A Semiparametric Approach to Dimension Reduction

Yanyuan MA and Liping ZHU

We provide a novel and completely different approach to dimension-reduction problems from the existing literature. We cast the dimension-reduction problem in a semiparametric estimation framework and derive estimating equations. Viewing this problem from the new angle allows us to derive a rich class of estimators, and obtain the classical dimension reduction techniques as special cases in this class. The semiparametric approach also reveals that in the inverse regression context while keeping the estimation structure intact, the common assumption of linearity and/or constant variance on the covariates can be removed at the cost of performing additional nonparametric regression. The semiparametric estimators without these common assumptions are illustrated through simulation studies and a real data example. This article has online supplementary material.

KEY WORDS: Estimating equations; Nonparametric regression; Robustness; Semiparametric methods; Sliced inverse regression.

1. INTRODUCTION

Dimension reduction has been an active field of statistical research for the last 20 years and continues to be important due to the increasingly large amount of available covariates in various scientific areas. The goal of dimension reduction is to identify one or multiple directions represented by a matrix $\boldsymbol{\beta}$, so that the response Y relates to the covariate vector $\mathbf{x}$ only through a few linear combinations $\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}$. When the conditional distribution depends on $\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}$, it is a problem of estimating the central space (Cook 1998); when the conditional mean $E(Y \mid \mathbf{x})$ depends on $\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}$, it is a problem of estimating the central mean space (Cook and Li 2002).

Started with the ingenious idea of sliced inverse regression (SIR) in the seminal article by Li (1991), many highly effective methods in the area of dimension reduction have been developed. For identifying the central space, see, for example, sliced average variance estimation (SAVE) (Cook and Weisberg 1991) and directional regression (DR) (Li and Wang 2007) and their variations such as kernel inverse regression (Zhu and Fang 1996), CANCOR analysis (Fung et al. 2002), and so on. These methods extend the inverse regression idea and are promising in recovering the central space. However, they all rely on certain conditions. These conditions mainly include the linearity condition, where $E(\mathbf{x} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$ is assumed to be a linear function of $\mathbf{x}$, and the constant variance condition, where $\mathrm{cov}(\mathbf{x} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$ is assumed to be a constant matrix. These conditions are not always satisfied, and sometimes could imply stringent assumptions on the joint distribution of $\mathbf{x}$. To be precise, SIR requires the linearity condition; SAVE and DR require both the linearity condition and the constant variance condition. If the covariates do not satisfy these two conditions, current practice often relies on transformation (Box and Cox 1964) or reweighting (Cook and Nachtsheim 1994), which can restore these conditions sometimes. Li and Dong (2009) and Dong and Li (2010) successfully remove the linearity condition from the dimension-reduction problems while their estimators remain to be the inverse regression type. This is no doubt a great advancement. The residual issue is that they assumed $E(\mathbf{x} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$ to be a polynomial function of $\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}$ and they still required the constant variance condition. These remaining requirements can still be stringent and difficult to check in practice. Zhu and Zeng (2006) introduced a dimension-reduction method to identify the central subspace through using Fourier transformations. Their method, however, requires one to estimate the joint probability density function (pdf) of $\mathbf{x}$, which is typically infeasible in a high-dimensional environment. To circumvent this difficulty, they assumed $\mathbf{x}$ to be multivariate normal in implementations. Adapting the idea of minimum average variance estimation (MAVE) (Xia et al. 2002), Xia (2007) proposed a similar procedure (dMAVE) to recover the central space, and Wang and Xia (2008) proposed sliced regression (SR) for dimension reduction. However, their methods estimate the distribution function nonparametrically and heavily rely on the implicit assumption that all the covariates are continuous. Because all the existing dimension-reduction methods impose either the above two conditional moment conditions or distributional assumptions on the covariate vector in one form or another, new dimension-reduction methods which are free of any of these assumptions are highly in demand, particularly when some covariates are discrete or categorical.

Similarly, a large amount of literature exists for identifying the central mean space. For example, when the central mean space is one dimensional, Li and Duan (1989) suggested using the ordinary least squares (OLS), assuming $\mathbf{x}$ to satisfy the linearity condition. Härdle and Stoker (1989) and Power, Stock, and Stoker (1989) proposed the average derivative estimation, which requires $\mathbf{x}$ to be continuous. Horowitz and Härdle (1996) proposed a method that allows some covariates to be discrete; however, the number of levels in discrete variables cannot be large. Ichimura (1993) and Härdle, Hall, and Ichimura (1993) suggested using nonlinear least squares, which are essentially special cases of our proposal which will be described in the following. When the central mean space is possibly two or more dimensional, Xia et al. (2002) proposed the minimum average variance estimation method, which, similar to the constructive

Yanyuan Ma is Professor, Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX 77843-3143 (E-mail: *ma@stat.tamu.edu*). Liping Zhu is the corresponding author and Associate Professor, School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai 200433, China (E-mail: *zhu.liping@mail.shufe.edu.cn*).
Yanyuan Ma's work was supported by the National Science Foundation (DMS-0906341) and the National Institute of Neurological Disorders and Stroke (R01-NS073671). Liping Zhu's work was supported by the Natural Science Foundation of China (11071077).

DOI: 10.1080/01621459.2011.646925

approach in Xia (2007) and the sliced regression in Wang and Xia (2008), requires $\mathbf{x}$ to be continuous. Li (1992) and Cook and Li (2002) proposed the method of principal Hessian directions which requires $\mathbf{x}$ to satisfy both the linearity condition and the constant variance condition, which are not always satisfied in practice. Assuming conditional normality of $\mathbf{x}$ on Y, Cook and Forzani (2009) proposed a likelihood-based method, and in the context of classification, Hernández and Velilla (2005) estimated the dimension reduction space via minimizing a criterion function which involves kernel density estimation. Yin and Cook (2005); Yin, Li, and Cook (2008); and Park, Sriram, and Yin (2010) proposed a method to recover the dimension-reduction space via minimizing a Kullback–Leibler distance.

In this article, we provide a completely different viewpoint for looking at the dimension-reduction problems. Our approach is through semiparametrics, which has not been considered in the literature. By casting the dimensional-reduction problem in the semiparametric framework, the dimension-reduction problems become semiparametric estimation problems. Therefore, powerful semiparametric estimation and inference tools become applicable. We use the geometric approach in Bickel et al. (1993) and Tsiatis (2006) to analyze these problems and derive the space of the influence functions. This enables us to construct a rich class of estimators. Many of the existing dimension-reduction methods turn out to be special cases in this class. In fact, the complete class of influence functions provide all the possible consistent estimators.

A direct consequence of the semiparametric analysis is the relaxation of the linearity condition and the constant variance condition. Using the semiparametric construction, we reveal that these conditions are not structurally necessary. Consequently, in all these existing dimension-reduction procedures, we can remove these two conditions, and instead replace the assumed quantity with nonparametric estimation of the corresponding conditional expectations. Thus, the semiparametric derivation allows us to obtain the dimension reduction spaces without any distributional assumption on the covariate vector. Another advantage of the semiparametric analysis is that we do not require all the covariates to be continuous.

In summary, the contributions of this article are:

1. We introduce a novel and drastically different approach to the dimension-reduction field. We anticipate to stimulate deeper and richer literature in this direction.
2. We derive the complete class of influence functions, which guarantees to yield all possible root-n consistent estimators for the column space of $\boldsymbol{\beta}$. We demonstrate how to obtain several most popular dimension-reduction methods from this class. This further reveals the underlying connection between these different methods and provides a different and natural motivation for their construction.
3. We completely eliminate the linearity condition, the constant variance condition, the condition on the quadratic form of the covariates, or, in fact, any moment conditions on the covariates at all.
4. We eliminate the redundant continuity conditions on the covariates. The new approach can be readily used even when some covariates are categorical or discrete.

The outline of this article is the following. In Section 2, we describe the semiparametric approach to the central space estimation and derive a rich class of estimators. We establish their link and generalization to several existing dimension-reduction methods in Section 3. The analysis for central mean space estimation is given in Section 4. We explain the implementation details on estimation and on selecting the dimension of the central space/central mean space in Section 5. Extensive simulation studies are conducted in Section 6 to demonstrate the practical performance and the method is implemented in a real data example in Section 7. We finish the article with a brief discussion in Section 8. Technical derivations are collected in an appendix and the online supplementary document.

2. ESTIMATING THE CENTRAL SUBSPACE VIA SEMIPARAMETRICS

Let $\mathbf{x}$ be a $p \times 1$ covariate vector and Y a univariate response variable. The goal of sufficient dimension reduction (Cook 1998) is to seek a matrix $\boldsymbol{\beta}$ such that

$$F(y \mid \mathbf{x}) = F(y \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}), \text{ for } y \in \mathbb{R}, \tag{1}$$

where $F(y \mid \mathbf{x}) \stackrel{\text{def}}{=} \Pr(Y \le y \mid \mathbf{x})$ denotes the conditional distribution function of Y given $\mathbf{x}$. (1) implies that the response variable Y relates to $\mathbf{x}$ only through linear combinations $\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}$. The column space of $\boldsymbol{\beta}$ satisfying (1) is called a dimension-reduction subspace. Because the dimension-reduction subspace is not unique, our primary interest is the central subspace, which is defined as the intersection of all dimension-reduction subspaces, provided that the intersection itself is a dimension-reduction subspace (Cook 1998).

Following the convention in the area of dimension reduction, we denote by $\mathcal{S}_{Y|\mathbf{x}}$ the central subspace and assume $\boldsymbol{\beta}$ to satisfy $\boldsymbol{\beta}^{\mathrm{T}}\mathrm{cov}(\mathbf{x})\boldsymbol{\beta} = \mathbf{I}_d$. Note that $\boldsymbol{\beta}$ is more restrictive than before, however to avoid introducing a new notation, we keep the same notation. We also assume $\mathcal{S}_{Y|\mathbf{x}}$ exists and is unique. Here, the number of columns in $\boldsymbol{\beta}$, denoted by d, is the dimension of $\mathcal{S}_{Y|\mathbf{x}}$ and is often referred to as the structural dimension. Our goal is to find $\mathcal{S}_{Y|\mathbf{x}}$ through finding $\boldsymbol{\beta} \in \mathbb{R}^{p\times d}$ which satisfies (1). To focus on the main issues of the dimension-reduction problems, we assume throughout our article that the covariate vector $\mathbf{x}$ satisfies $E(\mathbf{x}) = \mathbf{0}$ and $\mathrm{cov}(\mathbf{x}) = \mathbf{I}_p$. This assumption causes no loss of generality, thanks to an invariance property of the central subspace (Cook 1998, p. 106).

