\|f\| = \langle f, f\rangle^{1/2}$. The SDR extended for functional covariates assumes that

$$Y \perp\!\!\!\perp X \,|\, (\langle\beta_1, X\rangle, \ldots, \langle\beta_d, X\rangle), \tag{3.11}$$

where $\beta_1, \ldots, \beta_d \in H$ are linearly independent index functions. Because the dimension of $L^2(\mathcal{T})$ is infinite, the definition of a central subspace for the functional covariates requires additional assumptions, and SDR methods developed

for multivariate covariates cannot be applied directly to functional covariates. In particular, because $\langle \beta + g, X\rangle = \langle \beta, X\rangle$ for all $g \in \ker X = \{f \in H : \langle f, x\rangle = 0$ for all trajectories x of $X\}$, we can only identify the elements in the quotient space $H/\ker X$, instead of $\beta \in H$. Here, for simplicity, we assume that either $\ker X = 0$ or $\{\beta_1, \ldots, \beta_d\}$ is contained in the subspace spanned by all trajectories of X, to ensure that the SDR subspace is identifiable.

Ferré and Yao (2005) propose the functional SIR (fSIR), which extends the SDR for functional covariates, and discuss the identifiability of the SDR subspace. When X has mean zero and satisfies a linearity condition, they showed that $\mathrm{E}(X \,|\, Y)$ belongs to the subspace spanned by $\Gamma\beta_1, \ldots, \Gamma\beta_d$, where Γ is the covariance operator of X. Thus, the SDR subspace can be recovered using the eigenspace of $\Gamma^{-1}\Gamma_e$, where Γ_e is the covariance operator of $\mathrm{E}(X \,|\, Y)$. However, Γ^{-1} is, in general, unbounded. To overcome these difficulties, additional restrictions are imposed by Forzani and Cook (2007) and Ferré and Yao (2007). To allow for more general types of covariates, Hsing and Ren (2009) introduced another formulation of the functional inverse regression method when the trajectories of X are in a reproducing kernel Hilbert space. For these inverse-regression-type methods, Li and Hsing (2010) developed sequential χ^2 tests to determine the number d of linear projections in (1.1).

In many applications in which the functional covariates are measured only intermittently and sparsely, they become longitudinal covariates. In such cases, $X_i(t)$ can only be observed at $t = T_{i1}, \ldots, T_{iN_i}$ for the ith subject, and the collected data are $\{(Y_i, X_{i1}, \ldots, X_{iN_i}) : i = 1, \ldots, n\}$, where X_{ij} is the measurement of X_i at time T_{ij}. Jiang, Yu and Wang (2014) proposed an fSIR method by estimating $\mathrm{E}(X \,|\, Y)$ using two-dimensional kernel smoothing. More precisely, for given (t, y), consider the weighted sum of squares

$$\sum_{i=1}^{n}\sum_{j=1}^{N_i}\{X_{ij} - \alpha_0 - \alpha_1(t - T_{ij}) - \alpha_2(y - Y_i)\}^2 K_{h_1}(T_{ij} - t)K_{h_2}(Y_i - y), \quad (3.12)$$

where $K_h(u) = K(u/h)/h$ for generic h, K is a kernel weight function, and h_1 and h_2 are bandwidths. Then, $\mathrm{E}\{X(t) \,|\, Y = y\}$ is estimated by $\widehat{a}_0$, where $(\widehat{a}_0, \widehat{a}_1, \widehat{\alpha}_2)$ is the minimizer of (3.12). To obtain an estimator with a better convergence rate and to avoid selecting two different bandwidths, Yao, Lei and Wu (2015) showed that $\mathrm{E}(X \,|\, Y \le y)$ belongs to the subspace spanned by $\Gamma\beta_1, \ldots, \Gamma\beta_d$, for all y. Then, $\mathrm{E}(X \,|\, Y \le y)$ can be estimated by minimizing

$$\sum_{i=1}^{n}\sum_{j=1}^{N_i}\{X_{ij}I(Y_i \le y) - \alpha_0 - \alpha_1(t - T_{ij})\}^2 K_h(T_{ij} - t).$$

The validity of these inverse-regression-type approaches relies on a linearity condition of the covariates. This condition is satisfied when the distribution of the covariates is elliptically contoured. Although elliptically contoured distributions are more general than Gaussian processes (Cambanis, Huang and Simons (1981)), we still require methodologies that can deal with asymmetric distributions.