In model (1), the likelihood of one random observation $(\mathbf{x}, Y)$ is

$$\eta_1(\mathbf{x})\eta_2(Y, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}),$$

where η_1 is a probability mass function (pmf) of $\mathbf{x}$ or a pdf of $\mathbf{x}$, or a mixture, depending on whether $\mathbf{x}$ contains discrete variables, and η_2 is the conditional pmf/pdf of Y on $\mathbf{x}$. Treating η_1, η_2 as infinite-dimensional nuisance parameters while $\boldsymbol{\beta}$ as the parameter of interest, this can be viewed as a semiparametric estimation problem. The essential idea in semiparametrics is to construct estimators through deriving influence functions. Influence functions can be viewed as normalized elements in a so called nuisance tangent space orthogonal complement $\Lambda^{\perp}$. Thus, if one can successfully derive $\Lambda^{\perp}$, one at least has the hope of characterizing the influence functions and constructing estimators. Because semiparametrics is not a familiar tool in the dimension-reduction community, we give a general and more precise explanation in Appendix 0. It is not

a simple and straightforward tool to grasp and to master, and interested readers are encouraged to refer to Bickel et al. (1993) and Tsiatis (2006) for in-depth understanding. In model (1), using the semiparametric analysis, we characterize the space of the span of all possible score functions, had η_1, η_2 been replaced by all possible parametric submodels, to find the nuisance tangent space. We then derive its orthogonal complement to obtain

$$\begin{aligned}\Lambda^{\perp} &= \{\mathbf{f}(Y,\mathbf{x}) - E\{\mathbf{f}(Y,\mathbf{x}) \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}, Y\} : E\{\mathbf{f}(Y,\mathbf{x}) \mid \mathbf{x}\} \\ &= E\{\mathbf{f}(Y,\mathbf{x}) \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\},\ \forall\, \mathbf{f}(Y,\mathbf{x})\}.\end{aligned}$$

See the detailed derivation in Appendix 1. The form of $\Lambda^{\perp}$ permits many possibilities for constructing consistent estimating equations. For example, for any functions $\mathbf{g}(Y,\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$ and $\boldsymbol{\alpha}(\mathbf{x})$, we can choose $\mathbf{f}(Y,\mathbf{x})$ to be

$$[\mathbf{g}(Y,\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) - E\{\mathbf{g}(Y,\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\}]\,[\boldsymbol{\alpha}(\mathbf{x}) - E\{\boldsymbol{\alpha}(\mathbf{x}) \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\}].$$

Here, $\mathbf{f}(Y,\mathbf{x})$ satisfies $E\{\mathbf{f}(Y,\mathbf{x}) \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}, Y\} = \mathbf{0}$ and is thus a valid element in $\Lambda^{\perp}$. Therefore, a general class of estimating equations can be obtained using the sample version of

$$\begin{aligned}E([\mathbf{g}(Y,\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) - E\{\mathbf{g}(Y,\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\}] \\ \times [\boldsymbol{\alpha}(\mathbf{x}) - E\{\boldsymbol{\alpha}(\mathbf{x}) \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\}]) = \mathbf{0}. \qquad (2)\end{aligned}$$

The resulting estimate is obviously $\sqrt{n}$-consistent (Newey 1990).

Remark 1. Equation (2) is only one convenient way to construct elements in $\Lambda^{\perp}$. Other constructions are also possible. For example, an arbitrary linear combination

$$\begin{aligned}\sum_{i=1}^{k}&[\mathbf{g}_i(Y,\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) - E\{\mathbf{g}_i(Y,\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\}] \\ &\times [\boldsymbol{\alpha}_i(\mathbf{x}) - E\{\boldsymbol{\alpha}_i(\mathbf{x}) \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\}]\end{aligned}$$

also provides a $\sqrt{n}$-consistent estimator because it is a valid element in $\Lambda^{\perp}$.

Remark 2. It is easy to see that solving (2) does not necessarily yield a unique solution. Theoretically, as long as we choose $\mathbf{g}$ and $\boldsymbol{\alpha}$ so that the matrix $\mathbf{A}$ in Theorem 1 has rank $p(p-d)$ and $\mathbf{B}$ is bounded, solving (2) can yield a basis of $\mathcal{S}_{Y|\mathbf{x}}$, although the basis may not be unique; and as long as the dimension-reduction problem is identifiable, such $\mathbf{g}$ and $\boldsymbol{\alpha}$ always exist. Regarding the issue of multiple solutions in practice, there are two aspects to it. First, in a typical estimating equation approach, multiple roots issue presents a challenge. This problem almost always exists in a finite sample, even under the condition that at the population level a unique solution exists. There is no established method to handle it as far as we know. Almost in all the situations, empirical methods are used to select the most sensible root among several roots. Second, uniquely in the context of dimension reduction, even if the targeted central space $\mathcal{S}_{Y|\mathbf{x}}$ is unique, its basis—which is what we solve for—is not. Fortunately, this level of multiple roots issue is not a concern. As any particular choice of the basis will yield the same space, and the space is what we really aim for.

In Appendix 2, we show that (2) has a double robustness property, in that the consistency is doubly assured by the term $\mathbf{g} - E(\mathbf{g} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$ and the term $\boldsymbol{\alpha} - E(\boldsymbol{\alpha} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$. We can misspecify either $E\{\mathbf{g}(Y,\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\}$ or $E\{\boldsymbol{\alpha}(\mathbf{x}) \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\}$, the estimator obtained from (2) will still be consistent. Specifically, if we replace $E\{\boldsymbol{\alpha}(\mathbf{x}) \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\}$ with an arbitrary function $\mathbf{h}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$, then (2) becomes

$$E([\mathbf{g}(Y,\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) - E\{\mathbf{g}(Y,\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\}]\,\{\boldsymbol{\alpha}(\mathbf{x}) - \mathbf{h}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\}) = \mathbf{0},$$

which still yields a consistent estimator. Similarly, if we replace $E\{\mathbf{g}(Y,\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\}$ with an arbitrary function $\mathbf{h}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$, then (2) becomes

$$E(\{\mathbf{g}(Y,\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) - \mathbf{h}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\}\,[\boldsymbol{\alpha}(\mathbf{x}) - E\{\boldsymbol{\alpha}(\mathbf{x}) \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\}]) = \mathbf{0}, \qquad (3)$$

which also yields a consistent estimator.

Remark 3. To ensure the consistency of the estimation of $\boldsymbol{\beta}$, only one of the two expectations $E\{\mathbf{g}(Y,\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\}$ and $E\{\boldsymbol{\alpha}(\mathbf{x}) \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\}$ can be misspecified. The other expectation needs to be calculated consistently, this typically requires specifying a correct parametric model or performing nonparametric regression.

Remark 4. In contrast to the existing literature on sufficient dimension reduction (Cook 1998), when using (2) or its misspecified versions to identify $\mathcal{S}_{Y|\mathbf{x}}$, no additional assumptions are made on the covariate vector. This means that (i) we do not need to assume a specific joint distribution for $\mathbf{x}$; (ii) we do not need to assume the linearity condition or the constant variance condition; and (iii) we do not need to assume $\mathbf{x}$ to be continuously distributed. Although Cook and Li (2005) also considered noncontinuous covariates, they had to assume a parametric model for the distribution of the covariate vector $\mathbf{x}$ conditional on the response Y.

Remark 5. If we are willing to make additional parametric assumptions on $E\{\boldsymbol{\alpha}(\mathbf{x}) \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\}$ in (3), for example, $E\{\boldsymbol{\alpha}(\mathbf{x}) \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\} = \mathbf{C}\mathbf{x}$ or $E\{\boldsymbol{\alpha}(\mathbf{x}) \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\} = \mathbf{C}$ for a quantity $\mathbf{C}$ that may or may not depend on $\boldsymbol{\beta}$, then we will no longer need to perform a nonparametric estimation of $E\{\boldsymbol{\alpha}(\mathbf{x}) \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\}$ when using (3). Such assumptions greatly simplify the computation. We suspect this is the implicit motivation behind the linearity condition and the constant variance condition, which are widely used in the sufficient dimension reduction literature. We will explore this issue in detail in the next two sections.

The double robustness property further allows us to obtain a $\sqrt{n}$-consistent estimator without any undersmoothing requirement even when $d \geq 3$ through nonparametrically estimating both $E\{\mathbf{g}(Y,\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\}$ and $E\{\boldsymbol{\alpha}(\mathbf{x}) \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\}$. We state this result in Theorem 1 and provide the proof in the online supplementary document.

Theorem 1. Under conditions (C1)–(C4) given in Appendix 4, the estimator $\widehat{\boldsymbol{\beta}}$ obtained from the estimating equation

$$\begin{aligned}\sum_{i=1}^{n}&\left[\mathbf{g}(Y_i,\mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}) - \widehat{E}\{\mathbf{g}(Y_i,\mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}) \mid \mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}\}\right] \\ &\times \left[\boldsymbol{\alpha}(\mathbf{x}_i) - \widehat{E}\{\boldsymbol{\alpha}(\mathbf{x}_i) \mid \mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}\}\right] = \mathbf{0}\end{aligned}$$

satisfies

$$\sqrt{n}\mathbf{A}\mathrm{vec}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to \mathcal{N}(\mathbf{0}, \mathbf{B})$$

in distribution, where

$$\begin{aligned}\mathbf{A} &= E[\partial \operatorname{vec}([\mathbf{g}(Y, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) - E\{\mathbf{g}(Y, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\}] \\ &\quad \times [\boldsymbol{\alpha}(\mathbf{x}) - E\{\boldsymbol{\alpha}(\mathbf{x}) \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\}])/\partial\{\operatorname{vec}(\boldsymbol{\beta})\}^{\mathrm{T}}], \\ \mathbf{B} &= \operatorname{cov}\{\operatorname{vec}([\mathbf{g}(Y, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) - E\{\mathbf{g}(Y, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\}] \\ &\quad \times [\boldsymbol{\alpha}(\mathbf{x}) - E\{\boldsymbol{\alpha}(\mathbf{x}) \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\}])\}.\end{aligned}$$

Here $\operatorname{vec}(\mathbf{M})$ denotes the vector formed by concatenating the columns of $\mathbf{M}$.

3. CONNECTION WITH EXISTING METHODS

In this section, we will examine several popular existing sufficient dimension-reduction methods, and illustrate why they are special cases of the semiparametric estimation family. We will show that all these methods take advantage of the double robustness property. In addition, we will point out that the linearity condition and/or the constant variance condition are used in these methods to simplify the computation. To be specific, the linearity condition characterizes the mean of $\mathbf{x}$ conditional on $\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}$ by assuming

$$E(\mathbf{x} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) = \mathbf{P}\mathbf{x}, \tag{4}$$

and the constant variance condition characterizes the variance-covariance matrix of $\mathbf{x}$ conditional on $(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$ by assuming

$$\operatorname{cov}(\mathbf{x} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) = \mathbf{Q}, \tag{5}$$

where $\mathbf{P} = \boldsymbol{\beta}(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^{\mathrm{T}} = \boldsymbol{\beta}\boldsymbol{\beta}^{\mathrm{T}}$, $\mathbf{Q} = \mathbf{I}_p - \mathbf{P}$. Note that here the two conditions are given in the context where $E(\mathbf{x}) = \mathbf{0}$ and $\operatorname{cov}(\mathbf{x}) = \mathbf{I}_p$. Here, both $\mathbf{P}$ and $\mathbf{Q}$ are symmetric matrices.

Before presenting the specific analysis on these methods, We first highlight two simple linear algebra results that will be used frequently in the remaining context. These are simple linear algebra results; hence, we only sketch the proofs in the online supplementary document.

Lemma 1. Assume Λ is a $p \times p$ symmetric matrix of rank d. If and only if $\boldsymbol{\beta}$ satisfies

$$\Lambda - \mathbf{P}\Lambda\mathbf{P} = \mathbf{0},$$

then the span of the columns in $\boldsymbol{\beta}$ is the eigenspace of Λ corresponding to the d nonzero eigenvalues.

Lemma 2. Assume Λ is a $p \times p$ symmetric nonnegative definite matrix of rank d. If and only if $\boldsymbol{\beta}$ satisfies

$$\mathbf{Q}\Lambda\mathbf{Q} = \mathbf{0},$$

then the span of the columns in $\boldsymbol{\beta}$ is the eigenspace of Λ corresponding to the d nonzero eigenvalues.

3.1 Sliced Inverse Regression

The classic SIR (Li 1991) requires $\mathbf{x}$ to satisfy the linearity condition (4). It uses the eigenvectors associated with the d nonzero eigenvalues of the matrix $\Lambda_{\mathrm{SIR}} \stackrel{\mathrm{def}}{=} \operatorname{cov}\{E(\mathbf{x} \mid Y)\}$ to span $\mathcal{S}_{Y|\mathbf{x}}$. For ease of illustration, we assume that Λ_{SIR} has rank d hence excluding some degenerated cases.

To obtain SIR as a semiparametric estimator, we set $\mathbf{g}(Y, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) = E(\mathbf{x} \mid Y)$ and $\boldsymbol{\alpha}(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}$ in (2). The linearity condition promises a parametric form $E\{\boldsymbol{\alpha}(\mathbf{x}) \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\} = \mathbf{x}^{\mathrm{T}}\mathbf{P}$, hence we can use the misspecified version (3) while selecting $\mathbf{h}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) = \mathbf{0}$. This choice of $\mathbf{g}$, $\boldsymbol{\alpha}$, and $\mathbf{h}$ in (3) yields $E\{E(\mathbf{x} \mid Y)\mathbf{x}^{\mathrm{T}}\}(\mathbf{I}_p - \mathbf{P}) = \mathbf{0}$, or equivalently, $\Lambda_{\mathrm{SIR}}\mathbf{Q} = \mathbf{0}$. Because Λ_{SIR} is nonnegative definite and has rank d, hence $\Lambda_{\mathrm{SIR}}\mathbf{Q} = \mathbf{0}$ is equivalent to $\mathbf{Q}\Lambda_{\mathrm{SIR}}\mathbf{Q} = \mathbf{0}$. Lemma 2 indicates that this is equivalent to obtaining the eigenspace of Λ_{SIR} to span $\mathcal{S}_{Y|\mathbf{x}}$.

The above semiparametric derivation of SIR indicates clearly that the linearity condition is not structurally necessary for constructing SIR. When the linearity condition (4) does not hold, we simply lose the convenience of replacing $E\{\boldsymbol{\alpha}(\mathbf{x}) \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\}$ by $\mathbf{x}^{\mathrm{T}}\mathbf{P}$, everything else remains unchanged. Specifically, in this case, SIR becomes

$$\text{SIR}: \quad E[\{E(\mathbf{x} \mid Y) - E\{E(\mathbf{x} \mid Y) \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\}\}\{\mathbf{x} - E(\mathbf{x} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\}^{\mathrm{T}}] = \mathbf{0},$$

where $E(\cdot \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$ and $E(\cdot \mid Y)$ need to be estimated nonparametrically. This is what we propose as the semiparametric generalization of SIR in the absence of the linearity condition.

3.2 Sliced Average Variance Estimation

The SAVE (Cook and Weisberg 1991) assumes both the linearity condition (4) and the constant variance condition (5). Similar to SIR, SAVE uses the eigenvectors associated with the d nonzero eigenvalues of a matrix Λ_{SAVE} to span $\mathcal{S}_{Y|\mathbf{x}}$, where $\Lambda_{\mathrm{SAVE}} \stackrel{\mathrm{def}}{=} E[\{\mathbf{I}_p - \operatorname{cov}(\mathbf{x} \mid Y)\}^2]$.

To obtain SAVE from the semiparametric approach, we define $\mathbf{g}_1(Y, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) = \mathbf{I}_p - \operatorname{cov}(\mathbf{x} \mid Y)$, $\mathbf{g}_2(Y, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) = \mathbf{g}_1(Y, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})E(\mathbf{x} \mid Y)$ and $\boldsymbol{\alpha}_1(\mathbf{x}) = -\mathbf{x}\{\mathbf{x} - E(\mathbf{x} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\}^{\mathrm{T}}$, $\boldsymbol{\alpha}_2(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}$. As we have pointed out in Remark 1, $\sum_{i=1}^{2}\{\mathbf{g}_i(Y, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) - E(\mathbf{g}_i \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\}\{\boldsymbol{\alpha}_i(\mathbf{x}) - E(\boldsymbol{\alpha}_i \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\}$ is an element in $\Lambda^{\perp}$, hence it yields a valid semiparametric estimating equation. In this construction, taking advantage of the double robustness, we are allowed to misspecify $E(\mathbf{g}_1 \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) = \mathbf{0}$ and $E(\mathbf{g}_2 \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) = \mathbf{0}$. Some algebra then yields

$$\begin{aligned}\text{SAVE}: \quad & E(\{\mathbf{I}_p - \operatorname{cov}(\mathbf{x} \mid Y)\}[\{\mathbf{x} - E(\mathbf{x} \mid Y)\}\{\mathbf{x} - E(\mathbf{x} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\}^{\mathrm{T}} \\ & - \operatorname{cov}(\mathbf{x} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})]) = \mathbf{0}.\end{aligned} \tag{6}$$

The linearity condition (4) and the constant variance condition (5) further allow us to replace $E(\mathbf{x} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$ and $\operatorname{cov}(\mathbf{x} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$ with $\mathbf{P}\mathbf{x}$ and $\mathbf{Q}$, which directly simplifies (6) to

$$\Lambda_{\mathrm{SAVE}}\mathbf{Q} = \mathbf{0}.$$

Because Λ_{SAVE} is nonnegative definite, solving $\Lambda_{\mathrm{SAVE}}\mathbf{Q} = \mathbf{0}$ is equivalent to solving $\mathbf{Q}\Lambda_{\mathrm{SAVE}}\mathbf{Q} = \mathbf{0}$, which is equivalent to SAVE because of Lemma 2.

Relaxing the linearity condition and the constant variance condition is now obvious. Because (6) is obtained without using these two conditions, we can simply use (6) as the condition-free semiparametric generalization of SAVE, while in implementation, we need to estimate $E(\mathbf{x} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$ and $\operatorname{cov}(\mathbf{x} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$ nonparametrically.

3.3 Directional Regression

Like SAVE, the DR (Li and Wang 2007) also assumes both the linearity condition (4) and the constant variance condition (5). It uses the eigenvectors associated with the d nonzero eigenvalues of the matrix Λ_{DR} to span $\mathcal{S}_{Y|\mathbf{x}}$. Here, $\Lambda_{\mathrm{DR}} \stackrel{\mathrm{def}}{=} E[\{2\mathbf{I}_p - \mathbf{A}(Y, \widetilde{Y})\}^2]$, $\mathbf{A}(Y, \widetilde{Y}) = E\{(\mathbf{x} - \widetilde{\mathbf{x}})(\mathbf{x} - \widetilde{\mathbf{x}})^{\mathrm{T}} \mid Y, \widetilde{Y}\}$, and $(\widetilde{\mathbf{x}}, \widetilde{Y})$ is an independent copy of $(\mathbf{x}, Y)$.

To obtain DR from the semiparametric approach, we choose $\mathbf{g}_1(Y, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) = \mathbf{I}_p - E(\mathbf{x}\mathbf{x}^{\mathrm{T}} \mid Y)$, $\mathbf{g}_2(Y, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) = E\{E(\mathbf{x} \mid Y)E(\mathbf{x}^{\mathrm{T}} \mid Y)\}E(\mathbf{x} \mid Y)$, $\mathbf{g}_3(Y, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) = E\{E(\mathbf{x}^{\mathrm{T}} \mid Y)E(\mathbf{x} \mid Y)\}E(\mathbf{x} \mid Y)$, $\boldsymbol{\alpha}_1(\mathbf{x}) = -\mathbf{x}\{\mathbf{x} - E(\mathbf{x} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\}^{\mathrm{T}}$ and $\boldsymbol{\alpha}_2(\mathbf{x}) = \boldsymbol{\alpha}_3(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}$. Remark 1 indicates that $\sum_{j=1}^{3}\{\mathbf{g}_j(Y, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) - E(\mathbf{g}_j \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\}\{\boldsymbol{\alpha}_j(\mathbf{x}) - E(\boldsymbol{\alpha}_j \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\}$ is an element in $\Lambda^{\perp}$. Taking advantage of the double robustness property, we misspecify $E(\mathbf{g}_j \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) = \mathbf{0}$, for $j = 1, 2, 3$. The subsequent estimating equation is therefore the sample version of

$$\begin{aligned}&\text{DR}:\\ &E[\{\mathbf{I}_p - E(\mathbf{x}\mathbf{x}^{\mathrm{T}} \mid Y)\}\{-\mathbf{x}\mathbf{x}^{\mathrm{T}} + \mathbf{x}E(\mathbf{x}^{\mathrm{T}} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) + \operatorname{cov}(\mathbf{x} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\}]\\ &+ E[E\{E(\mathbf{x} \mid Y)E(\mathbf{x}^{\mathrm{T}} \mid Y)\}E(\mathbf{x} \mid Y)\{\mathbf{x}^{\mathrm{T}} - E(\mathbf{x}^{\mathrm{T}} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\}]\\ &+ E[E\{E(\mathbf{x}^{\mathrm{T}} \mid Y)E(\mathbf{x} \mid Y)\}E(\mathbf{x} \mid Y)\{\mathbf{x}^{\mathrm{T}} - E(\mathbf{x}^{\mathrm{T}} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\}] = \mathbf{0}. \quad (7)\end{aligned}$$

When both the linearity condition (4) and the constant variance condition (5) hold, we can insert $E(\mathbf{x} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) = \mathbf{P}\mathbf{x}$ and $\operatorname{cov}(\mathbf{x} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) = \mathbf{Q}$ in (7). Some algebra then leads to the equivalence between (7) and

$$\begin{aligned}\mathbf{0} &= E[\{\mathbf{I}_p - E(\mathbf{x}\mathbf{x}^{\mathrm{T}} \mid Y)\}^2]\mathbf{Q} + [E\{E(\mathbf{x} \mid Y)E(\mathbf{x}^{\mathrm{T}} \mid Y)\}]^2\mathbf{Q}\\ &\quad + E\{E(\mathbf{x}^{\mathrm{T}} \mid Y)E(\mathbf{x} \mid Y)\}E\{E(\mathbf{x} \mid Y)E(\mathbf{x}^{\mathrm{T}} \mid Y)\}\mathbf{Q}\\ &= 2^{-1}\Lambda_{\mathrm{DR}}\mathbf{Q},\end{aligned}$$

where the last equality is due to Li and Wang (2007). Because Λ_{DR} is nonnegative definite, solving $\Lambda_{\mathrm{DR}}\mathbf{Q} = \mathbf{0}$ is equivalent to solving $\mathbf{Q}\Lambda_{\mathrm{DR}}\mathbf{Q} = \mathbf{0}$, which is equivalent to DR because of Lemma 2.

Similar to SAVE, as (7) is obtained without any of the linearity or constant variance condition, it can thus be used as a semiparametric generalization of DR.

Li and Dong (2009) and Dong and Li (2010) extended SIR, SAVE, and DR to the case when the linearity condition (4) is violated while the constant variance condition (5) is true. Similar analysis shows that these are also special cases of the semiparametric approach. These results are available in the supplementary document.

4. SEMIPARAMETRIC ESTIMATION OF THE CENTRAL MEAN SUBSPACE

In situations when one only concerns about the conditional mean of the response given the predictors, Cook and Li (2002) introduced the notion of the central mean subspace. They defined the column space of $\boldsymbol{\beta}$ as a mean dimension-reduction subspace if $\boldsymbol{\beta}$ satisfies

$$E(Y \mid \mathbf{x}) = E(Y \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}).$$

The intersection of all mean dimension-reduction subspaces is defined as the central mean subspace, denoted by $\mathcal{S}_{E(Y|\mathbf{x})}$, if the intersection itself is also a mean dimension-reduction subspace. The conditional mean model assumes the mean of Y conditional on $\mathbf{x}$ relies on $\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}$ only. In other words, $\mathbf{x}$ contributes to the conditional mean of Y only through $\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}$. Our main interest is to estimate $\mathcal{S}_{E(Y|\mathbf{x})}$, or equivalently, a basis matrix $\boldsymbol{\beta}$ which spans $\mathcal{S}_{E(Y|\mathbf{x})}$.

To facilitate the semiparametric analysis, we write the conditional mean model as

$$Y = \ell(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) + \epsilon, \quad (8)$$

where $\ell(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) \stackrel{\mathrm{def}}{=} E(Y \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$ is an unspecified smooth function and $E(\epsilon \mid \mathbf{x}) = 0$. We emphasize that because we make no assumptions on ϵ other than conditional mean zero, (8) is equivalent to the central mean subspace model. For the conditional mean model (8), the likelihood of one random observation $(\mathbf{x}, Y)$ is

$$\eta_1(\mathbf{x})\eta_2\{Y - \ell(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}), \mathbf{x}\},$$

where η_1 has the same meaning as in Section 2, $\ell(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$ is the mean function of Y conditional on $\mathbf{x}$ (or equivalently, on $\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}$), and η_2 is the conditional pmf/pdf of the residual $\epsilon = Y - E(Y \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$ on $\mathbf{x}$. Here, η_2 satisfies $E(\epsilon \mid \mathbf{x}) = 0$, and is otherwise unconstrained. Similarly, treating η_1, η_2, and ℓ as nuisance parameters while $\boldsymbol{\beta}$ as the parameter of interest, following the semiparametric analysis in Appendix 3, we obtain the nuisance tangent space orthogonal complement to be the class of the form

$$\Lambda^{\perp} = \{\{Y - E(Y \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\}\{\boldsymbol{\alpha}(\mathbf{x}) - E(\boldsymbol{\alpha} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\} : \forall\boldsymbol{\alpha}\}.$$

Similar to (1), the form of $\Lambda^{\perp}$ allows us to take any $\boldsymbol{\alpha}(\mathbf{x})$ to obtain

$$\{Y - E(Y \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\}\{\boldsymbol{\alpha}(\mathbf{x}) - E(\boldsymbol{\alpha}(\mathbf{x}) \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\}$$

as a valid element in $\Lambda^{\perp}$. Hence, a general class of estimating equations can be obtained using the sample version of

$$E(\{Y - E(Y \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\}[\boldsymbol{\alpha}(\mathbf{x}) - E\{\boldsymbol{\alpha}(\mathbf{x}) \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\}]) = 0. \quad (9)$$

Similar to Appendix 2, we can easily show that (9) has a double robustness property, in that we can misspecify either $E(Y \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$ or $E\{\boldsymbol{\alpha}(\mathbf{x}) \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\}$, the resulting estimator from (9) will still yield a consistent estimating equation.