4. Applications to Incomplete Data

4.1. Survival analysis

In a survival analysis, the response variable $Y \in \mathbb{R}^+$ is the survival time, and is usually subject to right-censoring. Let C be a censoring time, $Z = \min(Y, C)$ be the last follow-up time, and $D = I(Y \le C)$ be the indicator of the non-censoring. In addition, $S_Z(t \,|\, \mathbf{x})$, $S_Y(t \,|\, \mathbf{x})$, and $S_C(t \,|\, \mathbf{x})$ denote the conditional survival functions of Z, Y, and C, respectively, given $\mathbf{X} = \mathbf{x}$. Under the independent censoring assumption $Y \perp\!\!\!\perp C \,|\, \mathbf{X}$, it can be shown that $S_Z(t \,|\, \mathbf{x}) = S_Y(t \,|\, \mathbf{x})S_C(t \,|\, \mathbf{x})$ and $\mathrm{P}(D = 1 \,|\, \mathbf{X} = \mathbf{x}) = \int_0^\infty S_C(t^- \,|\, \mathbf{x})\mathrm{d}_t\{1 - S_Y(t \,|\, \mathbf{x})\}$. It follows that $\mathcal{S}_{Z|\mathbf{X}} \subset \mathcal{S}_{(Z,D)|\mathbf{X}} \subset \mathcal{S}_{Y|\mathbf{X}} + \mathcal{S}_{C|\mathbf{X}}$. Although existing SDR methods can be applied directly to estimate $\mathcal{S}_{Z|\mathbf{X}}$ and $\mathcal{S}_{(Z,D)|\mathbf{X}}$, they do not always recover $\mathcal{S}_{Y|\mathbf{X}}$ successfully.

To estimate the inverse mean $\mathrm{E}(\mathbf{X} \,|\, Y = y)$ from right-censored survival data, Li, Wang and Chen (1999) first note that

$$\mathrm{E}\{\mathbf{X} \,|\, Y \in [y, y + \delta)\} = \frac{\mathrm{E}\{\mathbf{X}I(Y \ge y)\} - \mathrm{E}\{\mathbf{X}I(Y \ge y + \delta)\}}{\mathrm{E}\{I(Y \ge y)\} - \mathrm{E}\{I(Y \ge y + \delta)\}},$$

for all $\delta > 0$, and

$$\mathrm{E}\{\mathbf{X}I(Y \ge y)\} = \mathrm{E}\{\mathbf{X}I(Z \ge y)\} + \mathrm{E}\left\{\mathbf{X}I(Z < y, D = 0)\frac{S_Y(Z \,|\, \mathbf{X})}{S_Y(y \,|\, \mathbf{X})}\right\},$$

for all y. By plugging-in a suitable initial estimator for $S_Y(t \,|\, \mathbf{x})$, one can easily obtain a sample analogue of $\mathrm{E}(\mathbf{X} \,|\, Y = y)$. Then, we can apply the standard SIR procedure to recover $\mathcal{S}_{Y|\mathbf{X}}$. The only problem is that the estimation for $S_Y(t \,|\, \mathbf{x})$ usually involves multi-dimensional nonparametric smoothing, which is the original reason for considering dimension reduction. Thus, they suggest using $\mathcal{S}_{(Z,D)|\mathbf{X}}$ to obtain an initial estimator for $S_Y(t \,|\, \mathbf{x})$ in a possibly lower dimension of the envelope subspace.

Another method used to adjust for incomplete responses is the inverse censoring probability weighting (ICPW) introduced by Lu and Li (2011). They showed that

$$\mathrm{E}\left\{\frac{Dg(Z)}{S_C(Z\,|\,\mathbf{X})}\,|\,\mathbf{X}\right\} = \mathrm{E}\{g(Y)\,|\,\mathbf{X}\},$$

for all measurable transformations g. Based on this equation, all existing SDR methods developed for complete data can be applied directly to right-censored survival data, with the response variable inverse-weighted by $S_C(Z\,|\,\mathbf{X})$. In a similar spirit, Nadkarni, Zhao and Kosorok (2011) proposed a minimum discrepancy approach, coupled with the ICPW technique, to build a more efficient inverse regression estimator. However, these approaches require additional modeling on the censoring distribution, which may not be desirable in practice.