4.1 Ordinary Least Squares

We first inspect the OLS (Li and Duan 1989) method, where the linearity condition (4) is assumed to hold. The OLS method uses $\operatorname{cov}(\mathbf{x}, Y)$ to infer a subspace of the column space of $\boldsymbol{\beta}$ in (8).

From the semiparametric approach, we let $\boldsymbol{\alpha}(\mathbf{x}) = \mathbf{x}$ in (9). Taking advantage of the double robustness, we misspecify $E(Y \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) = 0$. Then (9) reduces to

$$\mathbf{0} = E(\mathbf{x}Y) - \mathbf{P}E(\mathbf{x}Y) = E(\mathbf{x}Y) - \boldsymbol{\beta}(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^{\mathrm{T}}E(\mathbf{x}Y).$$

Note that $(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^{\mathrm{T}}E(\mathbf{x}Y)$ is a $d \times 1$ vector. This directly yields $\operatorname{cov}(\mathbf{x}, Y)$ as a one-dimensional subspace of $\mathcal{S}_{E(Y|\mathbf{x})}$, which is exactly the OLS estimation.

When the linearity condition (4) does not hold, (9) has the form

$$E[\{Y - E(Y \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\}\{\mathbf{x} - E(\mathbf{x} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\}] = \mathbf{0}$$

and can still be used to estimate $\boldsymbol{\beta}$. A simple treatment is to set $E(\mathbf{x} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) = 0$ and solve the sample version of the above display to obtain $\boldsymbol{\beta}$.

4.2 Principal Hessian Directions

The principal Hessian directions (PHD) method (Li 1992) assumes both the linearity condition (4) and the constant variance condition (5). It uses the eigenvectors associated with d nonzero eigenvalues of Λ_{PHD} to form a basis of $\mathcal{S}_{E(Y|\mathbf{x})}$. Here,

$\Lambda_{\text{PHD}} \overset{\text{def}}{=} E[\{Y - E(Y)\}\mathbf{x}\mathbf{x}^{\mathrm{T}}]$. To obtain PHD from the semiparametric approach, we let $\boldsymbol{\alpha}(\mathbf{x}) = \mathbf{x}\mathbf{x}^{\mathrm{T}}$ in (9). Taking advantage of the double robustness, we misspecify $E(Y \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) = E(Y)$. The linearity condition (4) and the constant variance condition (5) yield a simplification $\boldsymbol{\alpha}(\mathbf{x}) - E\{\boldsymbol{\alpha}(\mathbf{x}) \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\} = \mathbf{x}\mathbf{x}^{\mathrm{T}} - \mathbf{Q} - \mathbf{P}\mathbf{x}\mathbf{x}^{\mathrm{T}}\mathbf{P}$, hence (9) reduces to

$$E[\{Y - E(Y)\}(\mathbf{x}\mathbf{x}^{\mathrm{T}} - \mathbf{P}\mathbf{x}\mathbf{x}^{\mathrm{T}}\mathbf{P})] = \Lambda_{\text{PHD}} - \mathbf{P}\Lambda_{\text{PHD}}\mathbf{P} = \mathbf{0}.$$

Lemma 1 indicates that this is equivalent to the PHD method.

When either (4) or (5) is not true, we can use

$$E[\{Y - E(Y \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\}\{\mathbf{x}\mathbf{x}^{\mathrm{T}} - E(\mathbf{x}\mathbf{x}^{\mathrm{T}} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\}] = \mathbf{0}$$

to estimate $\boldsymbol{\beta}$, where we calculate $E(Y \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$ and $E(\mathbf{x}\mathbf{x}^{\mathrm{T}} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$ nonparametrically. We can opt to misspecify $E(Y \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) = 0$ or preferably $E(\mathbf{x}\mathbf{x}^{\mathrm{T}} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) = \mathbf{0}$ to simplify the computation. The second simplification

$$\text{PHD}: \quad E[\{Y - E(Y \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\}\mathbf{x}\mathbf{x}^{\mathrm{T}}] = \mathbf{0}$$

will be considered as the semiparametric generalization of the PHD method without extra conditions.

5. IMPLEMENTATION

To focus on delivering the main message, we have avoided detailing the implementation details in practice, which we explain now.

The proposed semiparametric counterparts of SIR, SAVE, DR, PHD, and so on, which were respectively denoted by semi-SIR, semi-SAVE, semi-DR, and semi-PHD for ease of subsequent illustration, all have the similar components of estimating conditional expectation and solving estimating equations. Because the number of the estimating equations sometimes is larger than the number of parameters, the implementation is through minimizing their Frobenius norm. We use the familiar Newton–Raphson procedure to numerically obtain the minimizer. This essentially means that at the jth iteration $\boldsymbol{\beta}^{(j)}$, we perform nonparametric conditional estimation using, say, a kernel regression method to evaluate an estimating equation at $\boldsymbol{\beta}^{(j)}$, and use numerical difference to obtain the derivative of the estimating equation with respect to $\boldsymbol{\beta}$ evaluated at $\boldsymbol{\beta}^{(j)}$. We then update $\boldsymbol{\beta}^{(j)}$. This process iterates until the difference between two consecutive candidates are sufficiently small.

Because semi-DR has the most complicated form, we now outline an algorithm for semi-DR as a concrete illustration. We use $\boldsymbol{\beta}^{(j)}$ to denote the value of $\widehat{\boldsymbol{\beta}}$ at the jth iteration, and use $K(\cdot)$ to denote a kernel function. Let $K_h(\cdot) = K(\cdot/h)/h$ for any bandwidth h. We use the Epanechnikov kernel function in the implementation. Assume the observations are $(\mathbf{x}_i, Y_i)$ for $i = 1, \ldots, n$.

1. Nonparametrically estimate $E(\mathbf{x}\mathbf{x}^{\mathrm{T}} \mid Y)$ and $E(\mathbf{x} \mid Y)$ using
$$\widehat{E}(\mathbf{x}\mathbf{x}^{\mathrm{T}} \mid Y) = \frac{\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}K_h(Y - Y_i)}{\sum_{i=1}^n K_h(Y - Y_i)} \quad \text{and}$$
$$\widehat{E}(\mathbf{x} \mid Y) = \frac{\sum_{i=1}^n \mathbf{x}_i K_h(Y - Y_i)}{\sum_{i=1}^n K_h(Y - Y_i)}.$$
The bandwidth can be selected using the classic cross-validation procedure. We evaluate the above estimations at Y_i, for $i = 1, \ldots, n$.
2. Form $\mathbf{A}_i = \mathbf{I}_p - \widehat{E}(\mathbf{x}\mathbf{x}^{\mathrm{T}} \mid Y_i)$, $\mathbf{b}_i = n^{-1}\sum_{j=1}^n\{\widehat{E}(\mathbf{x} \mid Y_j)\widehat{E}(\mathbf{x}^{\mathrm{T}} \mid Y_j)\}\widehat{E}(\mathbf{x} \mid Y_i)$ and $\mathbf{c}_i = n^{-1}\sum_{j=1}^n\{\widehat{E}(\mathbf{x}^{\mathrm{T}} \mid Y_j)\widehat{E}(\mathbf{x} \mid Y_j)\}\widehat{E}(\mathbf{x} \mid Y_i)$.
3. Pick an arbitrary starting value $\boldsymbol{\beta}^{(1)}$, for example, the classical DR estimate.
4. At the jth iteration, form $\mathbf{t}_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{(j)}$ for $i = 1, \ldots, n$.
5. Nonparametrically estimate $E(\mathbf{x} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^{(j)})$ and $\text{cov}(\mathbf{x} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^{(j)})$ using
$$\widehat{E}(\mathbf{x} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^{(j)}) = \frac{\sum_{i=1}^n \mathbf{x}_i K_h(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^{(j)} - \mathbf{t}_i)}{\sum_{i=1}^n K_h(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^{(j)} - \mathbf{t}_i)}$$
and
$$\widehat{\text{cov}}(\mathbf{x} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^{(j)}) = \frac{\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}K_h(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^{(j)} - \mathbf{t}_i)}{\sum_{i=1}^n K_h(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^{(j)} - \mathbf{t}_i)} - \widehat{E}(\mathbf{x} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^{(j)})\widehat{E}(\mathbf{x}^{\mathrm{T}} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^{(j)}).$$
We use a same bandwidth to estimate the above two quantities to ensure that $\widehat{\text{cov}}(\mathbf{x} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^{(j)})$ is positive definite. The bandwidth can be selected using the cross-validation procedure. Evaluate the above estimations at $\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^{(j)} = \mathbf{t}_1, \ldots, \mathbf{t}_n$.
6. Form the sample version of the left-hand side of (7):
$$r(\boldsymbol{\beta}^{(j)}) = \frac{1}{n}\sum_{i=1}^n \big[\mathbf{A}_i\left\{-\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} + \mathbf{x}_i\widehat{E}(\mathbf{x}^{\mathrm{T}} \mid \mathbf{t}_i) + \widehat{\text{cov}}(\mathbf{x} \mid \mathbf{t}_i)\right\} + (\mathbf{b}_i + \mathbf{c}_i)\left\{\mathbf{x}_i - \widehat{E}(\mathbf{x}_i \mid \mathbf{t}_i)\right\}^{\mathrm{T}}\big].$$
Update $\boldsymbol{\beta}^{(j)}$ using the following Newton–Raphson steps.
7. Form the derivative of $\partial\{\|r(\boldsymbol{\beta})\|^2\}/\partial\{\text{vec}(\boldsymbol{\beta})\}$ evaluated at $\boldsymbol{\beta}^{(j)}$ through numerical difference. Specifically, let δ be a small number and decide an order for the elements in $\boldsymbol{\beta}^{(j)}$. Let $\boldsymbol{\beta}_{k+}^{(j)} = \boldsymbol{\beta}^{(j)} + \delta\mathbf{e}_k$ and $\boldsymbol{\beta}_{k-}^{(j)} = \boldsymbol{\beta}^{(j)} - \delta\mathbf{e}_k$. Here $\mathbf{e}_k$ has the same size as $\boldsymbol{\beta}$ but has 1 in the kth entry and zero elsewhere. Repeat the procedure in Step 5 to obtain $r(\boldsymbol{\beta}_{k+}^{(j)})$ and $r(\boldsymbol{\beta}_{k-}^{(j)})$. Set the kth row of $\partial\{\|r(\boldsymbol{\beta})\|^2\}/\partial\{\text{vec}(\boldsymbol{\beta})\}$ to be $\{\|r(\boldsymbol{\beta}_{k+}^{(j)})\|^2 - \|r(\boldsymbol{\beta}_{k-}^{(j)})\|^2\}/(2\delta)$. Repeat this for all the entries in $\boldsymbol{\beta}^{(j)}$. For simplicity, we denote the resulting first derivative $\partial\{\|r(\boldsymbol{\beta}^{(j)})\|^2\}/\partial\{\text{vec}(\boldsymbol{\beta})\}$. Following similar procedures, we obtain the second derivative $\partial^2\{\|r(\boldsymbol{\beta})\|^2\}/\partial\{\text{vec}(\boldsymbol{\beta})\}\partial\{\text{vec}(\boldsymbol{\beta})\}^{\mathrm{T}}$.
Practically, $\delta = 0.001$ is sufficiently small to obtain a close approximation of the derivatives. If more precision is desired, one can simply opt for smaller δ.
8. Update to obtain
$$\text{vec}(\boldsymbol{\beta}^{(j+1)}) = \text{vec}(\boldsymbol{\beta}^{(j)}) - \left[\frac{\partial^2\{\|r(\boldsymbol{\beta}^{(j)})\|^2\}}{\partial\{\text{vec}(\boldsymbol{\beta}^{(j)})\}\partial\{\text{vec}(\boldsymbol{\beta}^{(j)})\}^{\mathrm{T}}}\right]^{-1} \times \frac{\partial\{\|r(\boldsymbol{\beta}^{(j)})\|^2\}}{\partial\{\text{vec}(\boldsymbol{\beta}^{(j)})\}}.$$
9. Repeat Steps 4 to 8 until convergence.