To relax the linearity condition imposed for the inverse regression methods, Xia, Zhang and Xu (2010) proposed using inverse survival weighting and double kernel smoothing techniques in their hazard-based MAVE (hMAVE) method. However, this also requires an initial estimator for the conditional survival function, as in Li, Wang and Chen (1999). Another problem with the inverse-weighting techniques is that they often lead to unstable estimators in finite samples, especially when the values of the weights are close to zero. To overcome this difficulty, Huang and Chan (2020) proposed a least squares criterion by noting that

$$\mathrm{E}\{I(Z \le y, D = 1)\,|\,\mathbf{X}\} = \int_0^y S_Z(t\,|\,\mathbf{X})\mathrm{d}_t\Lambda(t\,|\,\mathbf{B}_0^{\mathrm{T}}\mathbf{X}),$$

where $\Lambda(t\,|\,\mathbf{B}_0^{\mathrm{T}}\mathbf{x})$ is the conditional cumulative hazard function of Y given $\mathbf{B}_0^{\mathrm{T}}\mathbf{X} = \mathbf{B}_0^{\mathrm{T}}\mathbf{x}$. The major advantage of this approach is that the induced response is not inverse-weighted by the weight function $S_Z(t\,|\,\mathbf{X})$. Thus, it provides a more stable estimator in finite samples.

4.2. Causal inference

A central topic in causal inference is the estimation of the causal effects of the treatments. Let $T \in \{0, 1\}$ be a treatment variable, $Y(1)$ be the response after receiving the treatment, and $Y(0)$ be the response without receiving the treatment. The individual causal effect is often defined as $Y(1) - Y(0)$. However, because only one of $Y(1)$ and $Y(0)$ can be observed for each individual, the individual causal effect can be viewed as an incomplete variable. In such cases, practitioners often focus on the average treatment effect $\tau = \mathrm{E}\{Y(1) - Y(0)\}$,

which can be estimated consistently using standard randomized controlled trials (RCTs). When only observational data are available, Rosenbaum and Rubin (1983) introduced the unconfoundedness assumption

$$\{Y(0), Y(1)\} \perp\!\!\!\perp T \,|\, \mathbf{X}, \tag{4.1}$$

for some covariates/confounders $\mathbf{X}$. Under (4.1), the average treatment effect can be identified nonparametrically as

$$\tau = \mathrm{E}\{\mathrm{E}(Y \,|\, T = 1, \mathbf{X}) - \mathrm{E}(Y \,|\, T = 0, \mathbf{X})\},$$

where $Y = TY(1) + (1 - T)Y(0)$ is the observed response. Thus, we can obtain consistent estimators for τ if we can estimate the conditional treatment effect $\mathrm{E}(Y \,|\, T = 1, \mathbf{X}) - \mathrm{E}(Y \,|\, T = 0, \mathbf{X})$ successfully.

Without parametric or semiparametric modeling assumptions, the estimation for $\mathrm{E}(Y \,|\, T = t, \mathbf{X} = \mathbf{x})$ $(t = 0, 1)$ usually involves nonparametric smoothing, and thus suffers from the curse of dimensionality when the dimension of $\mathbf{X}$ is large. To overcome this difficulty, the propensity score $\pi(\mathbf{X}) = \mathrm{P}(T = 1 \,|\, \mathbf{X})$ is commonly used to estimate τ, because it satisfies

$$\tau = \mathrm{E}[\mathrm{E}\{Y \,|\, T = 1, \pi(\mathbf{X})\} - \mathrm{E}\{Y \,|\, T = 0, \pi(\mathbf{X})\}].$$

Note that $\pi(\mathbf{X}) \in \mathbb{R}$. Thus, if $\pi(\mathbf{X})$ is known, $\mathrm{E}\{Y \,|\, T = t, \pi(\mathbf{X})\}$ $(t = 0, 1)$ can be easily estimated using one-dimensional smoothing estimators. However, in many applications, $\pi(\mathbf{X})$ is unknown and requires additional modeling or estimation. The effective balancing scores introduced by Hu, Follmann and Wang (2014) can be applied to reduce the curse of dimensionality in the nonparametric estimation for τ. They showed that

$$\tau = \mathrm{E}\{\mathrm{E}(Y \,|\, T = 1, \mathbf{B}^{\mathrm{T}}\mathbf{X}) - \mathrm{E}(Y \,|\, T = 0, \mathbf{B}^{\mathrm{T}}\mathbf{X})\},$$

for any basis matrix $\mathbf{B}$ of the central subspace $\mathcal{S}_{T|\mathbf{X}}$, $\mathcal{S}_{Y|\mathbf{X}}$, or $\mathcal{S}_{(T,Y)|\mathbf{X}}$. Thus, existing SDR methods can be applied directly to obtain suitable estimators for $\mathbf{B}$, and hence consistently estimate τ.