For identification, $\mathbf{x}$ must contain at least one continuous variable (Ichimura 1993; Horowitz and Härdle 1996). The above algorithm is applicable if the support of $\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}$ is connected. When

the support consists of several disjoint regions, we need to carry out the kernel estimation in Step 5 in each region.

Another important component in dimension reduction is to decide the dimension d of $\mathcal{S}_{Y|\mathbf{x}}$ or $\mathcal{S}_{E(Y|\mathbf{x})}$. To this end, the bootstrap procedure described by Dong and Li (2010) adapts the idea of Ye and Weiss (2003) by taking into account the variation of the covariates, and is very flexible. This is what we recommend for use in our context. Specifically, we can decide d through the following procedure. Let $\lambda_1, \ldots, \lambda_k$ be the nonzero eigenvalues of

$$\{\mathrm{var}(\mathbf{u})\}^{-1/2}\,\mathrm{cov}(\mathbf{u}, \mathbf{v}^{\mathrm{T}})\,\{\mathrm{var}(\mathbf{v})\}^{-1}\,\mathrm{cov}(\mathbf{v}, \mathbf{u}^{\mathrm{T}})\,\{\mathrm{var}(\mathbf{u})\}^{-1/2},$$

where $\mathbf{u}$ and $\mathbf{v}$ are two generic random vectors. In addition, we let

$$r^2(\mathbf{u}, \mathbf{v}) = k^{-1}\sum_{i=1}^{k}\lambda_i.$$

For any working dimension $k = 1, \ldots, p-1$, we let $\widehat{\boldsymbol{\beta}}_k$ be the estimate based on the original sample, and $\widehat{\boldsymbol{\beta}}_{k,b}$ be the estimate based on the bth bootstrap sample, for $b = 1, \ldots, B$. We then estimate the structural dimension d by maximizing over $k = 1, \ldots, p-1$

$$\bar{r}_k^2 = \frac{1}{B}\sum_{b=1}^{B} r^2\left(\widehat{\boldsymbol{\beta}}_k^{\mathrm{T}}\mathbf{x}, \widehat{\boldsymbol{\beta}}_{k,b}^{\mathrm{T}}\mathbf{x}\right). \tag{10}$$

Interested readers are referred to Dong and Li (2010) for more details.

6. SIMULATION STUDY

In this section, we conduct simulation studies to evaluate the performance of different estimation procedures. Unless otherwise stated, we repeat the experiments 500 times each with sample size $n = 200$ and reduced space dimension $d = 2$. Throughout the simulations, we used the Epanechnikov kernel and fixed $\widehat{\sigma}(3n/4)^{-1/(d+4)}$ as bandwidth, where $\widehat{\sigma}$ is the robust estimation of the standard deviation of $\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}$, which is the default bandwidth selector implemented in Matlab routine `ksdensity`. We choose the predictor dimension p to be 6 and 12, and consider the following two cases for the covariate vector $\mathbf{x} = (X_1, \ldots, X_p)^{\mathrm{T}}$.

Case 1: We generate $(X_1, X_2)^{\mathrm{T}}$ (corresponding to the case $p = 6$) and $\left(X_1, X_2, X_7, \ldots, X_p\right)^{\mathrm{T}}$ (corresponding to the case $p = 12$) from normal population with mean zero and variance-covariance matrix $(\sigma_{ij})_{(p-4)\times(p-4)}$ where $\sigma_{ij} = 0.5^{|i-j|}$. We generate X_3 and X_4 from nonlinear models: $X_3 = \mid X_1 + X_2 \mid + \mid X_1 \mid \epsilon_1$, and $X_4 = \mid X_1 + X_2 \mid^2 + \mid X_2 \mid \epsilon_2$, where ϵ_i's are independently generated from the standard normal population, X_5 from a Bernoulli distribution with success probability $\exp(X_2)/\{1 + \exp(X_2)\}$, and X_6 from another Bernoulli distribution with success probability $\Phi(X_2)$, where $\Phi(\cdot)$ denotes the cumulative distribution function of standard normal population. Note that in this case, both the linearity condition (4) and the constant variance condition (5) are violated.

Case 2: We generate $\mathbf{x}$ from normal population with mean zero and variance-covariance matrix $(\sigma_{ij})_{p\times p}$ where $\sigma_{ij} = 0.5^{|i-j|}$. Note that in this case, the covariates satisfy both the linearity condition (4) and the constant variance condition (5).

Let $\boldsymbol{\beta}$ be a basis matrix of $\mathcal{S}_{Y|\mathbf{x}}$ and $\widehat{\boldsymbol{\beta}}$ be its estimate. To assess the estimation accuracy of $\widehat{\boldsymbol{\beta}}$, we use the Euclidean distance between $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$, defined as the Frobenius norm of the matrix $\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})^{-1}\widehat{\boldsymbol{\beta}}^{\mathrm{T}} - \boldsymbol{\beta}(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^{\mathrm{T}}$. In both cases, the distance ranges from zero to two, and a smaller distance indicates a better estimate.

6.1 Example 1

We generate the response variable using the following four different models:

$$\begin{aligned}
&\text{model (I)}: \quad Y = \left(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_1\right)/\left\{0.5 + \left(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_2 + 1.5\right)^2\right\} + 0.5\varepsilon;\\
&\text{model (II)}: \quad Y = \left(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_1\right)^2 + 2 \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_2 \mid + 0.1 \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_2 \mid \varepsilon;\\
&\text{model (III)}: \quad Y = \exp\left(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_1\right) + 2\left(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_2 + 1\right)^2 + \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_1 \mid \varepsilon;\\
&\text{model (IV)}: \quad Y = \left(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_1\right)^2 + \left(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_2\right)^2 + 0.5\varepsilon,
\end{aligned}$$

where $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are $p \times 1$ vectors with their first six components being $(1, 1, 1, 1, 1, 1)^{\mathrm{T}}/\sqrt{6}$ and $(1, -1, 1, -1, 1, -1)^{\mathrm{T}}/\sqrt{6}$, respectively. When $p = 12$, the rest components of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are identically zero. The error term ε has a standard normal distribution. Models (I)–(IV) are chosen to compare, respectively, SIR, SAVE, DR, and PHD with their semiparametric counterparts. We also include dMAVE (Xia 2007) and MAVE (Xia et al. 2002) into our comparison as they are often used as a benchmark. To make a fair comparison on the core methodologies of these proposals, we estimate the kernel matrices of the classical SIR, SAVE, DR by using kernel smoothing rather than the usual slicing estimation. This allows us to avoid selecting the number of slices which usually adversely affects the performance. Thus, SIR, SAVE, and DR are implemented in their improved form.

The boxplots of the Euclidean distances are reported in Figure 1. The results under Case 1 are presented in panels (A) and (C), where we show the boxplots of Euclidean distances for different estimation procedures when both the linearity condition (4) and the constant variance condition (5) are violated. In this case, we can see that the semiparametric estimates are substantially more accurate than their classical dimension-reduction counterparts across all four models, indicating the significant improvement when the violation of these conditions is taken into account. Note that dMAVE has very large variability in these settings and sometimes even perform worse than the classical dimension-reduction methods. This is because some covariates are discrete which violates the continuity requirement of dMAVE. Because the semiparametric estimation procedures do not require continuous covariates, and they do not rely on any conditional moment or distributional assumptions, their performance dominates the competitors in all scenarios. The results for Case 2 are in panels (B) and (D), which contain the boxplots of Euclidean distances when both the linearity condition (4) and the constant variance condition (5) are satisfied. We can see that, surprisingly, our semiparametric proposals are still superior to their classical counterparts in these particular examples. This reminds us of a quite interesting phenomenon where sometimes, estimating a quantity even if it is known brings gain (Henmi and Eguchi 2004). However, whether this gain is real, in general, or just in these specific examples deserves further theoretical investigation. On the other hand, our estimators seem

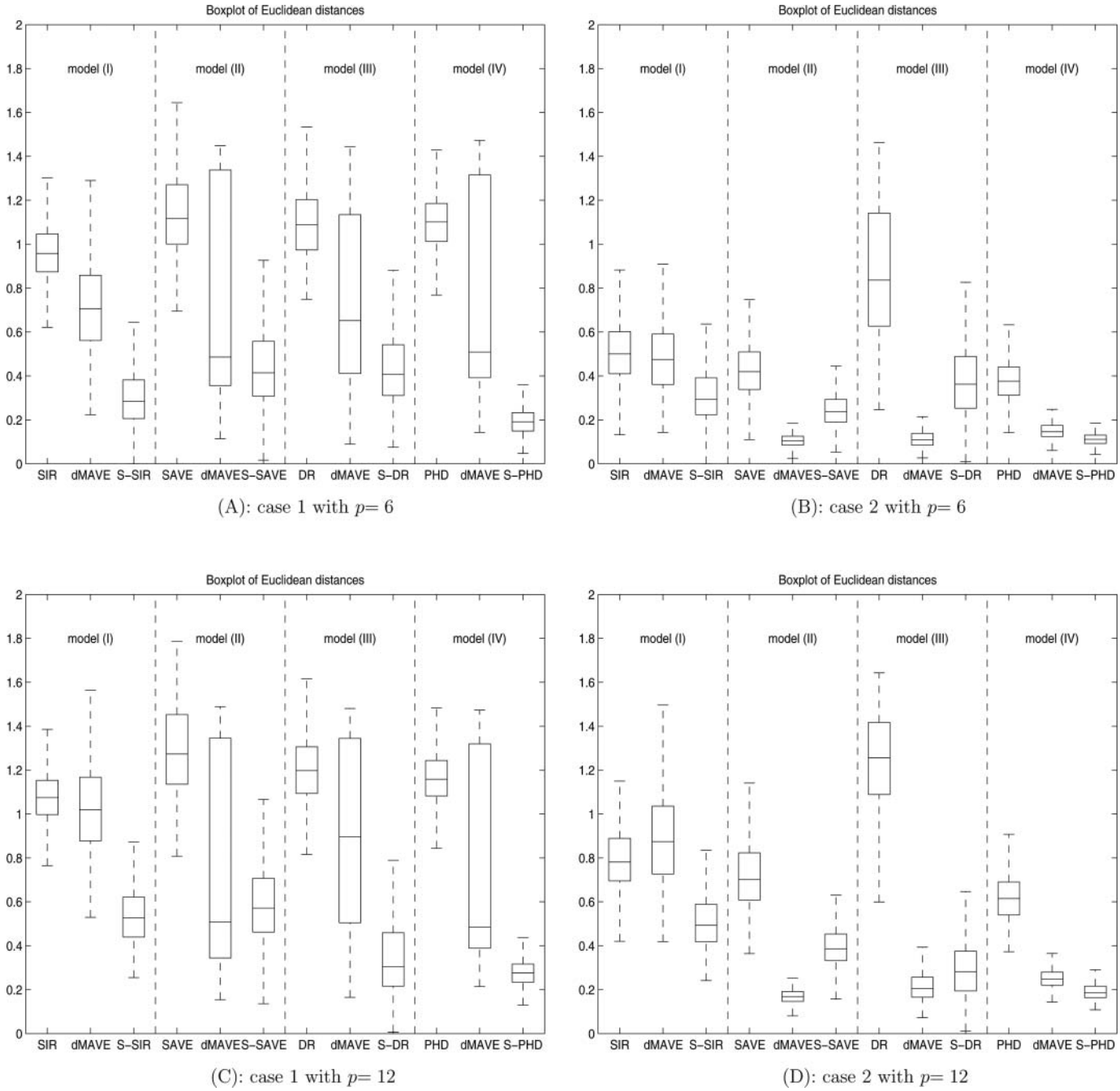

Figure 1. Boxplots of Euclidean distances for models (I)–(IV) with $p = 6$ and $p = 12$ in Example 1. Semi-SIR, Semi-SAVE, Semi-DR, and Semi-PHD are shorten to S-SIR, S-SAVE, S-DR, and S-PHD, respectively, in the labels.

to perform comparably with dMAVE, which usually produces very accurate estimates when all the covariates are continuous, such as the case here. Following the request of a referee, we also experimented the simulations with $p = 50$, which corresponds to 96 free parameters with the sample size $n = 200$. Unfortunately, none of the methods examined above can provide any reasonable results in this case. We believe that to handle p very large in comparison with n, additional assumptions such as sparsity is needed. The sparsity assumption assumes many rows of $\boldsymbol{\beta}$ are zeros, hence the corresponding variables in $\mathbf{x}$ simply have no effect on the response variable Y. Many existing methods are available to take advantage of the sparsity assumption (see, e.g., Li 2007; Wang and Wang 2010). Combining the sparsity techniques and the dimension reduction methods are promising future work.

6.2 Example 2

Next, we perform simulations to demonstrate the performance of the bootstrap procedure conjuncted with the semiparametric methods in estimating the structural dimension d. We continue to use models (I)–(IV) in Example 1, and generate the predictors $\mathbf{x}$ from case 1 when $p = 6$. We generated 100 datasets of sample size $n = 200$, with bootstrap size $B = 100$.

We report the relative frequency of the bootstrap selected dimensions for models (I)–(IV) in Table 1. It can be easily seen

Table 1. Relative frequency of the estimated dimension $\widehat{d}$

Model	Method	$\widehat{d}=1$	$\widehat{d}=2$	$\widehat{d}=3$	$\widehat{d}\geq 4$
(I)	Semi-SIR	1%	99%	0%	0%
(II)	Semi-SAVE	1%	97%	0%	2%
(III)	Semi-DR	3%	97%	0%	0%
(IV)	Semi-PHD	5%	95%	0%	0%

that with at least 95% accuracy the bootstrap method correctly chose the dimension, which we consider quite satisfactory.

7. REAL DATA APPLICATION

We illustrate further our semiparametric proposals through a dataset concerning the employees' salary in the Fifth National Bank of Springfield (Albright, Winston, and Zappe 1999). The aim of this study is to understand how an employee's salary associates with his/her personal characteristics. To this end, an employee's annual salary is the response variable Y, and six covariates are possibly associated with the salary: the employee's current job level, where a larger number indicates a higher rank (X_1); the employee's working experience at current bank, which is measured by the number of years of the employment (X_2); the employee's age (X_3); the experience of an employee at another bank prior to working at the Fifth National, which is measured by the number of years at other banks (X_4); the employee's gender (X_5); and a binary variable indicating whether the employee's job is computer related (X_6). We removed an obvious outlier in this dataset, leaving 207 observations in the subsequent analysis.

Under various working dimensions $k = 1, \ldots, 5$, we implement the proposed semi-DR method given in (7) in association with the bootstrap method described in Section 5, with bootstrap size $B = 1000$. The results are presented as boxplots in Figure 2. It is clear that bootstrap method favors $d = 1$, which indicates

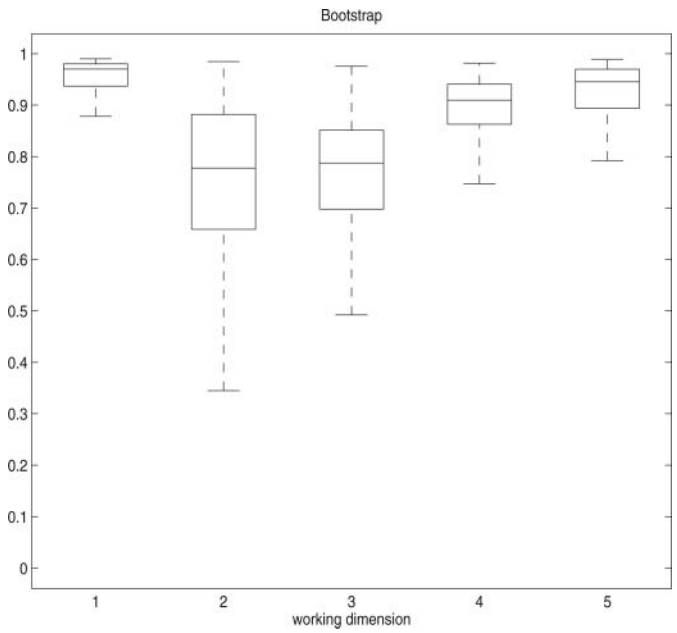

Figure 2. Boxplot of the $\bar{r}_k^2$ values defined in (10) from 1000 bootstrapped Semi-DR.

Table 2. Estimated coefficients of $\mathbf{x}$

Method	X_1	X_2	X_3	X_4	X_5	X_6
Semi-DR	0.7330	0.5794	0.2523	0.0147	0.1830	0.1723
dMAVE	0.8388	0.4404	0.2295	0.0417	0.1373	0.1711
DR	0.2673	0.8596	0.2829	−0.1023	0.3050	0.0782

that the six covariates affect the salary through one single linear combination direction. We present the estimation of this direction in Table 2. For comparison, we also implement the dMAVE and the classical DR with kernel smoothing. We emphasize here that dMAVE requires the covariates to be continuous, which is not the case for gender (X_5), and DR requires linearity condition and constant variance condition, which are not satisfied in this example (see Figure 3). Hence, both results should not be fully trusted.

To see the dimension-reduction effect from a different perspective, in Figure 4, we present the scatterplots of Y versus $\mathbf{x}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}$, where $\widehat{\boldsymbol{\beta}}$ denotes the estimate obtained from Semi-DR, dMAVE, and DR. We can see that the estimates obtained from Semi-DR and dMAVE exhibit a more obvious pattern than DR in that the data cloud appears more compact. We also include in Figure 4 three curves (the dashed lines) fitted from a quadratic function on $(\mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}, Y_i)$, $i = 1, \ldots, n$. These fit the data cloud rather well and illustrate a homoscedastic error pattern. Thus, we further used the fitted curves to perform a cross-validation procedure to calculate the prediction error. The resulting prediction errors from Semi-DR, dMAVE and DR are, respectively, 21.3191, 23.4481, and 47.0288, which further illustrate the advantage of the Semi-DR method. All these evidences support the conclusion that Semi-DR has successfully reduced the dimension from $p = 6$ to $d = 1$ and has found the right linear combination of the covariates in terms of establishing the association between salary and an employee's characteristic.

8. DISCUSSION

It is worth mentioning that (2) and (9) only differ in the first component. For (2), this component can be constructed from any function of Y and $\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}$, namely, any $\mathbf{g}(Y, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$, while for (9), the only valid choice is $\mathbf{g}(Y, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) = Y$. Thus, the family of estimators provided by (2) is much richer than the one provided by (9). In fact, it includes the family of (9) as a subfamily of estimators. This is easy to understand intuitively as model (1) assumes more structures than model (8), consequently one can have more ways to construct estimators for (1) and, therefore, model (1) has a larger subspace $\Lambda^{\perp}$. Because of this relation, any estimator for model (8) is necessarily also an estimator for (1). This also agrees with the dimension-reduction result that the conditional mean subspace $\mathcal{S}_{E(Y|\mathbf{x})}$ is a subspace of the conditional subspace $\mathcal{S}_{Y|\mathbf{x}}$, because if (1) holds for $\boldsymbol{\beta}$, then (8) also holds for $\boldsymbol{\beta}$ or a submatrix of $\boldsymbol{\beta}$.

For illustration, we have chosen to derive SIR, SAVE, and DR as examples for the central subspace problem and OLS and PHD as examples for the central mean subspace from the semiparametric approach, mainly due to their popularity in the dimension-reduction literature. Because $\Lambda^{\perp}$ contains all the influence functions, hence every root-n consistent method must correspond to a special choice of the functions in $\Lambda^{\perp}$. In this sense, the families given in Sections 2 and 4 are complete.

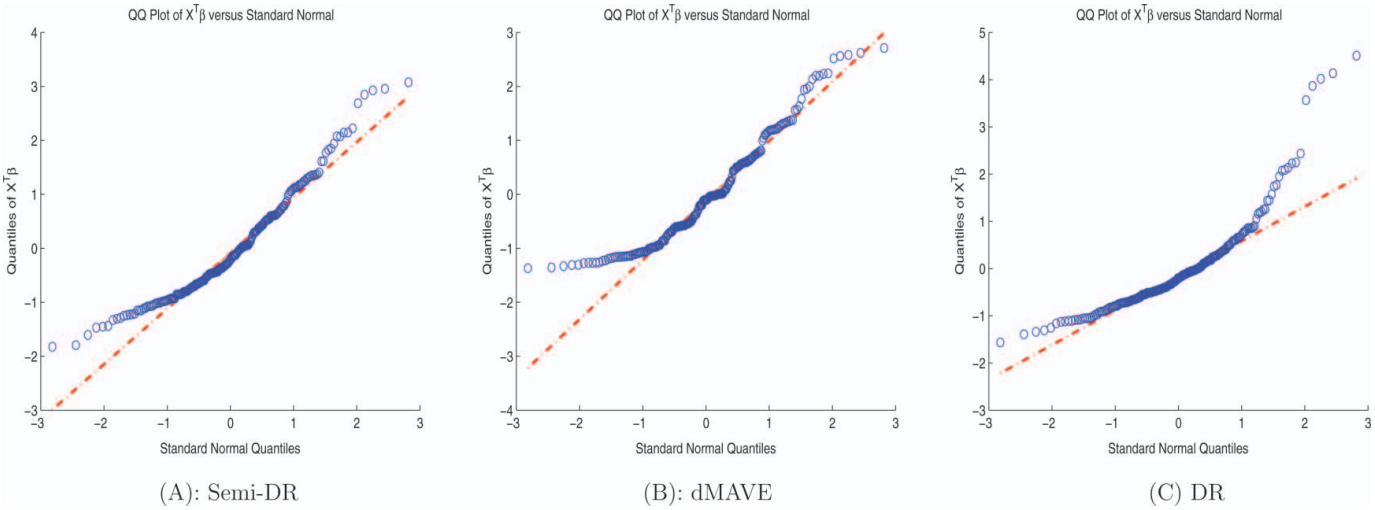

Figure 3. The Q–Q plot of $\mathbf{x}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_1$, with $\widehat{\boldsymbol{\beta}}_1$ estimated from Semi-DR, dMAVE, and DR. (The online version of this figure is in color.)

Finally, we acknowledge here that the linearity condition and the constant variance condition, when assumed to hold only at the true value of $\boldsymbol{\beta}$, can be mild and often hold approximately, especially when p is large (Hall and Li 1993). However, our simulations in Case 1 have shown that this approximation may not be sufficient to justify making and using these assumptions. In other words, if we assume the linearity condition and/or constant variance condition to hold exactly at the true $\boldsymbol{\beta}$ while, in fact, they are only approximately true, the subsequent estimation of the central subspace or central mean subspace can be quite different. In addition, even if the linearity and constant variance conditions do hold exactly, our limited numerical results seem to indicate that it might be still beneficial not to use them if one is concerned more about the estimation quality than the computational cost. Investigation on whether this is indeed a general phenomenon or some special isolated instances is a worthy endeavor.