The remaining issue is the estimation efficiency. Hahn (1998) showed that the projection of the log-likelihood score onto the true propensity score can be inefficient. In general, using the balancing scores obtained from $\mathcal{S}_{T|\mathbf{X}}$, $\mathcal{S}_{Y|\mathbf{X}}$, or $\mathcal{S}_{(T,Y)|\mathbf{X}}$ may not achieve the semiparametric efficiency bound of estimating τ. To solve this problem, Huang and Chan (2017) introduced a joint central subspace

with a basis matrix $\mathbf{B}$ that satisfies

$$T \perp\!\!\!\perp \mathbf{X} \,|\, \mathbf{B}^{\mathrm{T}}\mathbf{X}, \quad Y(0) \perp\!\!\!\perp \mathbf{X} \,|\, \mathbf{B}^{\mathrm{T}}\mathbf{X}, \quad Y(1) \perp\!\!\!\perp \mathbf{X} \,|\, \mathbf{B}^{\mathrm{T}}\mathbf{X}.$$

They showed that $\mathbf{B}^{\mathrm{T}}\mathbf{X}$ is still an effective balancing score, and that the estimation for τ based on $\mathbf{B}^{\mathrm{T}}\mathbf{X}$ can achieve the semiparametric efficiency bound.

References

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1998). *Efficient and Adaptive Estimation for Semiparametric Models*. Springer-Verlag, New York.

Cambanis, S., Huang, S. and Simons, G. (1981). On the theory of elliptically contoured distributions. *Journal of Multivariate Analysis* **11**, 368–385.

Carroll, R. J. and Li, K.-C. (1992). Measurement error regression with unknown link: Dimension reduction and data visualization. *Journal of the American Statistical Association* **87**, 1040–1050.

Chen, C.-H. and Li, K.-C. (1998). Can SIR be as popular as multiple linear regression? *Statistica Sinica* **8**, 289–316.

Chen, C.-H. and Li, K.-C. (2001). Generalization of Fisher's linear discriminant analysis via the approach of sliced inverse regression. *Journal of the Korean Statistical Society* **30**, 193–217.

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

Cook, R. D. (1994). On the interpretation of regression plots. *Journal of the American Statistical Association* **89**, 177–189.

Cook, R. D. (1996). Graphics for regressions with a binary response. *Journal of the American Statistical Association* **91**, 983–992.

Cook, R. D. (1998). *Regression Graphics*. John Wiley and Sons, Inc., New York.

Cook, R. D., Li, B. and Chiaromonte, F. (2007). Dimension reduction in regression without matrix inversion. *Biometrika* **94**, 569–584.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**, 849–911.

Ferré, L. and Yao, A.-F. (2005). Smoothed functional inverse regression. *Statistica Sinica* **15**, 665–683.

Ferré, L. and Yao, A. F. (2007). Reply to the paper by Liliana Forzani and R. Dennis Cook: "A note on smoothed functional inverse regression". *Statistica Sinica* **17**, 1683–1687.

Forzani, L. and Cook, R. D. (2007). A note on smoothed functional inverse regression. *Statistica Sinica* **17**, 1677–1681.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* **66**, 315–331.

Hilafu, H. (2017). Random sliced inverse regression. *Communications in Statistics-Simulation and Computation* **46**, 3516–3526.

Hilafu, H. and Yin, X. (2017). Sufficient dimension reduction and variable selection for large-p-small-n data with highly correlated predictors. *Journal of Computational and Graphical Statistics* **26**, 26–34.

Hsing, T. and Ren, H. (2009). An RKHS formulation of the inverse regression dimension-reduction problem. *The Annals of Statistics* **37**, 726–755.

Hu, Z., Follmann, D. A. and Wang, N. (2014). Estimation of mean response via the effective balancing score. *Biometrika* **101**, 613–624.

Huang, M.-Y. and Chan, K. C. G. (2017). Joint sufficient dimension reduction and estimation of conditional and average treatment effects. *Biometrika* **104**, 583–596.