APPENDIX: SOME TECHNICAL DERIVATIONS

A.1. General Introduction to Semiparametrics

Consider the Hilbert space $\mathcal{H}$ consisting of all the mean zero, finite variances, length m vector functions of $\mathbf{x}$, Y, where the inner product between two functions h, g is defined as $E(h^T g)$. Here and in the following definitions, all the expectations are calculated under the true distribution. The nuisance tangent space Λ is a subspace of $\mathcal{H}$ defined as the mean squared closure of all the elements of the form BS, where S is an arbitrary nuisance score vector function, and B is any conformable matrix with m rows. Here the nuisance score vector functions are calculated conventionally in every possible valid parameterization of the infinite-dimensional nuisance parameter, where a "valid parameterization" means that there exists one parameter value which yields the truth. Furthermore, $\Lambda^{\perp}$ is defined to be the orthogonal complement of Λ in $\mathcal{H}$.

Semiparametric theory ensures that every regular, asymptotic linear, root n consistent (RAL) estimator corresponds to an influence function, and every influence function is a normalized element in $\Lambda^{\perp}$, where the normalization is to ensure $E(\phi S_{\beta}^{\mathrm{T}}) = \mathbf{I}_m$. Here, S_{β} is the usual score vector with respect to the parameter of interest $\boldsymbol{\beta}$, and we use ϕ to denote an influence function and $\mathbf{I}_m$ to denote the size m identity matrix. This link between the complete family of RAL estimators and the space $\Lambda^{\perp}$ allows one to derive estimators through characterizing $\Lambda^{\perp}$ and identifying members in $\Lambda^{\perp}$. Every explicit identification of one function in $\Lambda^{\perp}$ ensures the discovery of one estimator. This is the semiparametric approach that drives all our derivations.

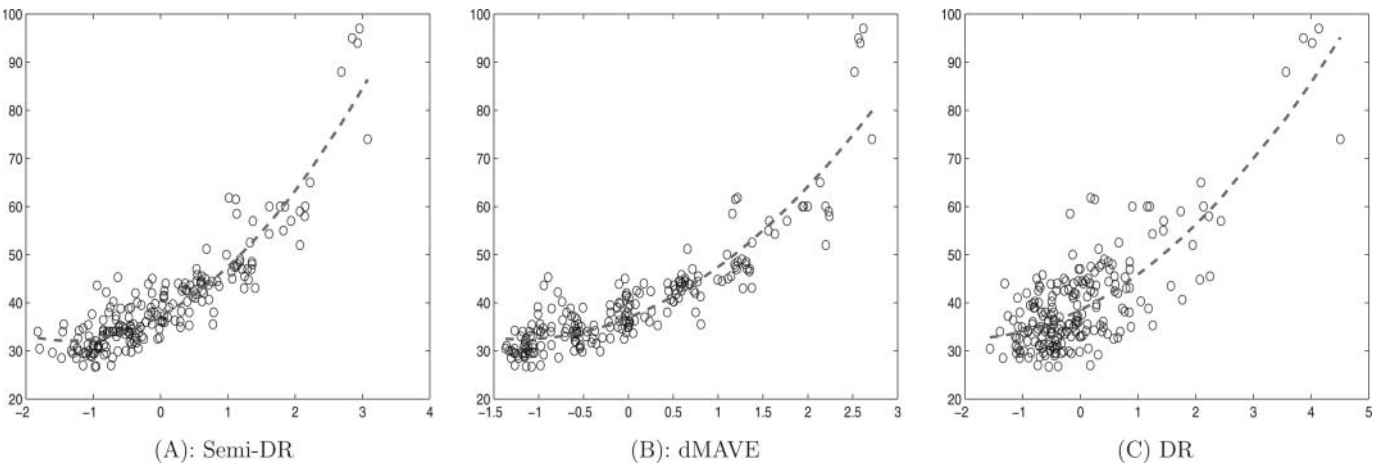

Figure 4. Scatterplot of Y versus $\mathbf{x}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_1$, with $\widehat{\boldsymbol{\beta}}_1$ estimated from Semi-DR, dMAVE, and DR. The dash lines are fitted curves based on quadratic regression modeling.

A.2. The Derivation of $\Lambda^\perp$ in Model (1)

Denote the nuisance tangent space corresponding to η_1 and η_2, respectively, Λ_1 and Λ_2. We have

$$\Lambda_1 = \{\mathbf{f}(\mathbf{x}) : \forall \mathbf{f} \text{ such that } E(\mathbf{f}) = \mathbf{0}\},$$
$$\Lambda_2 = \{\mathbf{f}(Y, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) : \forall \mathbf{f} \text{ such that } E(\mathbf{f} \mid \mathbf{x}) = E(\mathbf{f} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) = \mathbf{0}\}.$$

Obviously, $\Lambda_1 \perp \Lambda_2$, hence $\Lambda = \Lambda_1 \oplus \Lambda_2$. It is easy to see that $\Lambda_1^\perp = \{\mathbf{f}(Y, \mathbf{x}) : E(\mathbf{f} \mid \mathbf{x}) = 0\} \supseteq \Lambda_2$. We now show that $\Lambda_2^\perp = \{\mathbf{f}(Y, \mathbf{x}) : E(\mathbf{f} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}, Y)$ is a function of $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$ only.$\}$. Obviously, functions having the required conditional expectation property are certainly elements in $\Lambda_2^\perp$. To show that elements in $\Lambda_2^\perp$ have to satisfy the conditional expectation requirement, consider any $\mathbf{f}(Y, \mathbf{x}) \in \Lambda_2^\perp$. We let $\mathbf{g} = E(\mathbf{f} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}, Y) - E(\mathbf{f} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$. Obviously, $\mathbf{g} \in \Lambda_2$ hence $E(\mathbf{g}^{\mathrm{T}}\mathbf{f}) = 0$. On the other hand,

$$\begin{aligned} 0 = E(\mathbf{g}^{\mathrm{T}}\mathbf{f}) &= E\{\mathbf{g}^{\mathrm{T}}E(\mathbf{f} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}, Y)\} \\ &= E(\mathbf{g}^{\mathrm{T}}\mathbf{g}) + E\{\mathbf{g}^{\mathrm{T}}E(\mathbf{f} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})\} \\ &= E(\mathbf{g}^{\mathrm{T}}\mathbf{g}) + E\{E(\mathbf{g} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})^{\mathrm{T}}E(\mathbf{f} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})\} \\ &= E(\mathbf{g}^{\mathrm{T}}\mathbf{g}) + 0. \end{aligned}$$

Hence $\mathbf{g}$ itself should be zero. This means $\mathbf{f}$ indeed satisfies the conditional expectation requirement.

We now show that

$$\begin{aligned} \Lambda^\perp = \Lambda_1^\perp \cap \Lambda_2^\perp &= \{\mathbf{f}(Y, \mathbf{x}) - E(\mathbf{f} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}, Y) : E(\mathbf{f} \mid \mathbf{x}) \\ &= E(\mathbf{f} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) \forall \mathbf{f}\}. \end{aligned}$$

To see this, note that its form simply says for any element in $\Lambda^\perp$, it should satisfy the condition $E(\mathbf{g} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}, Y) = 0$. Let us denote the set $A = \{\mathbf{f}(Y, \mathbf{x}) - E(\mathbf{f} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}, Y) : E(\mathbf{f} \mid \mathbf{x}) = E(\mathbf{f} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) \forall \mathbf{f}\}$. Obviously, $A \subset \Lambda_2^\perp$. Because

$$\begin{aligned} &E\{\mathbf{f}(Y, \mathbf{x}) - E(\mathbf{f} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}, Y) \mid \mathbf{x}\} \\ &\quad = E\{\mathbf{f}(Y, \mathbf{x}) \mid \mathbf{x}\} - E(\mathbf{f} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}, Y) \mid \mathbf{x}\} \\ &\quad = E(\mathbf{f} \mid \mathbf{x}) - E\{E(\mathbf{f} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}, Y) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}\} \\ &\quad = E(\mathbf{f} \mid \mathbf{x}) - E(\mathbf{f} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) = 0, \end{aligned}$$

hence $A \subset \Lambda_1^\perp$ as well. Hence, $A \subset \Lambda^\perp$. On the other hand, for any $\mathbf{f} \in \Lambda^\perp$, because $\mathbf{f} \in \Lambda_2^\perp$, we have $E(\mathbf{f} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}, Y) = \mathbf{a}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$ for some $\mathbf{a}$. Writing this out, we obtain

$$\begin{aligned} \mathbf{a}(\boldsymbol{\beta}^{\mathrm{T}}x) &= \frac{\int_{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}=\boldsymbol{\beta}^{\mathrm{T}}x} \mathbf{f}(Y, \mathbf{x})\eta_1(\mathbf{x})\eta_2(Y \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})d\mu(\mathbf{x})}{\int_{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}=\boldsymbol{\beta}^{\mathrm{T}}x} \eta_1(\mathbf{x})\eta_2(Y \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})d\mu(\mathbf{x})} \\ &= \frac{\int_{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}=\boldsymbol{\beta}^{\mathrm{T}}x} \mathbf{f}(Y, \mathbf{x})\eta_1(\mathbf{x})d\mu(\mathbf{x})}{\int_{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}=\boldsymbol{\beta}^{\mathrm{T}}x} \eta_1(\mathbf{x})d\mu(\mathbf{x})}. \end{aligned}$$

Now, we have

$$\begin{aligned} \mathbf{a}(\boldsymbol{\beta}^{\mathrm{T}}x) &= \int \mathbf{a}(\boldsymbol{\beta}^{\mathrm{T}}x)\eta_2(Y \mid \boldsymbol{\beta}^{\mathrm{T}}x)d\mu(Y) \\ &= \int \frac{\int_{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}=\boldsymbol{\beta}^{\mathrm{T}}x} \mathbf{f}(Y, \mathbf{x})\eta_1(\mathbf{x})d\mu(\mathbf{x})}{\int_{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}=\boldsymbol{\beta}^{\mathrm{T}}x} \eta_1(\mathbf{x})d\mu(\mathbf{x})}\eta_2(Y \mid \boldsymbol{\beta}^{\mathrm{T}}x)d\mu(Y) \\ &= \frac{\int \int_{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}=\boldsymbol{\beta}^{\mathrm{T}}x} \mathbf{f}(Y, \mathbf{x})\eta_1(\mathbf{x})\eta_2(Y \mid \boldsymbol{\beta}^{\mathrm{T}}x)d\mu(\mathbf{x})d\mu(Y)}{\int_{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}=\boldsymbol{\beta}^{\mathrm{T}}x} \eta_1(\mathbf{x})d\mu(\mathbf{x})} \\ &= \frac{\int_{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}=\boldsymbol{\beta}^{\mathrm{T}}x} \int \mathbf{f}(Y, \mathbf{x})\eta_2(Y \mid \beta^{\mathrm{T}}\mathbf{x})d\mu(Y)\eta_1(\mathbf{x})d\mu(\mathbf{x})}{\int_{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}=\boldsymbol{\beta}^{\mathrm{T}}x} \eta_1(\mathbf{x})d\mu(\mathbf{x})} = 0 \end{aligned}$$

because $E(\mathbf{f} \mid \mathbf{x}) = 0$ due to $\mathbf{f} \in \Lambda_1^\perp$. Thus, elements in $\Lambda^\perp$ indeed have the form $\mathbf{f}(Y, \mathbf{x}) - E(\mathbf{f} \mid Y, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$. The requirement of these elements belonging to $\Lambda_1^\perp$ generates the second requirement of A, and we obtain $\Lambda^\perp \subset A$. This completes the derivation of $\Lambda^\perp$.