Huang, M.-Y. and Chan, K. C. G. (2020). Sufficient dimension reduction with simultaneous estimation of effective dimensions for time-to-event data. *Statistica Sinica* **30**, 1285–1311.

Hung, H. and Huang, S.-Y. (2019). Sufficient dimension reduction via random-partitions for the large-p-small-n problem. *Biometrics* **75**, 245–255.

Jiang, C.-R., Yu, W. and Wang, J.-L. (2014). Inverse regression for longitudinal data. *The Annals of Statistics* **42**, 563–591.

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**, 1177–1186.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association* **86**, 316–342.

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**, 1025–1039.

Li, K.-C., Aragon, Y., Shedden, K. and Thomas Agnan, C. (2003). Dimension reduction for multivariate response data. *Journal of the American Statistical Association* **98**, 99–109.

Li, K.-C., Lue, H.-H. and Chen, C.-H. (2000). Interactive tree-structured regression via principal Hessian directions. *Journal of the American Statistical Association* **95**, 547–560.

Li, K.-C., Wang, J.-L. and Chen, C.-H. (1999). Dimension reduction for censored regression data. *The Annals of Statistics* **27**, 1–23.

Li, L., Cook, R. D. and Tsai, C.-L. (2007). Partial inverse regression. *Biometrika* **94**, 615–625.

Li, Y. and Hsing, T. (2010). Deciding the dimension of effective dimension reduction space for functional and high-dimensional data. *The Annals of Statistics* **38**, 3028–3062.

Lu, W. and Li, L. (2011). Sufficient dimension reduction for censored regression. *Biometrics* **67**, 513–523.

Lue, H.-H. (2009). Sliced inverse regression for multivariate response regression. *Journal of Statistical Planning and Inference* **139**, 2656–2664.

Ma, Y. and Zhang, X. (2015). A validated information criterion to determine the structural dimension in dimension reduction models. *Biometrika* **102**, 409–420.

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

Ma, Y. and Zhu, L. (2013). Efficient estimation in sufficient dimension reduction. *The Annals of Statistics* **41**, 250–268.

Nadkarni, N. V., Zhao, Y. and Kosorok, M. R. (2011). Inverse regression estimation for censored data. *Journal of the American Statistical Association* **106**, 178–190.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.

Székely, G. J., Rizzo, M. L. and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics* **35**, 2769–2794.

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. Springer, New York.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

Wang, H. and Xia, Y. (2008). Sliced regression for dimension reduction. *Journal of the American Statistical Association* **103**, 811–821.

Xia, Y. (2007). A constructive approach to the estimation of dimension reduction directions. *The Annals of Statistics* **35**, 2654–2690.

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

Xia, Y., Zhang, D. and Xu, J. (2010). Dimension reduction and semiparametric estimation of survival models. *Journal of the American Statistical Association* **105**, 278–290.

Yao, F., Lei, E. and Wu, Y. (2015). Effective dimension reduction for sparse functional data. *Biometrika* **102**, 421–437.

Yin, X. and Hilafu, H. (2015). Sequential sufficient dimension reduction for large p, small n problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **77**, 879–892.

Yin, X. and Li, B. (2011). Sufficient dimension reduction based on an ensemble of minimum average variance estimators. *The Annals of Statistics* **39**, 3392–3416.

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

Zeng, P. and Zhu, Y. (2010). An integral transform method for estimating the central mean and central subspaces. *Journal of Multivariate Analysis* **101**, 271–290.

Zhu, L., Miao, B. and Peng, H. (2006). On sliced inverse regression with high-dimensional covariates. *Journal of the American Statistical Association* **101**, 630–643.

Zhu, L., Wang, T., Zhu, L. and Ferré, L. (2010). Sufficient dimension reduction through discretization-expectation estimation. *Biometrika* **97**, 295–304.

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**, 1455–1466.

Zhu, Y. and Zeng, P. (2006). Fourier methods for estimating the central subspace and the central mean subspace in regression. *Journal of the American Statistical Association* **101**, 1638–1651.

Ming-Yueh Huang

Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan.

E-mail: myh0728@stat.sinica.edu.tw

Hung Hung

Institute of Epidemiology and Preventive Medicine, National Taiwan University, Taipei 100, Taiwan.

E-mail: hhung@ntu.edu.tw

(Received May 2022; accepted August 2022)