A.3. The Double Robustness of (2)

Denote $E^*(\mathbf{g} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$ the misspecified function of $E(\mathbf{g} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$. We have

$$\begin{aligned} &E[\{\mathbf{g}(Y, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) - E^*(\mathbf{g} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\}\{\boldsymbol{\alpha}(\mathbf{x}) - E(\boldsymbol{\alpha} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\}] \\ &\quad = E([E\{\mathbf{g}(Y, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) \mid \mathbf{x}\} - E^*(\mathbf{g} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})]\{\boldsymbol{\alpha}(\mathbf{x}) - E(\boldsymbol{\alpha} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\}) \\ &\quad = E[\{E(\mathbf{g} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) - E^*(\mathbf{g} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\}\{\boldsymbol{\alpha}(\mathbf{x}) - E(\boldsymbol{\alpha} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\}] \\ &\quad = E[\{E(\mathbf{g} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) - E^*(\mathbf{g} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\}E\{E(\boldsymbol{\alpha} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) - E(\boldsymbol{\alpha} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\}] \\ &\quad = 0. \end{aligned}$$

Denote $E^*(\mathbf{a} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$ the misspecified function of $E(\mathbf{a} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$. We have

$$\begin{aligned} &E[\{\mathbf{g}(Y, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) - E(\mathbf{g} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\}\{\boldsymbol{\alpha}(\mathbf{x}) - E^*(\boldsymbol{\alpha} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\}] \\ &\quad = E([E\{\mathbf{g}(Y, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) \mid \mathbf{x}\} - E(\mathbf{g} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})]\{\boldsymbol{\alpha}(\mathbf{x}) - E^*(\boldsymbol{\alpha} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\}) \\ &\quad = E[\{E(\mathbf{g} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) - E(\mathbf{g} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\}\{\boldsymbol{\alpha}(\mathbf{x}) - E^*(\boldsymbol{\alpha} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\}] = 0. \end{aligned}$$

A.4. The Derivation $\Lambda^\perp$ in Model (8)

Denote the nuisance tangent space corresponding to η_1, η_2, and m, respectively, Λ_1, Λ_2, and Λ_m. Straightforward derivation yields

$$\Lambda_1 = \{\mathbf{f}(\mathbf{x}) : \forall \mathbf{f} \text{ such that } F(\mathbf{f}) = 0\},$$
$$\Lambda_2 = \{\mathbf{f}(\epsilon, \mathbf{x}) : \forall \mathbf{f} \text{ such that } E(\mathbf{f} \mid \mathbf{x}) = 0, E(\epsilon\mathbf{f} \mid \mathbf{x}) = 0\},$$
$$\Lambda_m = \left\{\frac{\eta'_{20,\epsilon}(\epsilon, \mathbf{x})}{\eta_{20}(\epsilon, \mathbf{x})}\mathbf{h}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) : \forall \mathbf{h}\right\}.$$

Consequently, we have

$$\Lambda_1^\perp = \{\boldsymbol{\alpha}(\epsilon, \mathbf{x}) : \forall \boldsymbol{\alpha} \text{ such that } E(\boldsymbol{\alpha} \mid \mathbf{x}) = 0\},$$
$$(\Lambda_1 + \Lambda_2)^\perp = \{\boldsymbol{\alpha}(\mathbf{x})\epsilon : \forall \boldsymbol{\alpha}\}.$$

To further derive $\Lambda^\perp \subset (\Lambda_1 + \Lambda_2)^\perp$, we first inspect an $\boldsymbol{\alpha}(\mathbf{x})\epsilon \in \Lambda^\perp$. For any $\mathbf{h}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$, $\boldsymbol{\alpha}(\mathbf{x})$ satisfies

$$\begin{aligned} 0 &= \int \boldsymbol{\alpha}(\mathbf{x})\epsilon\frac{\eta'_{20,\epsilon}(\epsilon, \mathbf{x})}{\eta_{20}(\epsilon, \mathbf{x})}\mathbf{h}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\eta_{10}(\mathbf{x})\eta_{20}(\epsilon, \mathbf{x})d\mu(\epsilon)d\mu(\mathbf{x}) \\ &= \int \boldsymbol{\alpha}(\mathbf{x})\mathbf{h}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\eta_{10}(\mathbf{x})\left\{\int \epsilon\eta'_{20\epsilon}(\epsilon, \mathbf{x})d\mu(\epsilon)\right\}d\mu(\mathbf{x}) \\ &= -\int \boldsymbol{\alpha}(\mathbf{x})\mathbf{h}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\eta_{10}(\mathbf{x})\left\{\int \eta_{20}(\epsilon, \mathbf{x})d\mu(\epsilon)\right\}d\mu(\mathbf{x}) \\ &= -E\left\{\boldsymbol{\alpha}(\mathbf{x})\mathbf{h}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\right\}. \end{aligned}$$

This implies $E\{\boldsymbol{\alpha}(\mathbf{x}) \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\} = 0$. Therefore,

$$\begin{aligned} \Lambda^\perp &= [\{\boldsymbol{\alpha}(\mathbf{x}) - E(\boldsymbol{\alpha} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\}\epsilon : \forall \boldsymbol{\alpha}] \\ &= [\{Y - E(Y \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\}\{\boldsymbol{\alpha}(\mathbf{x}) - E(\boldsymbol{\alpha} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\} : \forall \boldsymbol{\alpha}]. \end{aligned}$$

A.5. Regularity Conditions for Theorem 1

(C1) The univariate kernel function $K(\cdot)$ is Lipschitz, has compact support. It satisfies

$$\int K(u)du = 1, \int u^i K(u)du = 0, 1 \le i \le m-1,$$
$$0 \ne \int u^m K(u)du < \infty.$$

The d-dimensional kernel function is a product of d univariate kernel functions, that is, $K_h(\mathbf{u}) = K(\mathbf{u}/h)/h^d = \Pi_{j=1}^d K_h(u_j) = \Pi_{j=1}^d K(u_j/h)/h^d$ for $\mathbf{u} = (u_1, \ldots, u_d)^{\mathrm{T}}$.

Here, we abuse the notation and use the same K regardless of the dimension of its argument.

(C2) Let $\mathbf{r}_1(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) = E\{\boldsymbol{\alpha}(\mathbf{x}) \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\} f(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$ and $\mathbf{r}_2(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) = E\{\mathbf{g}(Y, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\} f(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$. The mth derivatives of $\mathbf{r}_1(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$, $\mathbf{r}_2(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$ and $f(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$ are locally Lipschitz-continuous.

(C3) The density functions of $\mathbf{x}$ and $\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}$, denoted, respectively, by $f_{\mathbf{x}}(\mathbf{x})$ and $f(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$, are bounded from below and above. Each entry in the matrices $E\{\boldsymbol{\alpha}(\mathbf{x})\boldsymbol{\alpha}^{\mathrm{T}}(\mathbf{x}) \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\}$ and $E\{\mathbf{g}(Y, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\mathbf{g}^{\mathrm{T}}(Y, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\}$ is locally Lipschitz-continuous and bounded from above as a function of $\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}$.

(C4) The bandwidth $h = O(n^{-\kappa})$ for $1/(4m) < \kappa < 1/(2d)$.

[Received January 2011. Revised November 2011.]

REFERENCES

Albright, S. C., Winston, W. L., and Zappe, C. J. (1999), *Data Analysis and Decision Making with Microsoft Excel*, Pacific Grove, CA: Duxbury Press. [176]

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993), *Efficient and Adaptive Estimation for Semiparametric Models*, Baltimore, MD: The Johns Hopkins University Press. [169]

Box, G. E. P., and Cox, D. R. (1964), "An Analysis of Transformations" (with discussion), *Journal of the Royal Statistical Society,* Series B, 26, 211–252. [168]

Cook, R. D. (1998), *Regression Graphics: Ideas for Studying Regressions through Graphics*, New York: Wiley. [168,169,170]

Cook, R. D., and Forzani, L. (2009), "Likelihood-Based Sufficient Dimension Reduction," *Journal of the American Statistical Association*, 104, 197–208. [168]

Cook, R. D., and Li, B. (2002), "Dimension Reduction for Conditional mean in regression," *The Annals of Statistics*, 30, 455–474. [168,172]

Cook, R. D., and Li, L. (2005), "Dimension Reduction in Regressions with Exponential Family Predictors," *Journal of Computational and Graphical Statistics*, 18, 774–791. [170]

Cook, R. D., and Nachtsheim, C. J. (1994), "Reweighting to Achieve Elliptically Contoured Covariates in Regression," *Journal of the American Statistical Association*, 89, 592–599. [168]

Cook, R. D., and Weisberg, S. (1991), Discussion of "Sliced Inverse Regression for Dimension Reduction," *Journal of the American Statistical Association*, 86, 28–33. [168,171]

Dong, Y., and Li, B. (2010), "Dimension Reduction for Non-Elliptically Distributed Predictors: Second-Order Moments," *Biometrika*, 97, 279–294. [168,172,174]

Fung, W. K., He, X., Liu, L., and Shi, P. (2002), "Dimension Reduction Based on Canonical Correlation," *Statistica Sinica*, 12, 1093–1113. [168]

Hall, P., and Li, K. C. (1993), "On Almost Linearity of Low Dimensional Projections from High Dimensional Data," *The Annals of Statistics*, 21, 867–889. [176]

Härdle, W., Hall, P., and Ichimura, H. (1993), "Optimal Smoothing in Single-Index Models," *The Annals of Statistics*, 21, 157–178. [168]

Härdle, W., and Stoker, T. M. (1989), "Investigating Smooth Multiple Regression by the Method of Average Derivatives," *Journal of the American Statistical Association*, 84, 986–995. [168]

Hernández, A., and Velilla, S. (2005), "Dimension Reduction in Nonparametric Kernel Discriminant Analysis," *Journal of Computational and Graphical Statistics*, 14, 847–866. [168]

Henmi, M., and Eguchi, S. (2004), "A Paradox Concerning Nuisance Parameters and Projected Estimating Functions," *Biometrika*, 91, 929–941. [174]

Horowitz, J. L., and Härdle, W. (1996), "Direct Semiparametric Estimation of Single-Index Models with Discrete Covariates," *Journal of the American Statistical Association*, 91, 1632–1639. [168,173]

Ichimura, H. (1993), "Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single-Index Models," *Journal of Econometrics*, 58, 71–120. [168,173]

Li, K. C. (1991), "Sliced Inverse Regression for Dimension Reduction" (with discussion), *Journal of the American Statistical Association*, 86, 316–342. [168,171]

——— (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. [168,172]

Li, L. (2007), "Sparse Sufficient Dimension Reduction," *Biometrika*, 94, 603–613. [174]

Li, B., and Dong, Y. (2009), "Dimension Reduction for Non-Elliptically Distributed Predictors," *The Annals of Statistics*, 37, 1272–1298. [168,172]

Li, K. C., and Duan, N. (1989), "Regression Analysis Under Link Violation," *The Annals of Statistics*, 17, 1009–1052. [168,172]

Li, B., and Wang, S. (2007), "On Directional Regression for Dimension Reduction," *Journal of the American Statistical Association*, 102, 997–1008. [168,171,172]

Newey, W. K. (1990), "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics*, 5, 99–135. [170]

Park, J. H., Sriram, T. N., and Yin, X. (2010), "Dimension Reduction in Time Series," *Statistica Sinica*, 20, 747–770. [168]

Power, J. L., Stock, J. H., and Stoker, T. M. (1989), "Semiparametric Estimation of Index Coefficients," *Econometrika*, 51, 1403–1430. [168]

Tsiatis, A. A. (2006), *Semiparametric Theory and Missing Data*, New York: Springer. [169]

Wang, J., and Wang, L. (2010), "Sparse Supervised Dimension Reduction in High Dimensional Classification," *Electronic Journal of Statistics*, 4, 914–931. [174]

Wang, H., and Xia, Y. (2008), "Sliced Regression for Dimension Reduction," *Journal of the American Statistical Association*, 103, 811–821. [168]

Xia, Y. (2007), "A Constructive Approach to the Estimation of Dimension Reduction Directions," *The Annals of Statistics*, 35, 2654–2690. [168,174]

Xia, Y., Tong, H., Li, W. K., and Zhu, L. X. (2002), "An Adaptive Estimation of Dimension Reduction Space" (with discussion), *Journal of the Royal Statistical Society,* Series B, 64, 363–410. [168,174]

Ye, Z., and Weiss, R. E. (2003), "Using the Bootstrap to Select one of a New Class of Dimension Reduction Methods," *Journal of the American Statistical Association*, 98, 968–979. [174]

Yin, X., and Cook, R. D. (2005), "Direction Estimation in Single-Index Regressions," *Biometrika*, 92, 371–384. [168]

Yin, X., Li, B., and Cook, R. D. (2008), "Successive Direction Extraction for Estimating the Central Subspace in a Multiple-Index Regression," *Journal of Multivariate Analysis*, 99, 1733–1757. [168]

Zhu, L. X., and Fang, K. T. (1996), "Asymptotics for Kernel Estimation of Sliced Inverse Regression," *The Annals of Statistics*, 3, 1053–1068. [168]

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. [168]